AMERICAN INSTITUTES FOR RESEARCH®

STATE OF NEW HAMPSHIRE
Department of Education

RESPONSE TO PROPOSAL: RFP 2017-073
DOE New Hampshire Statewide Assessments

TECHNICAL PROPOSAL

American Institutes for Research®
1000 Thomas Jefferson Street, NW
Washington, DC 20007-3835
Contact: Jon Cohen, President, Assessment
Contact Phone: 202-403-5420
Fax: 202-403-5303
Email: JCohen@air.org

April 26, 2017

COPY

Table of Contents

Table of Contents

Transmittal Form Letter

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

State of New Hampshire Proposal Transmittal Form Letter

Company Name American Institutes for Research (AIR)
Address 1000 Thomas Jefferson St. NW, Washington, DC 20007-3835

To: NH Department of Education
State Point of Contact: Sandie MacDonald
Telephone: (603) 271-3453
Email: saundra.macdonald@doe.nh.gov

RE: Proposal Invitation Name: **RFP 2017-073 DOE New Hampshire Statewide Assessments**
Proposal Due Date and Time: Wednesday, April 26, 2017 at 4:00 PM

Dear Project Manager:

Company Name: American Institutes for Research hereby offers to sell to the State of New Hampshire the services indicated in **RFP 2017-073 DOE New Hampshire Statewide Assessments** at the price(s) quoted in vendor Response Section VII: *Pricing Model*, and Appendix F: *Pricing Worksheets*, in complete accordance with all conditions of this RFP and all specifications set forth in the RFP and in the State of New Hampshire Terms and Conditions outlined in RFP Appendix H: *State of New Hampshire Terms and Conditions*.

[Company Signor:] Michael Casey, J.D. Contracts Officer is authorized to legally obligate [Company Name:] American Institutes for Research.

We attest to the fact that:

The company has reviewed and agreed to be bound by all RFP terms and conditions including but not limited to the *State of New Hampshire Terms and Conditions* in Appendix H, which shall form the basis of any Contract resulting from this RFP; No new terms and conditions have been added and no existing terms and conditions have been deleted in this RFP Proposal.

The proposal is effective for a period of 180 days or until the effective date of any resulting contract.

The prices quoted in the proposal were established without collusion with other eligible vendors and without effort to preclude the State of New Hampshire from obtaining the best possible competitive price; and

The vendor has read and included a copy of RFP 2017-073 DOE New Hampshire Statewide Assessments and any subsequent signed Addendum (a).

Our official point of contact is Jon Cohen
Title President, AIR Assessment
Telephone 202-403-5420 Email JCohen@air.org
Authorized Signature Printed Michael Casey, J.D. Contracts Officer

Authorized Signature

13

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

April 26, 2017

Sandie MacDonald
State of New Hampshire
Department of Education
101 Pleasant Street
Concord, New Hampshire 03301

RE: RFP 2017-073 DOE New Hampshire Statewide Assessments

Dear Ms. MacDonald,

The American Institutes for Research (AIR) is pleased to submit this proposal to deliver a system for administering, scoring, reporting, and providing related services for New Hampshire's Statewide Assessments in ELA, Mathematics, and Science.

Founded in 1946 and based in Washington, DC, AIR is a not-for-profit organization pursuing our mission to use the best social and behavioral sciences to improve people's lives. We are more than 1,800 people working in the areas of assessment, education research and technical assistance, health, human development, and international development.

In student assessment, pursuing our mission extends to collaborating with our clients to deliver assessments that

- assess the achievement level and learning gains of all students, as well as their educational strengths and needs;
- report timely or immediately actionable data for making decisions at the classroom and school levels; and
- support equal opportunity for all students by providing the tools needed to access and use the online tests effectively and without distraction.

AIR currently delivers approximately 30 million online tests in 24 states, plus the U.S. Virgin Islands, and delivers online, adaptive tests in 14 of those states, plus the U.S. Virgin Islands. Many clients are surprised to discover that AIR is now the undisputed leader in the United States for statewide summative assessment testing. In addition to delivering 43 different statewide alternate assessments, EOC assessments, science assessments, and English language proficiency testing programs, AIR delivers the core grades 3–8 ELA and mathematics testing in 17 states, serving 34% of these students nationwide. No other vendor serves more than 11 states or more than 21% of these students, and no other vendor tests anywhere near as many students online as AIR. We also deliver approximately 1.5 million science tests each year in California, Connecticut, Delaware, Hawaii, Idaho, Ohio, Oregon, Utah, Washington, and West Virginia.

Per RFP Section 4.10, Public Disclosure, AIR has designated approximately five pages of our proposal as confidential, in accordance with New Hampshire code RSA 91-A:5, IV and other applicable laws. We have identified approximately five pages of our proposal that contain trade secret information on the design of AIR's test delivery system. These pages qualify for exception because we take great effort to keep this information secret, as AIR derives actual independent economic value from this method and associated formulae for the design, configuration, and operation of our test delivery system. This trade secret information provides AIR the ability to support many more simultaneous users in an Internet-based test delivery system that works in schools than previous systems have been able to support. AIR is the only organization that has proven the ability to deliver adaptive tests to hundreds of thousands of simultaneous users in schools. I am authorized to respond to inquiries related to this confidential information, and my contact information is included herein.

We believe that our team's collective efforts bring the capabilities that New Hampshire will need to successfully deliver the statewide assessments in ELA, mathematics, and science. Should you have any questions or require any clarification, please contact me.

Sincerely,

Jon Cohen
President, AIR Assessment
American Institutes for Research
1000 Thomas Jefferson Street NW
Washington, DC 20007
Phone: 202-403-5420
Fax: 202-403-5303

Section I:
Executive Summary

Section I: Executive Summary

AIR is pleased to offer the New Hampshire Department of Education (NH DOE) a system of standards-based summative and interim assessments that are aligned to the New Hampshire College and Career Ready Standards. We propose to deliver these assessments using AIR's *Independent College and Career Readiness* (ICCR) item banks, which have been developed to measure college and career ready skills in English language arts and mathematics, and three-dimensional science skills described in the K–12 Framework. ICCR banks are administered in statewide summative assessments in Arizona, Florida, Ohio, Tennessee, and Utah, and, starting in 2018, Wyoming. This plan will offer New Hampshire shorter testing times, rigorous measurement of New Hampshire standards, and the ability to compare scores to many states, including some comparisons to New Hampshire's own Smarter Balanced scores from the past three years.

Many states that use the SAT in high school are surprised to learn that using AIR's online testing platform in lower grades will provide a seamless experience from elementary school through high school. The College Board has recently chosen AIR as its online test delivery partner to move the SAT Suite of tests online. The SAT will be offered online as soon as spring 2018, and will be delivered on AIR's platform, which has reliably served New Hampshire schools and students for the last three years.

We offer a system that New Hampshire knows from experience will work seamlessly in its schools. Using AIR's system, New Hampshire was able to transition from a primarily paper system to virtually 100 percent online in a single year. Similar success throughout the country, a rich selection of accessible item types, and the industry's most robust collection of operational accommodations and embedded supports contributed to the College Board's decision to partner with AIR to go online.

AIR values our relationship with New Hampshire, and we hope to continue serve New Hampshire's schools and students as the state transitions to independent measurement of its standards. New Hampshire's schools and students will experience seven key benefits from AIR's proposed assessment system.

- Our test delivery system is familiar and already reliably delivering summative assessments to all New Hampshire students. The College Board has recently selected this same test delivery system to administer the SATs, New Hampshire's new high school accountability assessment. Test coordinators, test administrators, and other educators are familiar with our systems. Importantly, this includes our reporting system, which will continue to look and feel the same, even as the data reported transition to new tests and new reporting categories.
- A test built on the ICCR bank supports cross-state comparisons. The ICCR bank of items in ELA and mathematics is currently being administered as part of state assessments in Arizona, Florida, Ohio, Tennessee, and Utah, and, starting in 2018, in Wyoming, providing a group of states to which comparisons can be made. A linkage to Smarter Balanced also allows some longitudinal comparisons to support the transition.
- In science, an alliance of states offers the opportunity for item sharing to reduce costs and get a more robust item pool for the science test. The ICCR bank of science items is developed in collaboration with a group of states developing common item and item cluster specifications designed to measure three-dimensional science.
- Educators will receive immediate, actionable data. AIR's adaptive assessment design for ELA and mathematics and matrix design for science proceed from the goal of reporting valid and reliable test scores at the most granular level, with test results reported immediately using our intuitive online reporting system to communicate results.
- Students and schools will experience reduced testing times. Actual testing times from adaptive administrations of standards-based summative assessments consistent with our proposed assessment system show 85th percentile testing times, consistent with New Hampshire's testing time goals.

- Our technology supports rich measurement of the breadth and depth of the standards. The ICCR item bank includes a wide array of machine-scored item types that allow the Department to measure the full range and depth of the New Hampshire College and Career Ready Standards using constructed-response items that cost-effectively measure more complex cognitive processes.
- School and district technology coordinators already know that they can deliver tests online using our system without investments in extra bandwidth, more servers, or extensive technology expertise required by many other systems.

In the pages that follow, we summarize our proposed assessment system and the proven technology used to deliver those assessments to provide the Department with a 21st century assessment system to validly, reliably, and fairly measure student achievement of New Hampshire's College and Career Ready Standards for students who are being raised with 21st century technology.

Assessment System

AIR is pleased to offer New Hampshire an assessment system solution that provides for both standards-based summative and interim assessments aligned to New Hampshire's College and Career Ready Standards. Our solution combines a state-of-the-art online item bank, designed from the ground up to use the full range of response mechanisms afforded by online testing, incorporating the industry's largest selection of operationally deployed accessibility tools and accommodations into the items, and automated scoring of student responses to support adaptive test administration and rapid scoring and reporting of test results. Our state-of-the-art test delivery system administers these items following any design that the Department may desire. We propose to deliver standards-based adaptive test administrations, but our test delivery system can just as easily administer fixed-form or stage-adaptive assessments as well. AIR's test delivery system is highly robust, and reliably delivers more than 30 million test administrations annually, 18.8 million of which are secure summative test administrations including New Hampshire's administration of the Smarter Balanced assessments the past three years. Moreover, the College Board has recently selected AIR to provide test delivery for the SAT assessments, so that should the Department award AIR the contract for New Hampshire's statewide assessments, the test delivery system will be common across the grades 3–8 and high school assessment systems.

ELA and Mathematics. For New Hampshire's summative assessments in ELA and mathematics, we propose to administer standards-based, computer adaptive tests based on our ICCR item banks, which have been developed to measure career and college-ready standards in statewide assessments.

AIR's ICCR item pools have been constructed explicitly to support statewide assessment programs and have been administered as part of the Arizona, Florida, Ohio, Tennessee, Utah, and Wyoming statewide assessments; and they have been embedded in Oregon's administration of the Smarter Balanced assessments. ICCR items are developed in conjunction with state departments of education, following a rigorous system of internal and external review procedures implemented by each of the participating states. Thus, all ICCR items have been reviewed by content review committees comprising educators in one or more states, as well as by bias and fairness review committees in those states. All ICCR items have been embedded within operational test administrations for field testing so that resulting item parameter estimates, based on large numbers of students participating in state summative assessments, are highly precise and stable. Following field-test administration, item statistics for all ICCR items are evaluated for discrimination, difficulty, and differential item functioning, and any items flagged for out-of-range statistics are forwarded for further review by AIR content staff, AIR psychometric staff, and state assessment staff, with items that performed poorly being rejected from the pool.

The ICCR item bank has been constructed to support a flexible range of blueprints, from which the state can craft and implement a custom blueprint that meets that state's unique requirements and needs. AIR will work with the Department to finalize a blueprint that aligns to the New Hampshire College and Career Ready Standards. This will yield test scores that are valid and reliable, both overall and for all domain reporting categories, and ensure that test administration times remain within desired limits.

Science Assessments. As the Department is aware, AIR is currently working with a consortium of states to develop item specifications to support the development of state-specific ICCR item clusters and discrete item types to measure the three-dimensional science framework that has informed New Hampshire's College and Career Ready Standards in science. As with the ICCR item pools in ELA and mathematics, the ICCR item pools in science will allow the Department to identify the performance standards for any consortium state on the common ICCR scale. ICCR science items will be available for administration in spring 2018. At that time, we propose to administer science items in an operational field-test design that will administer test forms that meet all blueprint specifications, allow for calibration and equating of science items to establish the ICCR science scale, and support identification and adoption of performance standards for New Hampshire's statewide assessments in science.

Interim Assessments

To support more truly formative uses of interim test results, we propose an interim test design around AIRWays, our innovative reporting tool for interim benchmark assessments. AIRWays is designed to leverage testing events to drive students and teachers to interact. The system reports only non-secure test results and shows the teacher both the item and each student's actual response. This provides a platform and opportunity for the teacher and student to begin exploring gaps in knowledge to support instruction. AIRWays fosters more truly formative assessments by allowing teachers and students to explore their responses to test items and to focus instruction in more meaningful ways. Truly formative testing requires teachers and students to work together to understand how and why students respond to test items, which precludes administration of secure item content.

The NH DOE may eventually wish to aggregate student interim results to replace a summative test administration for accountability purposes. Since any accountability use of test items requires rigorous test security, an assessment system in which interim assessments replace a summative assessment for accountability would have to be administered using the same ICCR item pool and same test administration procedures as the summative assessments, precluding non-secure reporting of test results in AIRWays. The ICCR item pools are sufficiently large to support both interim and summative test administrations, and AIR's adaptive algorithm is configured to ensure that students are not administered the same item across test administrations. However, until the U.S. Department of Education determines that summative assessments can be replaced by aggregating a series of interim assessments, we propose to deliver New Hampshire's interim assessments using one of several non-secure formative assessment banks.

AIR can offer one of a variety of career and college readiness item banks for the interim assessments, all of which are comparable to the ICCR items. AIR will work with the Department to embed interim items in the summative assessments to calibrate and equate the item pools to the New Hampshire scale, so that interim assessment results based on these item pools can be reported on the New Hampshire reporting scale.

We are currently negotiating with Utah to access its formative item pool populated with items already calibrated and equated to the ICCR scale. The item pools making up Utah's formative assessments have several very important strengths. Utah's formative item pool is composed of the same kind of items and item types used to administer Utah's Student Assessment of Growth and Excellence (SAGE) accountability assessments. SAGE items were originally written to item specifications virtually identical to ICCR item specifications. The items were developed following the same rigorous procedures described in Topic 2a Item Development for ICCR item development, and already have precise item parameters.

Technology

A next-generation assessment should begin online. Careers from the shop floor to the executive offices all rely on technology. As long ago as 2009, 97 percent of classrooms had computers for instructional use (U.S. Department of Education, 2010). In 2013, more than 92 percent of children lived in a household with a computer (including handheld devices, File & Ryan, 2014). Technology is ubiquitous for the current generation of children, students, and workers.

Authentic Measurement of Important Skills

Computer-based assessments hold the potential to create tests and items that provide more authentic evaluation of the skills and knowledge we want our children to learn. AIRCraft, our item authoring tool, was developed from the ground up to support a digital assessment. AIRCraft allows our test developers to go beyond pre-defined item types. A single item can integrate multiple different interactions (e.g., a graph and an equation and a selection from among choices), allowing for deeper, richer measurement of what students know and can do. Rather than simply matching a key for a correct response, scoring rules for AIRCraft items can search the student response for patterns and relationships that provide evidence of specific skills and standards being measured. In Topic 2, Item Development, we present annotated examples of items developed in AIRCraft for each subject area. Readers can interact with these and other samples at https://demo.tds.airast.org/nh.

Small Footprint

AIR has had great success helping our client states transition to online assessment. Some states have chosen to make a gradual transition online, including Arizona, Florida, and Ohio. But other states chose to make an immediate transition to online assessment. Seven states/jurisdictions with whom we are working, including New Hampshire, transitioned to virtually 100 percent online assessment in 2015. These states and jurisdictions are:

1. California
2. Connecticut
3. Idaho
4. New Hampshire
5. Vermont
6. South Dakota
7. U.S. Virgin Islands

These include large, midsize, and small states and states/jurisdictions with urban, rural, and island populations. These rapid-transition states/jurisdictions all experienced successful transition.

In the words of Michael Hock, the Director for Educational Assessment for Vermont,

> *A few weeks ago the Washington Post ran an article headlined "Technical Glitches Plague Computer-based Standardized Tests Nationwide?" The long list of problems experienced by both PARCC and Smarter Balanced states cataloged by the story might lead a reader to believe that the move to digital test delivery was premature and perhaps ill-advised. It gave only brief mention of the few states, including big California and little Vermont, which had good first experiences with the new technologies and were generally pleased with the outcomes. What we know, and what the article failed to mention, is that the "happy states" had one thing in common. They had all contracted with the American Institutes for Research to implement their new assessment systems.*
>
> *-Michael Hock*

The key to our success in helping states to transition to online assessment has been our respect for the infrastructure that exists in schools. We have designed our system with what we refer to as a small footprint. Our system works on the newest tablets as well as the oldest desktops, because we know that schools must work with the technology available. School technology should support teaching and learning, and assessments should work within that infrastructure. Our assessments require very little bandwidth, despite rich interactions and stimuli. They are engineered this way because we know that even schools with good Internet connections can sometimes have internal bottlenecks that throttle bandwidth. Most importantly, our assessments are designed to require virtually no technology expertise in schools.

We know that schools and districts often lack access to skilled technicians. Our system requires only a single piece of software—the AIR secure browser—and no special hardware. The secure browser can be installed quickly and simply, either on a single machine or across a large network. Once installed, it takes care of itself.

On many days in April this past year, we experienced peak simultaneous loads approaching or exceeding 500,000 students all testing at the same time. This number corresponds to about 1,000,000–1,500,000 tests administered in a day. Across the country, schools are succeeding with online testing, and we are proud to be part of that. We believe that we can help New Hampshire make this transition smoothly.

Summary

We believe that AIR's record of successfully transitioning states to online assessment has been unparalleled. We currently deliver assessments online in more than 20 states, most of which are completely or nearly completely online. Among those, seven, including New Hampshire, made an abrupt and successful transition from all paper to all online assessments in 2015.

As a not-for-profit organization, we take our mission seriously. In assessment, that means using the best science to build and deliver assessments that measure the skills that we really want students to have, and to do so in a way that is the least disruptive to schools. We are committed to constructive partnerships with our clients.

While we must manage to the contract scope, we continuously improve our services. Where our clients want something that will improve the student experience, improve assessment or education, or reduce the administrative burden in states, we often build those capabilities at our own expense. These improvements are offered to all clients as part of our mission.

References

U.S. Department of Education (2010). National Center for Education Statistics, Fast Response Survey System (FRSS). *Internet Access in U.S. Public Schools and Classrooms: 1994-2005 and Educational Technology in U.S. Public Schools: Fall 2008*; and unpublished tabulations.

File, T. and Ryan, C. (2014) Computer and Internet Use in the US: 2013. American Community Survey Reports, U.S. Census Bureau.

Section II:
Glossary of Terms and Abbreviations

Section II: Glossary of Terms and Abbreviations

AIR presents our Glossary of Terms and Abbreviations in two exhibits:

- Exhibit II-1: Glossary of Acronyms
- Exhibit II-2: Glossary of Terms

Exhibit II-1: Glossary of Acronyms

Acronym	Glossary Description
2PL	Two-parameter logistic model
3DES	Triple Data Encryption Algorithm
3PL	Three-parameter logistic model/Generalized Partial Credit IRT model
ADA	Americans with Disabilities Act
AES	Automated Essay Scoring engine
AES	Advanced Encryption Standard
AI	Artificial Intelligence
AIR	American Institutes for Research
APIP/APIP 2.2	Accessible Portable Item Protocol/ Accessible Portable Item Protocol 2.2
ARIA	Accessible Rich Internet Applications
ASL	American Sign Language
CCSSO	Council of Chief State School Officers
CLS	Common login system
COPPA	Children’s Online Privacy Protection Act
CSSC	Computer and Statistical Sciences Center
CSV	comma-separated value (file format)
DEI	data entry interface
DIF	Differential item functioning
DNS	Domain Name Server/System
DoK	Depth of Knowledge
DoR	Database of Record
DoS	Denial of Service
DDoS	Distributed Denial of Service
ELA	English language arts
ELL	English language learner
EM/EM algorithm	Expectation maximization/expectation maximization algorithm
EOC	End-of-Course
ESSA	Every Student Succeeds Act
FAQ	Frequently asked questions
FERPA	Family Educational Rights and Privacy Act
FISMA	Federal Information Security Management Act
FIPS/FIPS 140	Federal Information Processing Standard/Federal Information Processing Standard (FIPS) Publication 140-2
HTTPS	Hypertext transfer protocol secure

Exhibit II-1: Glossary of Acronyms (continued)

Acronym	Glossary Description
iAM	AIR's statistical software
ICC	Item characteristic curve
ICCR	Independent College and Career Readiness
ICMP	Internet Control Message Protocol
ID	Identification
IT	Information Technology
IP	Internet Protocol
IPS	Intrusion Prevention System
IPSec	Internet protocol security
IRT	item response theory
ITS	Item Tracking System
JAWS	Jobs Access With Speech screen reader
Kbps	Kilobits per second
KDS	Key Data Systems
MH	Mantel-Haenszel
NAEP	National Assessment of Educational Progress
NCES	National Center for Education Statistics
NH DOE	New Hampshire Department of Education
NOC	Network Operation Center
NTP	Network Time Protocol
OIB	Ordered Item Booklet/online Ordered Item Booklet
ORS	Online Reporting System
OS	Operating System
PARCC	Partnership for Assessment of Readiness for College and Careers
PDF	Portable Document Format
PII	Personally identifiable information
PISA	Programme for International Student Assessment
PLD	Performance level descriptors
QA	Quality Assurance
QC	Quality control
QM	Quality Monitor
QTI	Question and Test Interoperability
RAID	Redundant array of independent disks
RFP	Request for Proposal
RTS	Roster Tracking System
SAGE	Student Assessment of Growth and Excellence
SEM	Standard Error of Measurement
SEMCs	Standard error of measurement curves
SIF	simple interaction file (file format)
SIRVE	Secure Item Response Viewing Environment
SFTP	Secure File Transfer Protocol
SMD	Standardized mean difference
SME	Subject matter expert

Exhibit II-1: Glossary of Acronyms (continued)

Acronym	Glossary Description
SQL	Structured Query Language
SSDP	Simple Service Discovery Protocol
SSID	State student identifier
SSH	Secure shell
SSL	Secure sockets layers
SSO	Single sign-on
TA	Test Administrator
TA Interface	Test Administrator Interface
TAC	Technical advisory committee
TAM	Test Administration Manual
TCC	Test characteristics curves
TCM	Test Coordinator Manual
TCP	Transmission Control Protocol
TCP RST	Transmission Control Protocol Reset (flag)
TCP SYN	Transmission Control Protocol Synchronize (flag)
TDS	Test delivery system
TIC	Test information curve
TIDE	Test Information Distribution Engine
TIMSS	Trends in International Mathematics and Science Study
TTS	Text-to-speech
UAT	User acceptance testing
UDP	User Datagram Protocol
UI	User interface
USED	United States Department of Education
VPN	Virtual private network
WCAG/WCAG 2.0 AA	Web Content Accessibility Guidelines/Web Content Accessibility Guidelines 2.0 AA certification status
XML	Extensible Markup Language (file format)

Exhibit II-2: Glossary of Terms

Term	Glossary Description
8x8 Virtual Call Center	All telephone calls, e-mails, and chat interactions between Help Desk agents and users are recorded within the 8x8 Virtual Call Center system. The 8x8 system provides supervisors the ability to listen to recorded calls, monitor real-time call queue statistics, and retrieve case histories sorted by caller.
AIRCraft	AIR's proprietary item authoring tool, developed from the ground up to support a digital assessment; allows our test developers to go beyond pre-defined item types by integrating multiple different interactions, allowing for deeper, richer measurement of what students know and can do.
AIRWays	AIR's proprietary innovative reporting tool for interim benchmark assessments, designed to leverage testing events to drive students and teachers to interact. The system reports only non-secure test results and shows the teacher both the item and each student's actual response.
Bookmaps	Test maps
Database of Record (DoR)	Serves as our consistent repository of data. The extract system, which will interface with the state's systems to deliver test and item data, will draw data from the DoR. The DoR will send data to the Response Bank, our data system that supports the ORS, as well as to our Score Reporting team.
Data entry interface (DEI)	AIR's proprietary platform used by scribes to enter student testing data for students who cannot interact with the computer and require large print or braille accommodations.
Department	New Hampshire Department of Education
Diagnostic Tool	A readiness tool that uses a sophisticated statistical model of the testing process that models variation in instantaneous demand (e.g., the number of students who simultaneously press the Next button) and evaluates the likelihood that peaks will exceed network free capacity (not used by other processes) with a frequency likely to cause noticeable delays in testing, used to assess system readiness for testing. The tool is simple, effortless, and accurate.
Excel	Microsoft Excel
FormBuilder	AIR's software that assists our test developers as they construct operational forms. FormBuilder interfaces with ITS to extract item and test information and interactively evaluates the form's match to blueprint to ensure that test forms meet all blueprint specifications.
HTML5	A markup language used for structuring and presenting content on the World Wide Web. It is the fifth and current version of the HTML standard.
i4see	Platform used by New Hampshire and AIR to support the nightly exchange of student registration/enrollment files for the 2016-2017 Smarter Balanced assessments. This nightly exchange allows AIR to automatically import student demographic and accommodation information into TIDE or make updates to existing information in

	TIDE.
ICCR	AIR item banks, which have been developed to align with the college and career ready skills in English language arts and mathematics, and the three-dimensional science skills described in the K–12 Framework; can be used to implement a variety of assessment designs, including fixed-form assessments, stage adaptive assessments, and fully adaptive standards-based assessments.
Item Tracking System (ITS)	AIR’s proprietary system that facilitates item development and review processes.
KnowledgeTree	AIR proposes that all major project documents be stored in KnowledgeTree. KnowledgeTree is an online document management system on a secure cloud computing platform. KnowledgeTree manages document workflow and file sharing; alerts document followers to changes; and maintains version control, including authors and document history. Documents are checked in and checked out for editing. Access to documents can be controlled through a user roles and permissions system.
Online Reporting System (ORS)	AIR’s proprietary reporting platform that provides educators with a highly intuitive and powerful tool for navigating both summative and interim assessment results in the aggregate or at individual levels.
OpenAM	Open-source access management solution and a federation server platform that is highly scalable and offers industry-standard security capabilities; single sign-on system that enables all AIR online testing systems to appear to users as a single, integrated system. Once logged in, users can navigate the various components of the system securely and seamlessly.
Quality Monitor (QM)	After a test is administered to a student, the TDS passes the resulting data to our Quality Monitor (QM) system. The QM system rescores tests, checks that the tests meet blueprints, captures statistics on items, and runs a host of extensive quality checks. The QM system also runs a suite of analyses designed to detect cheating, which our psychometricians can access at any time. From the QM system, data are passed to the database of record (DoR).
Rackspace	Most of our data and test items are stored on AIR-dedicated servers at Rackspace, which has state-of-the-art security, including biometric access control to sensitive areas.
Rubric Evaluation and Verification for Items Scored Electronically (REVISE) software	A secure web-based application that selects and presents responses, gathers committee input, and updates our Item Tracking System (ITS) with the results.

Exhibit II-2: Glossary of Terms (continued)

Term	Glossary Description
Secure Browser	The only piece of software required for testing, can be installed quickly and simply; a special build of the Mozilla, or Firefox, browser that is modified to protect the security of the test and support certain accommodations; the student interface component of our test delivery system (TDS).
Superhelp	An AIR-designed online knowledge management tool designed to provide consistent responses to Help Desk calls and emails about the same issue by providing an electronic repository of FAQ documents, guides, manuals, and other resources. The use of this tool is a cornerstone in ensuring that agents provide up-to-date, accurate information to every user during every contact.
Test Administrator Interface (TA Interface)	The test administrator interface component of our test delivery system (TDS) used by test administrators to select the appropriate tests to administer, approve a student's or set of students' entry into a test session, set and approve accommodations (configurable), view student testing status, pause a student's test, or approve a print-on-demand request.
Test Information Distribution Engine (TIDE)	AIR's proprietary platform that provides an integrated system for gathering and managing student enrollment and pre-ID labels; adding, editing, and deleting users (e.g., district or school test coordinators) and granting them certain pre-defined authorizations based on a role hierarchy; monitoring testing progress; allowing role-based access to add, edit, or view student accommodations, interim assessment assignments and the creation or modification of class rosters.
Word	Microsoft Word

Section III: Responses to Requirements and Deliverables

Responses to RFP 2017-073 DOE New Hampshire Statewide Assessments IT Requirements Final BISSA Copy

Business Requirements

BUSINESS REQUIEMENTS					
State Requirements			Vendor		
Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
Appendix B					
	The vendor must address completely each Assessment Component bid on.	**M**	Yes	Standard	
	Each vendor must address Reporting Component E.	**M**	Yes	Standard	
B-1	**Submission Requiremnts**				
	Submission requirements • The proposal is date and time stamped before the deadline as defined in Section 2: Schedule of Events. • The vendor has sent the proper number of copies with the original version of the proposal marked "ORIGINAL" and the copies marked "COPY" as defined in Section 4.1: Proposal Submission, Deadline and Location Instructions. • The original proposal includes a signed Transmittal Letter accepting all terms and conditions of the RFP without exception.	**M**	Yes	Standard	
B-2	**Compliance with System Requirements**				
	System requirements and deliverables are listed in Appendix C: System Requirements and Deliverables in this RFP. The proposed vendor's solution must be able to satisfy all mandatory requirements listed.	**M**	Yes	Standard	
B-3	**Current Use of Vendor Proposed Software**				
	Components that constitute the vendor's proposed software suite must be fully implemented and operational in at least one (1) government entity comparable in size and complexity to the State. See Appendix E.	**M**	Yes	Standard	
B-4	**Project Components**				

	The vendor must have completed the vendor proposed software implementation for at least one (1) government clients comparable in size and complexity to the State within the last three (3) years. The specific vendor proposed software version and functionality must be described. See Appendix D and Appendix E.	**M**	Yes	Standard	
B-5	**Proposed Project Team**				
	The proposed Project Team must include individuals with substantial experience in statewide assessment design and administration, project management, data analysis and student privacy requirements. For the purpose of evaluating compliance with this requirement, the vendor team is permitted include subcontractors. In addition, one (1) team member may be identified to fulfill the experience requirement in multiple areas. See Appendix D and Appendix E.	**M**	Yes	Standard	
Appendix C	Vendors may bid on one or all of the components below. Each component must be fully addressed. Each component must include a reporting portal feature. Priority will be given to vendors who bid on all components in the project.		Yes	Standard	
	Submission Requiremnts				
Comp. A	A. Summative Assessments in ELA and Mathematics. The English language arts (to include AI scored writing component) and mathematics assessments will be administered annually in grades 3-8.	**M - if Bid On**	Yes	Standard	
Comp. B	B. Summative Assessments: Science. Science assessments will be administered annually in grades 5, 8, and 11.	**M - if Bid On**	Yes	Standard	
Comp. C	C. Interim Assessments: ELA and Mathematics. These online assessments are to be made available for local district use at least in the same grades included in the summative assessments described above. The purpose of the interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.	**M - if Bid On**	Yes	Standard	

Comp. D	D. Interim Assessments: Science. These online assessments are to be made available for local district use at least in the same grades included in the summative assessment described above. The purpose of the interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.	**M - if Bid On**	Yes	Standard	
Comp. E	E. Reporting Portal. The reporting portal must be capable of being upgraded across time. Initially, the system must be able to organize and present assessment data in a way that is easy for all users to read and understand. The Reporting Portal should allow for differentiated access to individual student, classroom, school, district and state assessment data, and they should integrate seamlessly with each other. Assessment results must be able to be provided to parents, among others, in a timely fashion, either on line or by printed report.	**M - if Bid On**	Yes	Standard	
	The vendor will be responsible for the development or procurement of all items included on the components of the NH statewide summative and interim assessments for ELA, mathematics, and/or science bid on.	**M**	Yes	Standard	
D-1	**Proposed Solution**		**Yes**	**Standard**	
D-1.1	**Test Construction**		**Yes**	**Standard**	
Topic 1 Test Design					
Topic 1 Test Design	Summative assessments will use a common-matrix design to support a) the reporting of student-level overall performance in terms of performance levels and scaled scores and b) the reporting of school- and district-level scores in a manner that reflects the depth and breadth of the academic standards.	**M**	Yes	Standard	
	Interim assessment frameworks should be based on the frameworks developed for the summative assessments.	**P**	Yes	Standard	
	Test blueprints must specify the numbers of each type assessment item to be used at each grade level in each content area, depth of knowledge, the numbers of items in each eventual test form (field and operational) and the total amount of testing time.	**M**	Yes	Standard	

	Reading assessments may include a combination of literary and informational passages. Literary texts include fiction, literary nonfiction and poetry. Informational texts include exposition, argumentation, persuasive texts, procedural texts and documents.	**P**	Yes	Standard	
	Writing assessments may include a combination of narrative, informative and persuasive purposes.	**P**	Yes	Standard	
	The vendor must provide its plan for ensuring a high level of consistency between the summative and interim item types and blueprints with adjustments made due to time and item type constraints.	**P**	Yes	Standard	
Topic 1.1 Stnd. Revision	The vendor must include in proposals a plan for assessment revision aligned to updated standards in 2018-2019.	**M**	Yes	Standard	
	All test design items must be updated each year to align to NH academic standards in ELA, mathematics, and science. Updated test blue prints aligned to items must accompany test designs and be approved by the NH DOE.	**M**	Yes	Standard	
Topic 1.2 Test Admin.	The summative assessment should be designed to take approximately two hours of testing time, per content area, for the vast majority of students. The test will be loosely timed, and allow for the accommodation of additional testing time into the schedule.	**P**	Yes	Standard	
	Each summative assessment component should be structured to be divided into two sessions. Both test sessions should also be available to schools to administer in one day if they chose to do so.	**P**	Yes	Standard	
Topic 1.3 Student Reg.	The vendor shall be responsible for managing the student/organization registration process. This process shall include the registration of students in private out of school district placements.	**M**	Yes	Standard	

	The vendor will be able to accommodate any student in home education or private school environments that wish to participate in the assessment process by identifying their registration separate from public school student registrations for school accountability purposes.	**M**	Yes	Standard	
	After the window for registration is complete, the NH DOE shall approve the registration counts. The vendor must allow the NH DOE the opportunity to review and amend registration information.	**M**	Yes	Standard	
	The vendor's response should include a recommended timeline for the registration process (relative to the test administration window) to ensure that accurate information is captured, appropriate quality checks occur, and to allow the NH DOE a sufficient window of time to review and amend the registration information.	**M**	Yes	Standard	
	The vendor's response should include a plan for allowing the registration of students enrolling in schools after the end of the test registration window and for students moving between schools during the testing window.	**M**	Yes	Standard	
Topic 1.4 Access & Fairness	The vendor's response must include a draft list of appropriate assessment accommodations separately for students with disabilities and English language learners. Accommodations listed must be supported by the most current research. The list shall describe the test accommodations and supports that allow access for students with disabilities and English learners to most fully participate in each assessment without interfering with the measurement of the constructs. Vendors shall also discuss accommodations which would threaten the validity of the assessment by interfering with the construct being measured.	**M**	Yes	Standard	
	The vendor's response must detail their plans for meeting accessibility requirements. The vendor's response should address how their proposed assessment system compares to the states' current systems and explain how it will address accessibility, accommodations, and fairness – while maintaining data privacy and security.	**M**	Yes	Standard	
	Vendors must include a description of how students with visual impairment will access on-line assessments or be provided with other accommodation, as appropriate.	**M**	Yes	Standard	

	Vendors must include a description of how students unable to access online assessments will participate in the assessments.	**M**	Yes	Standard	
Topic 2 Item Development					
Topic 2 Item Devel.	Item Development applies to the development of items, whether used to tailor an off-the-shelf solution in response to this RFP or in a customized solution. Per NH RSA 193-C:3, III (d), "teachers shall be involved in designing and using the assessment system."	**M**	Yes	Standard	
	Whether custom-developed for the NH assessments or obtained from other sources, all items included on the assessments will be subject to the review processes described in the RFP.	**M**	Yes	Standard	
	The vendor shall describe general procedures for item development including the use of outsourced or free-lance item writers, the use of items previously developed by the vendor, and/or the procurement of developed items from other sources. The vendor's response should include a description of how all item writers are trained on the content of the academic standards, in general, and on any specific content criteria related to the assessment component.	**M**	Yes	Standard	
	The vendor's response must reflect familiarity with the academic standards as well as current best practices and recommendations regarding the assessment of student achievement in each content area bid on.	**M**	Yes	Standard	
	The vendor's response must reflect familiarity with computer-based testing and the use of a variety of item types, including technology enhanced items (TEI) to assess students' higher order cognitive skills as well as their knowledge of core ideas and concepts.	**M**	Yes	Standard	
	The vendor's response must discuss the procedures that will be used to ensure that all assessment components are accessible to all students. The vendor's response should address the use of Universal Design (UD), Universal Design for Learning (UDL), and the use of development protocols such as the Accessible Portable Item Profile (APIP). The vendor's response should also address how technology will be used to enhance accessibility.	**M**	Yes	Standard	

	The vendor's response must indicate how the security of items will be maintained throughout the development and item review process, including procedures that will be taken to ensure the secure transfer of items and item information to/from states during the development process.	**M**	Yes	Standard	
	If vendors believe that tasks not specified in this RFP are critical to the development of quality items and a testing program, they should identify and describe the significance of those tasks in their response.	**O**	Yes	Standard	
	The writing component of the English language arts assessment must include constructed response item(s) as it relates to a reading passage to be included in the students overall score. The writing component must not simply ask students to reply to a writing prompt.	**M**	Yes	Standard	
Topic 2.1 Item Dev. Team	The vendor will appoint an Item Development Team, as necessary, responsible for the development of items for the each Assessment Component bid on. The vendor's response will identify key personnel and describe the proposed composition of the item development team, including describing the responsibilities and time commitments of the proposed members.	**M**	Yes	Standard	
	The vendor's response will include a description of how the specific needs of students with disabilities and English language learners will be accounted for within the proposed item development team. If the vendor proposes that a single person will fulfill multiple roles within the team (e.g., lead developer and grade level developer; developer at multiple grade levels) that must be clearly described in the vendor's response. The vendor's response must include a rationale to support the proposed composition of the item development team.	**M**	Yes	Standard	
	The vendor must include, as necessary, a description of the procedures, including the use of technology that will be used to facilitate interactions among the vendor's Item Development Team and the NH Instructional Support Team.	**M**	Yes	Standard	

	The vendor’s response must describe the type, number, and duration of in-person and virtual meetings between the item development team and the NH Instructional Support Team that will be needed throughout the development cycle to produce quality items for the tests.	**M**	Yes	Standard	
	All costs for proposed in-person development meetings (including travel costs and lodging) must be covered by the vendor.	**M**	Yes	Standard	
Topic 2.2 Item Dev. & Review	The vendor shall propose and describe the process that will be used to interact with the NH Instructional Support Teams throughout the item development and review process.	**M**	Yes	Standard	
	The vendor’s process for item development and review should reflect an understanding of the responsibilities of the NH DOE staff and NH educators and propose a process that avoids unnecessary travel, makes the most efficient use of their time, and allows sufficient turnaround time for review and approval of all items and related materials.	**P**	Yes	Standard	
	The vendor's plan will describe the type and number of in-person and virtual meetings that will be held during an annual development cycle.	**M**	Yes	Standard	
	The vendor’s plan should include a proposal for an initial in-person meeting between the NH Instructional Support Team for each assessment area (ELA, mathematics, science) and the vendor’s Item Development Team at the beginning of the project.	**M**	Yes	Standard	
	The vendor will support grade-level item review committees (3-8 ELA; 3-8 mathematics; 5, 8, and 11 science) for each of the assessments proposed. The NH DOE will determine the composition of committees.	**M**	Yes	Standard	
Topic 2.3 Item Review Com.	The vendor will support annual meetings of the Item Review Committees that will be jointly facilitated by the NH Instructional Support Team and the vendor’s Item Development Team.	**M**	Yes	Standard	

	The NH DOE is proposing an annual summer meeting due to the availability of educators and the timing of the meeting within the development cycle. If the vendor thinks that additional meetings of the Item Review Committee would be necessary and/or that the timing of the meetings should be changed, the vendor's response should include a proposal for an alternative meeting schedule. In particular, the vendor should indicate if they think that additional meetings will be needed during the initial year of the contract in preparation for the spring 2018 assessments.	**P**	Yes	Standard	
	The vendor's response should indicate the steps that will be taken to maximize efficiency throughout the item review process and, in particular, should describe how the vendor will make the most efficient use of the limited time available with the members of the Item Review Committees.	**P**	Yes	Standard	
	In preparing a response, the vendor should plan on supporting the item review committee meeting with the following specifications: • Each grade level committee will consist of 3-6 members. • Committee members will be paid a stipend of $150 per day for participation in the summer meetings. (If the vendor proposes meetings during the school year, the stipend will be replaced by a corresponding payment to the committee members' school district for substitute reimbursement). • Representatives from NH will also attend the meeting, including the NH Instructional Support Teams (up to a maximum of three (3) people total). • The vendor will support and arrange for lodging for committee members and NH representatives. The vendor's response should presume that all participants will require 3 nights lodging for the meeting (beginning one night prior to the meeting). • The vendor will also be responsible for travel expenses (e.g., mileage, airfare) for all participants to attend any out of state meetings. • The vendor will provide breakfast and lunch each day of the meeting and be responsible for dinner expenses ($25 per day) on days which require an overnight stay.	**P**	Yes	Standard	

	Meetings will be held at a hotel, conference center, or similar suitable location in NH. The meeting location may vary around the state or may be held in a central location to minimize travel requirements and expenses for committee members. The vendor will work with the NH DOE to develop a timeline to ensure that the NH Instructional Support Teams have sufficient time to review and provide feedback on all materials and items prepared for the Item Review Committee meetings.	**P**	Yes	Standard	
	The vendor will schedule an additional meeting day following the conclusion of the Item Review Committee meeting for a meeting of the vendor's Item Development Team and the appropriate NH Instructional Support Team to reconcile item feedback.	**P**	Yes	Standard	
	The vendor will produce a written report documenting the meeting and recommendations within two weeks of each committee meeting.	**M**	Yes	Standard	
Topic 2.4 Bias Sensitivity Review Com.	The vendor will support a Bias/Sensitivity Review Committee consisting of external educators and experts recruited and selected by the NH DOE to review the content of passages, other stimuli, and test items for potential bias and sensitivity. The NH DOE will determine the composition of the committee.	**M**	Yes	Standard	
	The Bias/Sensitivity Review Committee meetings will be facilitated by the vendor in coordination with a representative from the NH DOE. The committee will meet in NH annually during the summer. The vendor may propose additional meetings, if necessary during the initial year of the project.	**M**	Yes	Standard	

	In preparing a response, the vendor should plan on supporting the bias/sensitivity review committee meeting with the following specifications: The committee will consist of five (5) members. • Committee members will be paid a stipend of $150 per day plus travel expenses for participation in the summer meeting. (If additional meetings are proposed during the year, the $150 stipend will be replaced by a corresponding payment to districts for substitute reimbursement for any committee members who are employed by local education agencies.) • The meeting will also be attended by a representative of the NH DOE (1 person). • The vendor will support and arrange for lodging for committee members and the state's representative attending the meeting. • The vendor will provide breakfast and lunch each day of the meeting and be responsible for dinner expenses ($25 per day) on days which require an overnight stay.	**P**	Yes	Standard	
	The Bias/Sensitivity Review Committee will focus on review of stimuli proposed for the development of new field test items, review of newly developed items recommended for field test, and beginning in 2018 after the initial field test, review of items recommended for inclusion in the operational item bank that have been flagged for Differential Item Functioning (DIF) analysis.	**M**	Yes	Standard	
	If feasible, the Bias/Sensitivity Review Committee meetings may be scheduled concurrently with the Item Review Committee. The vendor's response should propose a process that will help avoid the costs and lack of efficiency of having assessment items go through development and be flagged for bias/sensitivity (content review, not empirical DIF analyses) only after substantial investment in development effort.	**O**	Yes	Standard	
	The vendor will produce a written report documenting the committee meeting and recommendations within two weeks of each meeting.	**M**	Yes	Standard	

Topic 2.5 Content Review	The vendor must ensure that the content of all items recommended for field-testing is accurate and reflects the current state of knowledge in the appropriate field. The vendor's response must describe their methods and procedures for meeting this requirement within the item development process.	**M**	Yes	Standard	
	The vendor shall work in collaboration with the NH DOE to ensure all assessments adhere to current and future standards for ELA, mathematics and science. The State has begun a process for review and revision of the ELA, mathematics and science and expects to field test in the 2018-2019 school year.	**M**	Yes	Standard	
Topic 2.6 Item Types & Number of Items	All items on the ELA, mathematics and science assessments must be machine/AI-scorable.	**M**	Yes	Standard	
	Machine/AI-scorable items, however, must not limit the use of item types to traditional multiple-choice or selected-response items.	**M**	Yes	Standard	
	The vendor's response must describe the variety of item types that could be included on the assessment components bid on, including item types that require students to generate or produce a response as well as select a response. The vendor's response must include a description of the vendor's experience with each of the item types proposed and provide access to sample items to allow the NH DOE to review proposed item formats.	**M**	Yes	Standard	
Topic 2.7 Number of Items	The vendor must include a proposed plan for the number of items of various types that will need to be developed for the Spring 2018 administration and subsequent operational test administrations.	**M**	Yes	Standard	
	The vendor should address the number of items that will be administered to an individual student as well as the total number of items that will be administered across matrix-sampled forms on each assessment component.	**M**	Yes	Standard	

	The vendor's response should reflect an understanding of the NH DOE's intended use of the results from each Assessment Components, the type of scores that will be reported, the plan to release items, as well as an understanding of academic standards.	**P**	Yes	Standard	
	The vendor's response should also reflect an awareness of the testing time and cost constraints discussed throughout this RFP.	**P**	Yes	Standard	
Topic 2.8 Item Release	The vendor will propose a plan for the annual release of a representative sample of test items.	**M**	Yes	Standard	
	The annual release of items will consist of up to 25 percent of the items (points) on a single student test form. The vendor's response should address how common items in a common-matrix design may be used to support the release of items.	**P**	Yes	Standard	
	Release of items will begin with the first operational administration of the summative assessment. All item types should be represented in the release of items. The balance of item types should be proportional to their use on the assessment. Items will be released in a digital format that enables local educators to interact with the items in the same manner that the items would be encountered on the assessment. Released items will be accompanied by supporting materials including relevant item statistics, information about the knowledge and skills assessed by the item, information on how the item was scored, and information regarding correct and common incorrect responses to the item.	**M**	Yes	Standard	

Topic 2.9 Rotation of Common Items	The vendor will propose a plan for the rotation of common items across years. The plan should address issues related to security, item exposure, maintaining content balance, and stability of assessment forms across years. At a minimum, the proposed plan should address: ● The number (or percentage) of items that should remain in place for consecutive years. ● The number (or percentage) of items that should be replaced after each test administration (including released items). ● The number of years before the common items on an assessment are totally refreshed. ● The maximum number of years, if any, that an item should be included in the operational test bank.	**M**	Yes	Standard	
	The NH Instructional Support Teams and Management Teams should have appropriate access to generate reports and/or view items and item information, as needed. The vendor’s response will include a description of its existing software to meet this requirement or describe plans to develop or procure the software necessary to meet this requirement.	**M**	Yes	Standard	
	The vendor’s response will describe steps that will be taken to ensure the security of the items.	**M**	Yes	Standard	
	The vendor’s response will describe how items and item information developed for the NH assessments through contracts awarded by the NH DOE will be accessible by the NH DOE at the conclusion or termination of those contracts.	**M**	Yes	Standard	

Topic 2.9.2 Item Use	The vendor will describe how the interests of the NH DOE will be assured and protected if items from other sources are included on the assessments. In particular, the vendor's response will describe: ● How items eligible for use on the NH assessments will remain secure, including any procedures in place to ensure that items are not released by other assessment programs or used for any other non-secure purposes. ● How license agreements will be structured to ensure that items may be used on the NH assessments for multiple administrations. ● The vendor's experience in handling any restrictions that may be placed on the use of items from other sources that would negatively impact the NH DOE.	**M**	Yes	Standard	
D-1.2	**Solution Technology**		**Yes**	**Standard**	
Topic 3 Technology Requirements					
Topic 3 Tech. Req.	The vendor will provide the test delivery platform, hosting site, test administration application, server, and application management services for the NH summative and interim assessments.	**M**	Yes	Standard	
	The vendor will be responsible for the maintenance of the full system; including code updates and/or patches, technical support, hosting, management, coordination, and support for customer-facing administration activities.	**M**	Yes	Standard	

	The vendor's response must provide a full description of its proposed computer-based test administration solution. The vendor's response should address each of the following: 1. Requirements for the use of any software (and supporting devices) should be clearly documented and explained. 2. The minimum and preferred technology infrastructure needed to support online testing should be documented and explained. 3. The technical support documents should include information about suggested computer lab configurations. 4. Information on computer-based assistive technologies should be provided to the client so that the client can determine which they may allow; data on use of these technologies should be collected. 5. Practice and training tests should be provided to allow students to become familiar with keyboarding and navigation techniques and tools that will be used during the live assessment.	**M**	Yes	Standard	

	6. Procedures for uploading student demographic data in the online assessment system, including any necessary accessibility tools and supports, should be provided, as well as instructions and procedures for modification of enrollment data, where permitted by the client. 7. Procedures for maintaining the security of the online testing environment should be documented. 8. Descriptions of training protocols to be provided at the local level on the test administration procedures should be provided. 9. In the first two years of the program, the vendor will be responsible for providing up to four (4) one-half day regional trainings on system use and test administration procedures, to be supplemented by an on-line webinar and other online training materials (e.g., slide deck from webinar, FAQ document). In subsequent years, in-person training sessions may be replaced by a series of webinars. 10. Technical support should be available via telephone and electronically with tools such as help desk and/or email. (see additional details in the Support Center section below). 11. Metrics for monitoring and documenting systems performance should be identified and described. 12. Documentation should be provided regarding the capacity of the system to support the current and potential future range of item types. Provide documentation regarding the application's capacity to import and export as applicable: items, student item response data, student registration, demographics, and data regarding eligible and utilized accommodations.	**M**	Yes	Standard	
Topic 4 Assessment Delivery Platform					

Topic 4 Asmnt. Delivery Platform	The vendor will ensure that the assessment delivery platform provides the technical infrastructure necessary to manage and administer assessments across the state. The vendor's response will address each of the following subcomponents and functionalities: • Administrative portal; • Test registration and scheduling; • Test administration (administrator interface); • Test delivery (student interface); • Test client; • Key-based and rule-based machine scoring; • Assessment delivery data storage; and • Student toolset (e.g., virtual calculators, protractor, ruler, notepad, highlighter).	**M**	Yes	Standard	
	The vendor's response will also address: • Authentication/User Identity Management: internal user management, user authentication, role-based authorization.	**M**	Yes	Standard	
	The vendor's response will also address: • Logging and Audit: A centralized capability for logging, log analysis and audit support, capturing and recording all system and testing activities at sufficient detail to detect conformance and compliance issues, and track errors. Logging is also used to capture data for analytics and secondary analyses.	**M**	Yes	Standard	
	The vendor's response will also address: • System Monitoring and Alerting: A centralized system for monitoring all processes and systems (network, hardware, software) in the assessment system and sending alert notifications whenever behavior fall outside of nominal ranges. Also, a system for monitoring and alerting support system data and test security.	**M**	Yes	Standard	

	The vendor's response will also address: • Common ID system: A centralized system for assigning and managing persistent, unique identifiers to all users (educators and students) of the NH assessment system. The purpose of this service is to assure the integrity of student data, including to avoid multiple creation of the same ID number, and to prevent the mismatching of students to assessment results.	**M**	Yes	Standard	
	The vendor will provide a detailed description of the interfaces and the System components used for processing.	**M**	Yes	Standard	
	The vendor will describe the software platform that the system operates on (code base, database, etc.). Note any third party platform components. Indicate the need for the State to purchase licenses.	**M**	Yes	Standard	
	The vendor will include details of the proposed software Solution including the database management system, licensed software suggested for data retrieval and reporting, proposed approach to developing any custom-built software components.	**M**	Yes	Standard	
	The vendor will discuss plans for anticipated future release of System software. Address any impact on System users or interfaces.	**M**	Yes	Standard	
Topic 5 Data Exchange and Process					
Topic 5 Data Exchange & Process	The vendor will provide a detailed description of the mechanism and tools included in the proposed System to enable the specified data sharing between the vendor and the State. • Identify the type of interface/mechanism/tool and the frequency of data exchange between the State and the vendor with a full explanation of the processes involved in the exchange. • Identify the format of the data the vendor will provide.	**M**	Yes	Standard	

	The vendor will provide a detailed description of how State staff will track status of the data submissions and follow to view a record of: • Administrative actions: Login, Logout, Password reset, IP address, batch file transmission; • Data submission including user, date, time, and IP address; • Users viewing validations by date, time, file, IP address; • Users viewing completeness measures by date, time, file, IP address; • Users making corrections by date, time, file, record, element, and IP address.	**M**	Yes	Standard	
	The vendor will provide a detailed description of the process the State will follow to submit special requests for research. The vendor may wish to include a sample scenario with the format of returned results.	**M**	Yes	Standard	
	The vendor will describe and provide samples of the available documentation supporting the system and the asset verification service.	**M**	Yes	Standard	
Topic 6 Data Privacy and Security					
Topic 6 Data Privacy & Security	The vendor will be expected to comply with Federal laws data privacy and security that include how data are accessed, stored, and exchanged; and how the vendor's employees are managed and trained regarding data security protocols.	**M**	Yes	Standard	
	The vendor's response must detail their privacy and security plans. The vendor's response should address how the proposed solution and associated activities will employ security protocols and design features to meet the states' rigorous security needs for data encryption, identity management, data access, and redundant layers of data protection.	**M**	Yes	Standard	

Topic 6.1 Info Tech. Stnds.	The vendor will provide a detailed description of how the proposed solution complies with established information technology standards. The proposed solution must comply with Open Standards and Open Data Formats as mandated by RSA 21-R (HB418 2012). Include the following: • A description of the degree to which the solution complies with information technology standards on the State of NH web site: http://www.nh.gov/doit/internet/vendors.php • A description of Open Source Software • A description of Open Data Format? • A statement of compliance with privacy and confidentiality standards including HIPAA, NIST, FERPA.	**M**	Yes	Standard	
Topic 7 Technical Compatibility					
Topic 7 Tech Compatibility	Solutions need to provide optimal performance in high-technology capability settings that have current generation computers and large bandwidth networks, but that still function without sacrificing performance in low-technology capability settings. This core principle includes a “device agnostic” approach to assessment content and assessment technology development.	**M**	Yes	Standard	
	All assessment components must be designed to function comparably across a range of devices using commonly deployed web browsers, including desktops, laptops, netbooks, and tablets (9.5” or larger) running Windows, Mac, Linux, Apple iOS, Android, and Chrome operating systems.	**M**	Yes	Standard	
	The vendor’s response must describe how it will ensure and verify that its system functions comparably across a range of devices. The vendor’s response should address how it will ensure that the system is not impacted by upgrades or other changes to devices or operating systems.	**M**	Yes	Standard	

Topic 7.1 Interop.	The vendor's response must detail their plans for ensuring interoperability. The vendor's response should address its compliance with industry-recognized, open-licensed interoperability standards and the processes and procedures used to verify and validate interoperability.	**M**	Yes	Standard	
D-1.3	**Security and Protection of Data**		**Yes**	**Standard**	
Topic 8 Security and Forensics					
Topic 8.1 Test Security	The vendor's response should address the following areas in general test security: • Develop and implement a comprehensive plan to ensure the security of test items, materials, and student data throughout the assessment cycle. • Develop and implement training procedures and materials regarding test security, and confidentiality of student data and personally identifiable information • Develop protocols for the secure collection, storage and destruction of secure and confidential teacher and student information. • Develop and implement uniform policies and procedures for identifying and dealing with possible security breaches and testing irregularities.	**M**	Yes	Standard	
	The vendor's response should address the following areas in general test security: • Develop implement procedures to account for and protect secure materials at all stages of distribution, receipt, storage, and return. Note: This requirement has general implications, but applies specifically to paper-based test forms. • Chain of Custody for materials shipped or transported: Develop and implement policies, guidelines and sign-off procedures for State, District, and School officials to establish and document a chain of custody for hand-offs to ensure that documents are received, accounted for, and distributed and returned. • Provide a secure architecture to protect the development and administration environment from network-based attacks.	**M**	Yes	Standard	

Topic 8.2 Data Forensics	The vendor will apply procedures to monitor, detect, and evaluate the assessments for potential cheating, and provide documentation to the NH DOE. The vendor's response should describe plans and procedures to provide continuous updates that capture a variety of data including but not limited to: • Time of testing, • All student answer choices including the final choice used for scoring; • Response latency; • Tracking the movement of the examinee through the test; • Student response times; • Accessibility options used by the student; and analysis of student gains over time; and • Differential performance on common and matrix-sampled items, if applicable.	**M**	Yes	Standard	
Topic 8.3 Test Mon.	The vendor shall describe in detail the steps that it would take to monitor the fidelity with which the test administration and security procedures are being applied.	**M**	Yes	Standard	
Topic 9 System Security					
Topic 9 System Security	The vendor shall provide a detailed description of the security design and architectural features incorporated into the proposed system.	**M**	Yes	Standard	
	The vendor shall describe the practices employed to ensure that your system and staff comply with FERPA regulations.	**M**	Yes	Standard	
	The vendor shall describe the system assurance provisions incorporated into the proposed system. At a minimum, discuss the following: a) What process or methodology is employed within the proposed system to ensure data integrity? b) To what degree does the approach rely on system assurance capabilities of the relational database management system (RDMS)? c) If multiple databases are employed, what extra procedures are employed to ensure synchronization among databases?	**M**	Yes	Standard	

	security testing: a) The identification and authentication methods used to ensure that users and any interfacing applications are identified and that their identities are properly verified. b) The authorization methods used to ensure that users and client applications can only access data and services for which they have been properly authorized. c) The immunity methods used to ensure that unauthorized malicious programs (e.g., viruses, worms and Trojan horses) do not infect the application.	**M**	Yes	Standard	
	The vendor shall discuss the company's practices pertaining to the following security testing: d) The methods used to ensure that communications and data integrity are not intentionally corrupted via unauthorized creation, modification or deletion. e) The methods used to ensure that the parties to interactions with the application cannot later repudiate or rebut those interactions. f) The intrusion detection methods used to ensure the detection, recording and review of attempted access or modification by unauthorized individuals.		Yes	Standard	
	The vendor shall discuss the company's practices pertaining to the following security testing: g) The privacy methods used to ensure that confidential data and sensitive communications are kept private. h) The system maintenance methods used to ensure that unauthorized system maintenance does not unintentionally disrupt the security mechanisms of the application or supporting hardware. i) The testing methods conducted to load and stress test your system to determine its ability to withstand Denial of Service (DoS) attacks.		Yes	Standard	

	The vendor shall discuss the company's practices pertaining to the following security testing: j) Your software patch schedule employed to protect the software from new security vulnerabilities as they arise. k) The ability of your system's software to be installed in a "locked-down" fashion so as to turn off unnecessary features (user accounts, operating system services, etc.) thereby reducing the software's security vulnerabilities and attack surfaces available to system hackers and attackers.		Yes	Standard	
Topic 10 Backup and Reporting					
Topic 10 Backup& Reporting	The vendor will: • Describe the tools used for backup and recovery of applications and data. • Describe the impact of the proposed backup process on the operation of the System.	**M**	Yes	Standard	
	The vendor will address the following: • Use of and method for logging and journalizing; • Single points of failure and recommended approaches for their elimination; and • Approach to redundancy.	**M**	Yes	Standard	
	The vendor will describe options to have the collected data stored at the vendor's site in addition to sending results along to the State. Include a proposed retention schedule.	**M**	Yes	Standard	
Topic 11 Assurance of Business Continuity					
Topic 11 Assurance of Business Continuity	The vendor shall provide a detailed description of the business continuity plan that mitigates risk to the State.	**M**	Yes	Standard	
	The vendor shall provide information on business continuity plans in the event that the hosting site becomes unavailable.	**M**	Yes	Standard	
	The vendor shall discuss plans for moving operations to a remote site if the hosting site is incapacitated.	**M**	Yes	Standard	

	The vendor shall discuss recovery time objectives and how the company will continue to meet federally required response metrics.	**M**	Yes	Standard	
	The State believes that additional software license fees solely related to redundancy for assurance of business continuity would be inappropriate. If the proposal differs from this standard, describe and provide rationale for the difference.	**O**	Yes	Standard	
D-1.4	**Training and Support**		**Yes**	**Standard**	
Topic 12 User Manuals and Guides					
Topic 12 Training & Support	The vendor will be responsible for providing the training and support required to ensure the administration of the NH assessments, including maintaining a support center to provide quality customer service and support to districts and schools throughout the registration, testing, and reporting cycles.	**M**	Yes	Standard	
	The vendor will develop test coordinator and test administrator manuals to ensure effective administration of the NH assessments.	**M**	Yes	Standard	
Topic 12 User Manuals & Guides	The vendor shall develop and produce an Online User's Guide. The guide shall provide technical specifications for use of the online platform used for testing. Information shall include but not be limited to: hardware specifications, proctor caching requirements if needed, student data upload process, data editing information, detailed information on the use of the assessment tools, and other technical guidelines as necessary. Thumbnail art shall be included as much as possible. Separate guides may be provided with focuses for technical and assessment staff. The guide shall be provided in PDF format for posting to NH DOE and vendor websites and in Word for use by the NH DOE in creating training and informational materials.	**M**	Yes	Standard	

Topic 12.1 Test Coord. Manual	The Test Coordinator Manual will focus on the tasks that must be completed at the district and school level, including scheduling, meeting technology requirements, student registration, accessibility, maintaining security, and the training of Test Administrators on test administration policies and procedures as well as security policies and protocols.	**M**	Yes	Standard	
Topic 12.2 Test Admn. Manual	The vendor shall develop and produce a TAM for each assessment per administration. The Test Administration Manual (TAM) will provide all directions needed by the test administrator to prepare for and administer the assessments, including security procedures.	**M**	Yes	Standard	
	The TAM shall contain general instructions for administering the assessment including, but not limited to, information about checking materials, planning testing schedules, organizing classrooms, preparation of students, use of standardized testing procedures, administering practice activities, security of materials, completing the student demographic portion of the response documents, accessibility instructions, test directions, administration of the test, assembly of materials for scanning and processing, checklists for class, school and district level procedures and information for returning materials. Thumbnail to full scale images of documents, forms, and other ancillary materials as needed with illustrations and explanatory diagrams shall be used extensively. The TAM will include scripts necessary to administer the assessments, and procedures and scripts necessary for accommodated testing outside of the assessment delivery system.	**M**	Yes	Standard	
	The vendor shall review the TAM prior to each administration and revisions shall be made to reflect changes related to the program, State and/or federal guidelines. The vendor shall make these documents available in printed form and for downloading from the Internet on a secure site.	**M**	Yes	Standard	
Topic 13 Training Materials					

Topic 13 Training Materials	The vendor will provide training and training materials to support the efficient and secure handling of materials as well as standardized administration activities. All proposed training materials and activities will be subject to NH DOE approval.	**M**	Yes	Standard	
	The vendor must provide training and training materials for district/school assessment coordinators, test administrators and district/school technology coordinators. As appropriate, the training must include information about student registration procedures; administration protocols; security policies, protocols, and procedures; the assessment delivery system; and accessibility and accommodations policies and protocols. The vendor should design training modules to enhance efficiency across types of users.	**M**	Yes	Standard	
	All training materials will be provided in formats that will permit them to be accessed via the internet. Posted documents must be available for viewing and downloading and must be provided in ADA compliant format.	**M**	Yes	Standard	
	Training materials should for test administrators should include the opportunity to practice all steps necessary to administer the assessment, including experiencing the assessment from the student's perspective.	**M**	Yes	Standard	
	The vendor will be responsible for providing annually up to four (4) one-half day regional trainings throughout NH on system use and test administration procedures, to be supplemented by an on-line webinar and other online training materials (e.g., slide deck from webinar, FAQ document). In subsequent years, in-person training sessions may be replaced by a series of webinars.	**M**	Yes	Standard	
	The vendor's response must describe the type and amount of training that the vendor feels is necessary to ensure the administration of the NH assessments (Summative, Formative, Reporting). The vendor's response should address the type of training materials that will be used including narrated PowerPoint web presentations, WebEx or other similar webinar tool, or videos, in addition to hard copy documents.	**M**	Yes	Standard	

	The vendor’s response must propose recommended methods and procedures for ensuring that test coordinators, test administrators, and technology coordinators have accessed the relevant training materials, have participated in and completed the required training, and/or are certified to fulfill their responsibilities in administering the assessment.	**M**	Yes	Standard	
Topic 13.1 Teacher Directions	The vendor shall develop and produce teacher’s directions for each assessment per administration. The teacher’s directions shall contain specific instructions for the administration of each grade level and/or content area per assessment. The teacher’s directions shall include information related to test administration including but not limited to, test security, the timing of tests and/or subtests, and the number of items on each assessment part.	**M**	Yes	Standard	
	A script for the administration of each content area shall be included to ensure consistent and appropriate directions are given to students to begin the test. The teacher’s directions shall be reviewed prior to each administration and revisions shall be made to reflect changes related to the program. Directions shall be provided electronically.	**M**	Yes	Standard	
Topic 14 Practice Tests & Student Materials					
Topic 14 Practice Tests & Student Materials	The vendor shall provide practice tests for each of the Summative Assessment Components bid.	**M**	Yes	Standard	

	• The practice tests will include all item types and/or response formats that a student may encounter during testing. • The practice tests will include all support and accessibility features and functionalities that a student will have access to during testing. • The items on the practice tests will include a range of content, depth of knowledge, and rigor. • The practice test should require approximately 30 minutes, but no more than 45 minutes, for students to complete. • Student scores on each item should be provided to students at the conclusion of the practice test. • The practice tests will be updated, as needed, to incorporate new item types, response formats, or other assessment features and functionalities.	**M**	Yes	Standard	
	The vendor's response should describe written materials, online tutorials, or other supports that may be developed to ensure that students are prepared to function within the online testing environment.	**M**	Yes	Standard	
Topic 15 Software Implementation Training					
Topic 15 Software Implmnt. Training	The vendor will provide a detailed summary of proposed training approach to include: • Recommended training approach (instructor led vs. computer based) • Training evaluation tools • Training coordination • Description of training materials and plans for revision • Training timeline	**M**	Yes	Standard	
Topic 15.1 Training & Prof. Dev.	Training and support for the NH assessments shall be provided by the vendor to NH educators as needed for each assessment component. The vendor must include in its proposal a detailed plan of action and timeline that describe how and when each of the training and support tasks will be accomplished.	**M**	Yes	Standard	

Topic 15.2 Tech. Director Trng.	The vendor shall describe its training plan for district technology directors. This training may include training on the operation and features of the online assessment system. It may include training on the physical and electronic security of assessments, system requirements for implementing the online assessment and troubleshooting of technology issues at the school or district site. Training may include a visual as well as oral presentation and may include other types of interactive technology. The delivery mode for these activities must be identified. The vendor shall discuss the role NH DOE will have in previewing each training session and webinar.	**M**	Yes	Standard	
Topic 15.3 Asmnt. Admn. Training	The vendor shall describe its assessment administration training plans for district test coordinators and test administrators. This training may include how to sign up for the interim assessment program, as well as how to enroll students in the summative assessment. Training related to the actual test administration should also be discussed. Training may include a visual as well as oral presentation and may include other types of interactive technology. The delivery mode for these activities must be identified. The vendor shall discuss the role NH DOE will have in previewing each training session and webinar.	**M**	Yes	Standard	
Topic 15.4 Asmnt. Results Trng.	The vendor shall describe its plans for providing educators with tools to evaluate and analyze assessment results in order to make informed instructional and programming decisions. Training may include a visual as well as oral presentation and may include other types of interactive technology. The delivery mode for these activities must be identified. The vendor shall discuss the role NH DOE will have in previewing each training session and webinar.	**M**	Yes	Standard	
Topic 15.5 Online Training Support	The vendor shall describe its plan to provide training and customer support specific to online assessment. The description should include training with an easy to understand set of directions, including screenshots, for operating the online assessment software. The vendor may also include other beneficial training materials in its response such as e-learning modules and online tutorials for users.	**M**	Yes	Standard	
D-1.5	**Assessment Scoring, Analysis and Equating**		**Yes**	**Standard**	

Topic 16 Machine Scored Items					
Topic 16 Machine Scored Items	The vendor’s response must include a description of their experience scoring all item types proposed for use on the NH assessments as well as a detailed description of the methods that will be used to ensure and verify the accuracy of scores from each type of item.	**M**	Yes	Standard	
	The vendor’s response must include a description of type of information that will be collected and available to states related to scored student responses, particularly for items that require students to generate a response, make multiple selections, or have complex scoring algorithms.	**M**	Yes	Standard	
Topic 16.1 Automated Scoring of Student Generated Response	The vendor’s response should address the current and near-term feasibility of using automated scoring to score student-generated text responses of varying lengths (e.g., single word, 1-2 sentences, paragraph, and extended essay). In addition to issues related to technical quality and accuracy of scoring, the vendor’s response should address, if applicable, issues such as cost, development time required, testing time required, and impact on the breadth and depth of content coverage on the assessment.	**P**	Yes	Standard	
	The vendor must indicate in detail its experience in developing or using artificial intelligence (AI) software in scoring student responses. The description shall illustrate the vendor’s experience with using AI scoring for each of the proposed item types, as well as limitations to the use of the vendor’s artificial intelligence scoring engine for each of these item types. This includes past and current projects, the software used in each, the manner in which the vendor proposes to use its software for this assessment component, the issues that it anticipates in using its software in NH, as well as the areas in which it anticipates that its software will not be effective.	**M**	Yes	Standard	

	The vendor shall describe how its AI engine functions, including how it is trained in relationship to content. The vendor shall provide its projected plans, if any exist, for improving its AI scoring capacity, including a description of why the company believes that this is a realistic goal. This description shall include specific time frames and must be considered within the context of the projected online implementation schedule of each content area in the summative assessment component.	**M**	Yes	Standard	
	For all constructed response items it is proposing to bring to the NH Assessment System, the vendor shall present its current procedures for development and selection of training papers for scoring of constructed response items and training of the artificial intelligence scoring system.	**M**	Yes	Standard	
	The vendor shall also present its plan for development and selection of training papers for scoring of constructed response items and training of the artificial intelligence scoring system for items developed specifically for NH.	**M**	Yes	Standard	
	The vendor shall discuss the role, if any, of NH educators in validating the rubrics and scoring of the training papers should be discussed.	**O**	Yes	Standard	
Topic 17 Analysis & Psychometric Support					
Topic 17 Analysis & Psychometric Support	The vendor is responsible for designing and conducting all analyses necessary to report student, school, district, and state results from the NH assessments and for ensuring that the NH assessments meet standards of technical quality for high-quality state assessments.		Yes	Standard	
	The vendor is responsible for designing and conducting all analyses necessary to provide evidence that the assessment program meets relevant U.S. Department of Education Peer Review requirements.	**M**	Yes	Standard	
	The vendor shall describe how the different types of scores it is proposing, individual scale scores and subscores, will be produced and verified.	**M**	Yes	Standard	

	The vendor must include scores produced strictly on items which are computer-scoreable and scores produced based on a combination of the computer-scored and hand-scored items. The limitations in interpretation of both of these scores must be discussed.	**M**	Yes	Standard	
	The vendor shall also describe how the Interim Assessment reports will provide predictive information regarding expected performance on the summative assessment. The timeline for providing this information must be discussed.	**M**	Yes	Standard	
Topic 18 Caliberation and Scaling					
Topic 18 Calib & Scaling	The vendor will calibrate test items and develop a scale(s) for each of the NH assessments using appropriate item response theory model(s).	**M**	Yes	Standard	
	The vendor's response must propose a recommended model(s) for item calibration and scaling and provide a rationale for the recommendation that includes: • A discussion of the benefits/advantages and limitations of the proposed model(s); • Its appropriateness for the type of items that will be included on the NH assessments; • Its appropriateness based on anticipated initial student performance on items aligned to NH Academic Standards for ELA, Math and Science; and • Its appropriateness for the type of scores that will be reported from the NH assessments.	**M**	Yes	Standard	
	The vendor's response must identify the software that will be used to perform item calibration and scaling and include a description of the vendor's familiarity and experience with the software.	**M**	Yes	Standard	
	If the vendor is proposing the use of proprietary or open-source software, the vendor's response must include a description of the steps that will be taken to ensure and verify the accuracy and reliability of the software.	**O**	Yes	Standard	
Topic 18.1 Calib. Plan	The vendor's response must include a description of how items from the Spring 2018 Field Test will be calibrated and placed on a common scale.	**M**	Yes	Standard	

	The vendor's response should address how a matrix-sampling test design will impact and be accounted for in the calibration process for the Spring 2018 Field Test and future operational test administrations.	**M**	Yes	Standard	
	The vendor's response must include a description of how embedded field-test items on operational administrations of the NH assessments will be calibrated and placed on NH assessment scoring scales.	**M**	Yes	Standard	
	The vendor's response must demonstrate an understanding of the NH DOE's desire to take advantage of the matrix-sampling design to produce school-level results.	**M**	Yes	Standard	
	The vendor's response should address the feasibility of producing independent subscales for particular domains or dimensions within each of the NH assessments.	**M**	Yes	Standard	
	The vendor's response must include a description of the calibration process and/or established concordance table that will align the student results on the proposed NH Summative Assessments for ELA and mathematics with the current Smarter Balanced reporting scale.	**M**	Yes	Standard	
Topic 19 Equating					
Topic 19 Equating	The vendor will design and conduct all analyses required to equate the NH assessments from year to year.	**M**	Yes	Standard	
Topic 19.1 Equating Plan	The vendor's response must include a description of how it proposes to equate the NH assessments from year to year.	**M**	Yes	Standard	
	The vendor's response should address how a common-matrix sample test design will factor into the equating design and also indicate the equating model that will be used.	**M**	Yes	Standard	
Topic 19.2 Equating Verif.	The vendor will support an independent real-time review of the equating process, analyses, and results by independent vendor(s), identified by the NH DOE.	**M**	Yes	Standard	
	The vendor's response will include a description of the steps that will be taken to ensure the accuracy of equating results.	**M**	Yes	Standard	

	The vendor(s) will support this effort by providing the consultant(s) with the necessary data files and other materials in a timely manner during the equating process.	**M**	Yes	Standard	
Topic 19.3 Equating Report	The vendor will produce an annual report documenting the equating process and results. The report will be available for use by the NH DOE in evaluating and approving the results of the equating process prior to reporting.	**M**	Yes	Standard	
Topic 20 Assessment Evaluation					
Topic 20 Asmtn. Eval	The vendor will design and conduct all analyses required to evaluate the quality and performance of all items developed for and/or included on the NH assessments.	**M**	Yes	Standard	
	The vendor's response must include a description of item statistics that will be generated and other analyses that will conducted.	**M**	Yes	Standard	
	The vendor's response should address how the appropriateness of items for all students will be examined and how the use of matrix-sampling may impact item evaluation.	**P**	Yes	Standard	
Topic 20.1 Field-Test Item Eval.	The vendor's response must include a description of the processes that will be used to generate appropriate information to support the evaluation of field test items.	**M**	Yes	Standard	
Topic 20.2 Operational Test Item Eval.	The vendor's response must include a description of the processes that will be used to generate appropriate information to support the evaluation of items that will be used to generate student and school scores and items that will be used to equate tests from year to year.	**M**	Yes	Standard	
Topic 20.3 Test Const. Eval.	The vendor will conduct analyses and provide psychometric support necessary to support the construction of technically sound test forms that meet all of the purposes and intended uses of results from the NH assessments.	**M**	Yes	Standard	
	The vendor's response should address how it proposes to use item statistics and information from psychometric analyses to support test construction.	**M**	Yes	Standard	

	The vendor will design and conduct all analyses necessary to produce accurate results student, school, district, and state reports.	**M**	Yes	Standard	
Topic 20.4 Addtl. Analysis	The vendor shall describe its proposed procedures for assuring that the assessments will be scored in a reliable and valid manner. Reliable and valid scoring for subgroups must also be discussed.	**M**	Yes	Standard	
	The vendor shall describe its system's capabilities in identifying unusual responses such as those that indicate abuse or potential for student self-harm and in flagging those responses on Interim Assessments for the teacher to review, and on Summative Assessments to notify the NH DOE.	**M**	Yes	Standard	
D-1.6	**Reporting**		**Yes**	**Standard**	
Reporting	The vendor is responsible for the accurate and timely reporting of results of the NH assessments.	**M**	Yes	Standard	
	The vendor shall describe the process it has used or will use to develop the formatting of the reports.	**M**	Yes	Standard	
	The reporting system should be designed to complement instruction and to facilitate the use of assessment results to improve student achievement. Reports should reflect areas of strength as well as areas that need to be targeted for instruction.	**M**	Yes	Standard	
	The vendor shall describe the process it has used for collecting, monitoring and reporting longitudinal growth data and provide sample reports.	**P**	Yes	Standard	
	Results from the first operational administration will be reported following standard setting.	**M**	Yes	Standard	
	Complete results from subsequent operational administrations of the NH assessments will be reported in a timely fashion at the completion of testing.	**M**	Yes	Standard	
	The vendor's response must include a detailed description of the processes that will be used to ensure the production of accurate color reports at the student, school, district, and state levels, including information on quality assurance and quality control procedures that will be used to ensure and verify the accuracy of reported results.	**M**	Yes	Standard	

	The vendor's response must also include a plan for the design, review, approval, and production of color reports. The plan should describe how the vendor will interact with the NH DOE throughout the design process.		Yes	Standard	
	The vendor must discuss the availability of reports in languages other than English.		Yes	Standard	
	The vendor will describe and provide examples of the various report forms distributed to students, schools, and districts. It will contain supportive information related to interpreting the test results, including but not limited to: reporting categories assessed and definitions for technical assessment terms.		Yes	Standard	
	The comprehensive interpretive guides shall be developed for use by schools and districts and shall be posted on the vendor's information portal and available electronically for the NH DOE, schools and districts to download.		Yes	Standard	
	The vendor shall collaborate in the development of the interpretive guides to ensure accurate information related to the assessment design is clearly provided. From the comprehensive guide, a smaller Parent Brochure shall be developed containing information pertinent to student level reports. The parent brochure shall be translated into one or more languages for distribution to non-English background parents as requested by the NH DOE. The parent brochure shall be distributed with the printed test scores and posted on the vendor's information portal.		Yes	Standard	
	The specifications for the interpretive guides and Parent Brochure include but are not limited to: • Available in electronic format that is accessible via the Internet. • Include thumbnails and larger images of selected reports. • Interpretive guides are developed for the purpose of providing schools and districts with an understanding of the reports that are available. • Interpretive guides include training information for use of online reporting tools. • Parent Brochure: developed for the purpose of providing test awareness for parents and students, shall include training information for use of online reporting tools. • Parent Brochure: provided in print format on a 1:1 ratio to student reports per content area per assessment.	**M**	Yes	Standard	

Topic 21 Assessment Scores					
Topic 21 Asmnt. Scores	The vendor's response should propose and describe options for the type of additional student scores that could be supported by the proposed design of the assessment.	**M**	Yes	Standard	
	The NH DOE is interested in providing reporting measures that contain actionable information, such that teachers and parents can use results to connect students with targeted instructional and academic materials that meet and challenge the student's abilities, interests and learning objectives. Should the vendor propose to utilize the scores of a subcontractor that can help identify appropriate reading materials, those costs must be listed as an option.	**M**	Yes	Standard	
Topic 21.1 Aggregate School, District, & State Scores	In addition to aggregate student scores such as mean scaled score and the percentage of students performing at each achievement level, school and district reports should contain detailed information about performance on critical aspects of the NH academic standards.	**M**	Yes	Standard	
	The vendor's response should describe how matrix sampling will be used to produce reliable subscores that provide useful information and support valid inferences about school and district performance at one or more levels below overall achievement.	**M**	Yes	Standard	
	The vendor's response should also indicate whether such school and district scores will be reported on their own subscales which can be linked across years to allow comparisons in performance from one year to the next.	**M**	Yes	Standard	
Topic 22 Reports					
Topic 22 Reports	The vendor shall propose a process for the design of reports that includes participation of the NH DOE management team and the NH Content Teams.	**M**	Yes	Standard	
	The vendor's response shall describe the process and procedures that will be used to generate initial design specifications and concepts, to facilitate review and revision, and for the approval of report designs.	**M**	Yes	Standard	
	The vendor's response should address the feasibility of obtaining external feedback on proposed report designs.	**M**	Yes	Standard	

Topic 22.1 Types of Reports	The vendor will be responsible for producing a variety of reports intended for use by a variety of audiences.	**M**	Yes	Standard	
Topic 22.2 Student Reports	The vendor will produce hard copy student reports (one per student) that will be shipped directly to schools.	**M**	Yes	Standard	
	The vendor will also produce a printable, digital version of the student report that may be printed by the district or school.	**M**	Yes	Standard	
	Topic 22.3 School, District, and State Reports: The vendor will produce school-, district-, and state-level reports in printable, digital format.	**M**	Yes	Standard	
Topic 22.3 School, District, & State Reports	The vendor's response will propose a system for providing schools and districts with efficient and secure access to confidential and non-confidential reports.	**M**	Yes	Standard	

	Examples of the school-, district-, and state-level reports to be produced include: • Rosters providing individual student-level results at the school or classroom level (dependent upon data available). Rosters may also include item-level results for released items. • Summary aggregating results from the Roster at the school or classroom level. May include school, district, and state comparisons. • School Report Package containing information on school participation and performance including performance level results, use of accommodations, subgroup results as required by the USED and subscore results. The report may also include selected results from the released items, district and state comparisons, and comparisons with previous years. • District Report Package containing the same information as the school report aggregated at the district level. • State Report Package containing the same information as the school report aggregated at the state level. o School Summary Report providing summary participation and performance information across grade levels tested within the school. o District Summary Report providing the same information as the school summary report aggregated at the district level. o State Summary Report providing the same information as the School Summary Report aggregated at the state level.	**P**	Yes	Standard	
Topic 23 Data Files					
Topic 23 Data Files	the vendor will provide the information contained in all reports in a data file in an agreed upon format that can be imported into the NH DOE and schools' reporting systems.	**M**	Yes	Standard	
	To the extent possible, all reports should contain embedded information to support and promote the proper interpretation and use of the results provided on the report.	**P**	Yes	Standard	
	The vendor will also propose the development of supplemental materials to assist in the interpretation and use of NH assessment reports by the parents and students, local educators, and the public.	**M**	Yes	Standard	
	Interpretive materials will be developed in digital form for web-based delivery.	**M**	Yes	Standard	
	Vendors may propose options for printable text materials as well as materials in other media such as videos or interactive graphics.	**O**	Yes	Standard	

	The vendor will also support two in-person reporting workshops following the first two operational administrations of the NH assessments.	**M**	Yes	Standard	
D-1.7	**Standard Setting**		**Yes**	**Standard**	
Standard Setting	Student results from the NH assessments will be reported according to performance levels. Details on the number and names of performance levels will be determined during 2017-2018.	**M**	Yes	Standard	
	The vendor shall describe its standard setting methodology for summative and interim assessments. Use of empirical data, including summative data, as well as the any inclusion of NH educators in the process must be discussed.	**M**	Yes	Standard	
	The vendor’s plan to ensure continuity between grade levels must be described.	**M**	Yes	Standard	
	Cut scores indicating the level of student performance required to attain each performance level classification will be determined in the summer following the first operational administration of the NH assessments.	**M**	Yes	Standard	
	The vendor shall recommend methods of validating cut scores across time, including approaches to revising as evidence indicates.	**M**	Yes	Standard	
	The vendor shall suggest how to approach communication with the field regarding potentially changing cut scores.	**O**	Yes	Standard	
	The vendor will support the NH DOE in all activities related to establishing performance standards for the NH assessments.	**M**	Yes	Standard	
Topic 24 Performance Levels					
Topic 24 Perf. Levels	The vendor will be responsible for proposing, organizing, and supporting a process for supporting the NH DOE in determining the number and names of performance levels appropriate for the NH assessments. Such performance levels shall be set, at a minimum, to meet federal assessment reporting requirements.	**M**	Yes	Standard	
	The vendor will be responsible for proposing, organizing, and supporting a process for developing appropriate performance level descriptions for the NH assessments.	**M**	Yes	Standard	
	The vendor’s response will include a description of the processes that are proposed to determining the number and names of performance levels and then to develop appropriate performance level descriptions.	**M**	Yes	Standard	

	The vendor’s response should include a timeline of major activities and/or milestones in the process.	**M**	Yes	Standard	
	The vendor will be responsible for all costs associated with proposed meetings, including expenses, stipends and/or reimbursement costs for external panelists.	**M**	Yes	Standard	
	Vendors should use $150 per day per panelist for stipends/substitute reimbursement when budgeting for these meetings.	**P**	Yes	Standard	
Topic 25 Cut Scores					
Topic 25 Cut Scores	The vendor will be responsible for proposing, organizing, and supporting a process for determining performance level cut scores (i.e., thresholds) on each of the NH assessments.	**M**	Yes	Standard	
	The vendor’s response will propose a standard setting method that is appropriate for use with the NH assessments.	**M**	Yes	Standard	
	The vendor’s response will include a rationale for the use of the proposed method and will address how the method will be applied with the matrix-sampled design of the assessments.	**M**	Yes	Standard	
	The vendor’s response will include a description of the processes and procedures necessary to implement the proposed standard setting method.	**M**	Yes	Standard	
	The vendor will be responsible for all costs associated with standard setting meetings, including expenses, stipends and/or reimbursement costs for standard setting panelists.	**M**	Yes	Standard	
	Vendors should use $150 per day per panelist for stipends/substitute reimbursement when budgeting for standard setting meetings.	**P**	Yes	Standard	
	The vendor’s response should describe the role of the states in setting performance level cut scores before, during, and after any proposed standard setting panel meetings.	**M**	Yes	Standard	
Topic 26 Standard Setting Report					
Topic 26 Standard Setting Report	The vendor will prepare a report describing and documenting the entire Standard Setting Process. The report will be delivered in digital format no later than one month following the completion of the standard setting process.	**M**	Yes	Standard	

Topic 26.1 Standard Setting Validation	The vendor's response should include a plan for conducting analyses to validate the performance standards following the second operational administration of the NH assessments.	**M**	Yes	Standard	
D-1.8	**Reporting Portal**		**Yes**	**Standard**	
Reporting Portal	The NH Assessment Reporting Portal should provide a platform that seamlessly integrates data from state summative and interim assessments, providing stakeholders with a user-friendly interface that increases access to results.	**M**	Yes	Standard	
	The NH DOE prefers a system that would allow users to customize particular aspects of their individual dashboard profiles. The vendor should describe the features of its Reporting Portal, including the extent to which its system includes the preferred features.	**P**	Yes	Standard	
	The system must be designed to allow the state access to high level information and would ideally allow students and parents to access detailed information. In addition, educators, school administrators and district administrator roles must be included. For costing purposes included in the Pricing Model, state costs for making the system operable to the educator level should be provided.	**M**	Yes	Standard	
	Vendors may choose to provide a separate Pricing Model for a parent portal option and a separate Pricing Model for student portal option.	**O**	Yes	Standard	
	The optional parent portal should provide parents with a user- friendly platform that allows them to access their child's assessment results, as well as other classroom and school information. The vendor shall indicate whether or not its system includes features for parents. The vendor shall describe the parent- related features of its dashboard system, including the extent to which its system includes the preferred features. Pricing for the parent portal should be included separately as an option.	**O**	Yes	Standard	
	Vendors may describe how the student portal will allow for students to customize their individual pages. Pricing for the student portal should be included separately as an option.	**O**	Yes	Standard	
D-2	**Corporate Overview and Project Management**		**Yes**	**Standard**	

D-2.1	Corporate Qualifications		Yes	Standard	
Topic 27 Corporate Overview (Appendix E)					
Topic 27 Corp. Overview	E-1.1.1 Corporate Overview Identify the proposed role of the firm on the project. Describe the major business areas of the firm. Provide a high-level description of the firm's organization and staff size. Discuss the firm's commitment to the public sector, experience with this type of project implementation and experience in New Hampshire.	**M**	Yes	Standard	
	E-1.1.2 Financial Strength Provide at least one of the following: 1 The current Dunn & Bradstreet report on the firm; or 2 The firm's two most recent audited financial statements; and The firm's most recent un-audited, quarterly financial statement; or 3 The firm's most recent income tax return	**M**	Yes	Standard	
	E-1.1.3 Litigation Identify and describe any claims made by clients during the last ten (10) years. Discuss merits, current status and, if available, outcome of each matter.	**M**	Yes	Standard	
	E-1.1.4 Prior Project Descriptions Provide descriptions of no more than three (3) similar projects completed in the last three (3) years. Each project description should include: 1. An overview of the project covering type of client, objective, project scope, role of the firm and outcome; 2. Project measures including proposed cost, actual project cost, proposed project schedule and actual project schedule; 3. Names and contact information (name, title, address and current telephone number) for one or two references from the client; and 4. Names and project roles of individuals on the proposed team for the New Hampshire project that participated in the project described.B287	**M**	Yes	Standard	
	E-1.1.4.1 Components that constitute the vendor's proposed software suite must be fully implemented and operational in at least one (1) government entity comparable in size and complexity to the State of New Hampshire.	**M**	Yes	Standard	

	E-1.1.5 Subcontractor Information Vendors must provide information on any subcontractors proposed to work on this project. Required information shall include but not be limited to: 1. Identification of the proposed subcontractor and a description of the major business areas of the firm and their proposed role on the project; 2. A high-level description of the subcontractor's organization and staff size; 3. Discussion of the subcontractor's experience with this type of project; 4. Resumes of key personnel proposed to work on the project; and 5. Two references from companies or organizations where they performed similar services (if requested by the State).	**M**	Yes	Standard	
	E-2 Team Organization and Designation of Key Vendor Staff Provide resumes of key personnel proposed to work on the project and an organizational chart depicting the vendor project team. This chart should identify key staff required from the vendor, any subcontractors, and the State. Define the responsibilities and length of assignment for each of the roles depicted in the organizational chart. Identify the positions that should be designated key staff. Ensure that designation of key vendor staff includes subject matter experts in the following areas: A single team member may be identified to fulfill the experience requirement in multiple areas.	**M**	Yes	Standard	
	E-2.1 State Staff Resource Worksheet Append a completed State Staff Resource Worksheet to indicate resources expected of organization. Expected resources must not exceed those outlined in Section A 4.2.	**M**	Yes	Standard	

	E-3 Candidates for Project Manager Although the State recognizes that staff availability is somewhat uncertain, qualifications of the Project Manager are particularly critical. Therefore, the State requires that the Project Manager be identified with some degree of certainty. For the Project Manager candidate, provide a resume not to exceed five (5) pages in length addressing the following: • The candidate's educational background; • An overview of the candidate's work history; • The candidate's project experience, including project type, project role and duration of the assignment; • Any significant certifications held by or honors awarded to the candidate; and • At least three (3) references, with contact information, that can address the candidate's performance on past projects.	**M**	Yes	Standard	
	E-4 Candidates for key vendor staff Roles Provide a resume not to exceed 2 pages for each key vendor staff position on the project team. Each resume should address the following: • The individual's educational background; • An overview of the individual's work history; • The individual's project experience, including project type, project role and duration of the assignment; • Any significant certifications held by or honors awarded to the candidate; and • At least three (3) references, with contact information, that can address the individual's performance on past projects.	**M**	Yes	Standard	
Topic 28 Vendor Experience					
Topic 28 Vendor Exp.	The vendor shall provide the company's history, including the number of years that it has been in business, buyouts, takeovers, IPO's, bankruptcies, litigations and claims, etc. within the last five (5) years, or for that period which the firm has been in business, if less than five (5) years. Situations arising in assessed liquidated damages (LDs) must be described with their resolution, along with the amount of the LDs or provided additional services.	**M**	Yes	Standard	

	The description shall also identify the number of employees in the company and the company's location(s), including any presence in New Hampshire. The overall capacity of the vendor's organization(s) and the resources that it will commit to the work for the project (by name and role in project) shall be discussed.	**M**	Yes	Standard	
	The description shall also outline the vendor's overall position in the State assessment market, including the length of time, states served, addition/loss of states over the past five (5) years.	**M**	Yes	Standard	
	A general description of the vendor's capabilities and capacities related to development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities shall be included. Responses must demonstrate that the vendor meets, at a minimum, the mandatory qualifications presented at the beginning of this component.	**M**	Yes	Standard	
	Specific examples of the vendor's work products such as test and item specifications, items, forms, technical manuals, research reports, technical services, etc. should be identified under the relevant requirements and specifications and provided in attachments as appropriate. NH DOE expects to receive the same or better quality of work throughout the contract, including any extensions, as the examples that are provided in the proposal.	**M**	Yes	Standard	
Topic 28.1 Relevant Exp.	In tabular format, the vendor shall provide a listing and descriptions of all work in similar projects that it and its proposed subcontractors have carried out or are carrying out for other clients. The table shall include client, program name, content area, grades, administration mode (paper-pencil or computer-based), use of scoring (human and artificial intelligence), length of contract and number of students.	**M**	Yes	Standard	
	For computer-based testing, the vendor shall include the total number of tests administered and the highest number of on current testers. For each such project, the vendor must provide the name of the state or other organization, name of client contact person, this individual's telephone, email and fax numbers, and e-mail address	**M**	Yes	Standard	

	Current Use of Vendor Proposed Software – Current Implemented Sites of Vendor Proposed Software	**M**	Yes	Standard	
	Components that constitute the vendor's proposed software suite must be fully implemented and operational in at least one (1) government entity comparable in size and complexity to the State of New Hampshire.	**M**	Yes	Standard	
D-2.2	**Project Management**		**Yes**	**Standard**	
Topic 29 Project Management Team					
Topic 29 Mang. Team	The vendor shall provide a list of key staff, including but not limited to, the program manager, program coordinator, lead psychometrician, content development lead, content specific area lead, technology lead, special populations consultant, scoring manager(s), production manager(s), and publication staff, as well as all staff assigned 0.20 FTE or greater to each component. Each staff member's assigned responsibilities and time allocated to the project must be provided. Time expected to be allocated to other projects must also be indicated. In no case should an individual be assigned to more than one full-time equivalent position.	**M**	Yes	Standard	
	The vendor shall affirm in the response to this request for proposal that should the contract be awarded, all key personnel proposed shall be released from any concurrent responsibilities that would impede their availability to assume the work as proposed.	**M**	Yes	Standard	
	The vendor shall assign one person to function as the Program Manager. That person must be responsible for all activities required by the project and will serve as the main contact person between the vendor and NH DOE. The Program Manager shall have the authority to make decisions and commitments on behalf of the vendor, subject to NH DOE approval.	**M**	Yes	Standard	

	NH DOE requires that a Technology Consultant be identified. This individual shall be responsible for a number of tasks, including but not limited to, assisting NH DOE and the districts with the transition to assessment system; working with NH DOE information staff to ensure the accurate and efficient transfer of data to and from NH DOE; creating, defining and reviewing file layouts; providing assistance in the verification of demographic data; and assisting NH DOE and district assessment coordinators with the use of vendor websites and functionality.	**M**	Yes	Standard	
	NH DOE reserves the right to interview and approve all key staff including subcontractor staff. Throughout the life of this contract, and any extensions, changes to the assigned program manager, program coordinator, lead psychometrician, content development lead, content specific area lead, special populations consultant, and technology consultant, except for those resulting from separation of services, will require prior written consent by NH DOE. In the event that NH DOE requests removal of specific vendor personnel, the vendor shall provide acceptable replacement(s) with no impact to the project. Replacement(s) shall have qualifications which meet or exceed the original staff member proposed or the staff member holding the position previously and shall be approved by NH DOE.	**M**	Yes	Standard	
	All personnel who will work on-site at NH DOE or school sites may be required to be pre-approved for site access via a criminal background check paid for by the vendor.	**M**	Yes	Standard	
Topic 30 Staff Qualifications					
Topic 30 Staff Qual. & Exper.	Qualifications of all key personnel shall be presented in the vendor's proposal, including subcontractors. Supporting resumes outlining education/training, employment history, and experience in conducting work similar to what is expected under this contract shall be included as an appendix.	**M**	Yes	Standard	

	NH DOE requires a psychometric team that will not only execute routine functions, but will also be able to provide a sophisticated level of expertise to guide the psychometric decisions that will need to be made and re-evaluated as the program evolves. The expectation is that the team will be able to provide psychometric options with strengths and challenges and its recommendations along with rationale. In addition, especially in the event of unexpected challenges, the team must include someone with both extensive experience and psychometric knowledge, as well as the decision-making authority to quickly address and remedy the situation.	**M**	Yes	Standard	
	For all meetings involving educators, the vendor must indicate the qualifications of the facilitators. General qualifications for training and meeting facilitators must be included in the response to this request for proposals. Facilitators must be familiar with best practices, as well as state and federal laws, procedures and regulations concerning assessment. As applicable, facilitators must also be familiar with academic instruction of students and the educational and assessment landscape. Facilitators must be able to clearly articulate spoken English and create easily understood written materials and visual training aids. Facilitators must have demonstrated experience in leading large-group trainings including webinars and meetings as fit their responsibilities.	**M**	Yes	Standard	
	Organizational charts, including identification of Program Manager and key personnel, for the vendor as a whole and for the NH DOE project team specifically, including subcontractors where applicable, must be provided. The charts shall clearly indicate lines of authority and communication within and among the vendor’s departments and subcontractors, where appropriate.	**M**	Yes	Standard	
	The vendor shall also describe its escalation process for resolving any vendor/client disagreements.	**M**	Yes	Standard	

	The Program Manager directly in charge of overseeing the NH project shall be identified. This manager shall be available both during and outside of normal business hours to assist with any urgent situations. Contact information for this individual shall be provided at the time of contract award.	**M**	Yes	Standard	
	Changes to the assigned Program Manager, except for those resulting from separation of services, require prior written consent by NH DOE. The replacement shall have qualifications which meet or exceed the original staff member proposed or the staff member holding the position previously and shall be approved by NH DOE.	**M**	Yes	Standard	
	Qualifications of key executive personnel must be presented. A supporting resume outlining education/training, employment history, and experience in conducting work similar to what is expected under this contract shall be included as an appendix.	**M**	Yes	Standard	
	The vendor shall fulfill this requirement and all requirements listed in Appendix E and Appendix H.	**M**	Yes	Standard	
D-2.3	**Project Plan**		**Yes**	**Standard**	
	The vendor shall describe the planned project management activities as they pertain to the three phases, planning, Implementation, and operations. In addition to addressing the components listed in Appendix D2.3 Work Plan, the vendor shall provide an example of status reports prepared for another similar project. Names of the project and of any individuals involved may be removed.		Yes	Standard	
Topic 31 Work Plan					
Topic 31 Work Plan	proposal addressing all work offered in their proposal. The State will evaluate the proposed project Work Plan contained in the Proposal to determine how well it will serve the needs of State Project leaders.		Yes	Standard	

	The State sees a Work Plan as essential to reaching a comprehensive agreement with a vendor. Consequently, the State will seek to refine the proposed Work Plan during contract finalization with the selected vendor and to incorporate the refined Work Plan by reference into a contract. In addition, the State will require the selected vendor to update the Work Plan in consultation with the State during the term of the project.		Yes	Standard	
Topic 31.1 Prelim. Work Plan	Provide a preliminary Work Plan for the planning and implementation phases of the engagement. The vendor's preliminary proposed Work Plan includes a description of the schedule, tasks, deliverables (with pricing), major milestones, task dependencies, and a payment schedule. The Work Plan shall also address resource allocations (both State and vendor team members). Include sufficient detail that the State will be able to identify departures from the plan in sufficient time to seek corrective action. In particular provide information about staffing. Identify and discuss the following: • All assumptions upon which the work plan is based; • Descriptions of recommended roles by activity and time required for both State and vendor members of the project team; • Assignments of members of the vendor's team identified by role to specific tasks; and • Critical success factors for the project.	**M**	Yes	Standard	
	The vendor Work Plans should include information pertaining to resource allocation, update frequency, financial check points and a graphic overview.	**M**	Yes	Standard	
Topic 31.2 Project Plan and Schedule	Proposals shall include a detailed schedule reflective of the Work Plans that describe how each of the requirements and specifications described in the proposal will be accomplished. The schedule shall at a minimum identify the tasks, subtasks, beginning date, end date and the party/functional group responsible for each step in the process. The schedule must be included as a separate attachment to the proposal.	**M**	Yes	Standard	

	The proposed plans and schedule shall clearly identify and include: • Key activities related to the field (ordering of materials, receipt of materials, test dates, return of materials, demographic clean-up window, release of individual student scores, final individual student, school and district score file release, and receipt of paper reports) • Key transfer dates between the vendor and NH DOE related to development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities.	**M**	Yes	Standard	
	At the beginning of the project and by the beginning of each fiscal year, the vendor will develop a detailed project plan and schedule for the coming fiscal year.	**M**	Yes	Standard	
	Activities related to the development for the next year's assessment and reporting for the prior year's assessment must be clearly distinguishable from activities related to the current year's assessment.	**M**	Yes	Standard	
	Joint review of this schedule followed by NH DOE's approval for the first contract period should occur within two weeks of the contract award. The vendor and NH DOE shall mutually agree upon final dates. Joint monitoring of the schedule shall occur on an on-going basis. The vendor shall ensure that all schedule adjustments allow for final deliverable dates to be met. If necessary, timelines and schedules may be revised with prior approval of NH DOE and an executed contract amendment for all deliverables subject to liquidated damages.	**M**	Yes	Standard	
	A revision of a timeline on the part of the vendor exempts the vendor from meeting a contractual deadline only if (1) the vendor and NH DOE mutually agree upon and document through a contract amendment an extension of the deadline as executed through a contract amendment or (2) the vendor is able to prove that the deadline was not met due to NH DOE's failure to meet a contractual deadline resulting in the vendor's inability to adhere to the schedule for delivery of products and services.	**M**	Yes	Standard	
	The vendor shall alert NH DOE as soon as it believes a deliverable subject to liquidated damages is at risk of not meeting its delivery date.	**M**	Yes	Standard	

	NH DOE must be notified whenever the New Hampshire contract is included in vendor's internal meetings focused on programs at-risk.	**M**	Yes	Standard	
	For the contract beginning after July 1, 2017, the review of the schedule should occur within the first two weeks of the initial contract. For each following contract year, by May 1, the vendor shall provide an updated detailed Work Plan and project schedule that specifies all activities leading to products or services deliverable to either NH DOE or local school districts for the following assessment year.	**M**	Yes	Standard	
	The development of the project plan and schedule will follow a review of the current project status and contract specifications by the vendor and NH DOE. Any foreseeable changes to contract requirements and/or costs will be discussed and agreed upon during this process and reflected in the project plan and schedule.	**M**	Yes	Standard	
Topic 32 Manatement Meetings					
Topic 32 Mngt. Meetings	The vendor will be responsible for organizing and supporting regular management meetings with the NH DOE project management team. The vendor will be responsible for costs associated with management meetings.	**M**	Yes	Standard	
	An initial two-day, in-person management meeting will be held shortly after the contract is awarded. Participants will include key vendor staff and State project leaders. This meeting will enable leaders to become acquainted and establish any preliminary project procedures.	**M**	Yes	Standard	
	• Status Meetings: Participants will include project leaders from the vendor and the State. These meetings, which will be conducted at least twice monthly, will address overall project status and any additional topics needed to remain on schedule and within budget. A status report from the vendor will serve as the basis for discussion.	**M**	Yes	Standard	
	• Special Meetings: Need may arise for a special meeting with State leaders or project stakeholders to address specific issues.	**M**	Yes	Standard	

	• In Year 1 of the project, weekly phone calls between pertinent NH DOE staff and the vendor's Program Manager and other key vendor staff shall be held between in-person project meetings to keep NH DOE current on project status, discuss issues as they arise, and to plan upcoming activities.	**M**	Yes	Standard	
	• Exit Meeting: Participants will include Project leaders from the vendor and the State. Discussion will focus on lessons learned from the Project and on follow-up options that the State may wish to consider.	**M**	Yes	Standard	
	The vendor shall include the following when describing the meeting process: • Timing, duration, recommended participants and agenda for the kickoff meeting; • Frequency and standard agenda items for status meetings; • Availability for special meetings; and • Agenda for the exit meeting.	**M**	Yes	Standard	
	The State expects the vendor to prepare agendas and background for and minutes of meetings. Background for each status meeting must include an updated Work Plan. Drafting of formal presentations, such as a presentation for the kickoff meeting, will also be a vendor responsibility.	**M**	Yes	Standard	
	Vendor shall submit reports in accordance with the schedule and terms of the contract. All reports shall be prepared in formats approved by the State. The vendor's project manager shall produce reports related to project management as reasonably requested by the State. Vendor shall produce project status reports, which shall contain, at a minimum, the following: • Project status as it relates to Work Plan • Deliverables status • Accomplishments during weeks being reported • Planned activities for the upcoming two (2) week period • Future activities • Issues and concerns requiring resolution • Report and remedies in case of falling behind schedule	**M**	Yes	Standard	

Topic 33 Project Communication					
Topic 33 Project Comm.	The vendor will propose a communication plan to ensure effective communication among key project stakeholders.	**M**	Yes	Standard	
Topic 33.1 Ongoing Comm.	Communication between the vendor and NH DOE personnel is essential. Telephone calls, telephone conference calls, emails, overnight courier service, facsimile correspondence, and other communication procedures will be at the vendor's expense.	**M**	Yes	Standard	
	Toll-free numbers shall be provided by the vendor for telephone communication including conference calls and webinars.	**M**	Yes	Standard	
	The vendor shall make all written communication or summaries of communications with any subcontractor(s) identified in this proposal available to NH DOE at its request. In addition, NH DOE expects to be able to participate during all appropriate and applicable meetings and trainings between the vendor and any subcontractor(s) identified in this proposal.	**M**	Yes	Standard	
Topic 33.2 Timeliness of Com.	The Program Manager shall return calls from NH DOE staff and respond to email messages within 24 hours. If the Program Manager is not available to take calls and return messages, NH DOE shall be notified in advance. In the event that the Program Manager is not available, the vendor shall notify NH DOE as to whom to contact in his or her absence, and shall provide contact information for such individual. The vendor shall confirm its agreement to meet this requirement.	**M**	Yes	Standard	
	The vendor's response should address any technology that will be proposed to support effective communication, any regular written communication or reports that are proposed, and processes and procedures that will be taken to monitor and evaluate the effectiveness of project communication.	**M**	Yes	Standard	
Topic 33.3 Monthly Reports	The vendor shall provide a monthly report that summarizes actions taken, issues that arose, issue resolution that occurred, outstanding issues and when they will be resolved, upcoming deadlines, work that will occur in the next month and beyond, and so forth. These reports shall be sent monthly to NH DOE by the third business day of the following month.	**M**	Yes	Standard	

Topic 34 Program Improvement Plans					
Topic 34 Program Imp. Plans	For each phase of the program including development, production, shipping and receipt, administration (of paper- based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities, the vendor shall provide a report that addresses the relevant phase by detailing the activities completed and by providing recommendations for improvement for the next assessment cycle. The report shall also detail errors, problems and/or discrepancies by district and by school. The report will allow NH DOE to detect any patterns in the errors, problems, or discrepancies noted in the report and to use that information to clarify instructions in the Assessment Administration and/or Coordinator Manuals. This report shall be completed within one month of completing the relevant phase.	**M**	Yes	Standard	
Topic 35 Risk Management and Quality Assurance					
	The vendor shall provide a detailed description of the proposed approach to timely identification and effective action on issues and risks.	**M**	Yes	Standard	
	Vendors shall specifically address timeline issues, risks, and mitigation and contingency plans for all aspects of the project. These plans should include: • Description of the proposed approach to managing risks and issues. • A sample tracking document. • Methodology to ensure that the State staff is involved in the process. • Description of known risks and proposed steps to mitigate them.	**M**	Yes	Standard	
	Additional details may be provided in the response to relevant requirements and specifications. The vendor should highlight its and its proposed subcontractors proven ability to document and enact risk management strategies – especially as they relate to the development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities of high-visibility assessments.	**O**	Yes	Standard	

	The vendor should submit sample risk assessment documentation used in an existing program to demonstrate the comprehensiveness of its ability to conduct contingency planning for a variety of conditions. This risk assessment documentation may be submitted as an attachment to the proposal. This documentation should also highlight internal procedures and protocols for quality assurance in all aspects of delivering large-scale, statewide assessments – including test development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, and reporting.	**M**	Yes	Standard	
D-3	**Project Execution**		**Yes**	**Standard**	
D-3.1	**Implementation and Operation**		**Yes**	**Standard**	
Topic 36 Implementation Approach					
	The vendor shall provide a detailed description of the roles and responsibilities of vendor staff and State staff during pre-implementation, Implementation, and operational phases of the engagement. The description shall include the amount of time required of each staff member and when their time is needed during the implementation and operational phases of the project.		Yes	Standard	
Topic 36 Imp. Approach	The vendor shall provide one or more feasible implementation plans and user readiness. For each plan provided: • Identify timeframes for major milestones. • Discuss cost implications of the plan, including impact on maintenance fees; and • Address the level of risk associated with the plan.		Yes	Standard	
Topic 37 User Acceptance Testing					
Topic 37 User Accpt. Testing	The vendor shall provide a detailed description of the support the vendor will supply to assist State during user acceptance testing of the configured System for New Hampshire.		Yes	Standard	

	State staff will conduct Acceptance Testing, but support from the selected vendor is required. To define the type of support that will be provided, address the following questions: • Describe your testing methodology and include a proposed test plan. • Include the time the State will need to complete User Acceptance Testing of a component. • Include a description of the support will be provided to prepare State staff during Acceptance testing. • Include the preparation required for testing the configured Software. • Include the documentation that will be available to the testing team for the configuration. • Include any defects likely to be encountered. This information should be based on previous experience and include metrics from other projects to support the response. • Include time frames for investigation of planned or suspected defects. • Include time frame for defect correction. • Provide a sample User Acceptance Test Plan from a completed project as an appendix.		Yes	Standard	
D-3.2	**Ongoing Operations**		**Yes**	**Standard**	
Topic 38 Implementation Approach					

Topic 38 Help Desk Support	The vendor shall provide a detailed description of support available to the State to help them with the process of uploading and receiving files and other aspects of data validation and correction. The vendor shall: • Describe support for the State to assist with the process of uploading files and receiving files. Include hours of operation, response times, problem classification, and escalation procedures. • Describe your electronic problem tracking process and tools used. • Describe how user account management will be handled. • Describe how general support and maintenance skills are transferred to State technical support personnel for knowledge sharing. • Describe how are support and maintenance issues are tracked detailing methodology and if any additional software is required. • Describe process for maintenance of the general knowledge base. • Describe any particular procedures required to handle escalation and emergency calls • Detail the plan for preventive maintenance and for upgrade installations • Detail the types and frequency of support tasks required	**M**	Yes	Standard	
Topic 39 Support Center					
Topic 39 Support Center	The vendor's response must describe processes and procedures used to ensure timely and accurate assistance; measures used to monitor and document the efficiency and accuracy of the service provided; expected standards for performance and customer service (e.g., wait time, quality of service); and procedures to measure customer satisfaction with the services provided.	**M**	Yes	Standard	
	The vendor's response should address the processes, procedures, or systems that will be used to ensure that all interactions with districts and schools are documented and maintained in a system that allows for efficient access and review.	**M**	Yes	Standard	

	The vendor will provide for provide customer support to districts and schools throughout the registration, testing, and reporting cycles, with an emphasis on service provided at key periods such as registration of students and test administration.	**M**	Yes	Standard	
	The vendor will guarantee that help desk staffing will increase and/or decrease based on call volume and wait time/caller. When staffing increases/decreases will be determined in consultation with and with approval from the NH DOE management team.	**M**	Yes	Standard	
	The vendor will provide help desk and technical support via toll-free phone, e-mail, and/or other online methods Monday through Friday from 7:00 a.m. EST/EDT through 5:00 p.m. EST/EDT. This includes a dedicated technical support line for NH districts, schools, and state representatives.	**M**	Yes	Standard	
	The vendor will provide tiered levels of customer support to district and school administrators and educators. The vendor and states will agree upon the type of questions and issues that will be addressed by the vendor, what actions the support center and other vendor staff will take to resolve and/or answer those questions and issues, and the type of questions and issues that will be forwarded to the NH DOE for resolution. Support center staff must have the ability to reopen accidentally closed tests.	**M**	Yes	Standard	
	The vendor will ensure that all support center staff and other vendor staff are qualified and have been trained to provide the level of support required by their position.	**M**	Yes	Standard	

	5. The vendor must develop a Service Level Agreement (SLA) to ensure that the system specifications, performance, and support are appropriate and acceptable. The SLA should have Level 1 (basic level), Level 2 (intermediate level), and Level 3 (technical level) services. The SLA and support processes, shall include at a minimum the following: ● Availability; ● Reliability; ● Latency; ● Disaster recovery plan; ● Server backup plan; ● Recovery point objective; ● Issue resolution times; ● Maintenance windows; ● Service reporting; ● Support hours; ● Support contact information; ● Escalation; ● Errata notice template suitable for electronic posting and distribution (subject to state approval); and ● Change management.	**M**	Yes	Standard	
	The vendor must include a plan for timely electronic notification to district and school administrators and test administrators through email, posting a notice on the online system, and/or direct calling, of any issues affecting test administration.	**M**	Yes	Standard	
	The vendor must develop an errata notice template that includes a description of the issue, the timeline for resolution, and any required actions that need to be taken by district or school administrators and/or test administrators.	**M**	Yes	Standard	
Topic 40 Technical Reporting					

Topic 40 Technical Reporting	The vendor will produce and maintain adequate documentation of all technical processes, procedures, and analyses conducted on an ongoing basis throughout the registration, testing, and reporting cycles. One purpose of the documentation will be to enhance quality assurance and quality control. The technical documentation will be produced in a format that is accessible to the NH DOE and conveys useful information to the NH DOE about the technical quality of the assessment program.	**M**	Yes	Standard	
Topic 40.1 Technical Report	The vendor will design, develop, and produce an annual Technical Report that documents and provides the necessary evidence to demonstrate the quality of the technical processes and procedures related to the design, development, administration, and reporting of results from the NH assessments. As appropriate, the annual Technical Report must also provide evidence that the planned processes and procedures were implemented for the given year.	**M**	Yes	Standard	
	The Technical Report is one piece of evidence produced to demonstrate that each of the NH assessments and the assessment program as a whole serve their intended purposes and meet accepted professional standards for educational testing.	**M**	Yes	Standard	
	The NH DOE will approve the table of contents, design, and format for the Technical Report.	**M**	Yes	Standard	
	The annual Technical Report will not replace or fulfill the general requirement of ongoing technical documentation of the NH Assessment Program or for task-specific technical documentation specified in this RFP.	**M**	Yes	Standard	
	A final draft of the annual Technical Report will be delivered to the NH DOE no later than three months following the release of assessment results from operational assessments or three months following the completion of the administration of the Spring 2018 Field Test.	**M**	Yes	Standard	

	The annual Technical Report will be delivered to the NH DOE in a digital format suitable for posting and distribution through the NH DOE website.	**M**	Yes	Standard	
	The vendor's response must include a Technical Report that it has prepared for a large-scale state assessment program.	**M**	Yes	Standard	
	If applicable, a link to a publicly available Technical Report can be provided in the vendor's response to fulfill this requirement.	**O**	Yes	Standard	
Topic 41 Technical Advisory Committee					
Topic 41 Technical Advisory Committee	The vendor will support two meetings per year of a NH Assessment System Technical Advisory Committee.	**M**	Yes	Standard	
	The vendor will be represented at the meetings by the project director, lead psychometrician assigned to the project, and additional staff as needed based on the agenda for the meeting.	**M**	Yes	Standard	
	The NH DOE will select members of the Technical Advisory Committee and will be responsible for facilitating all meetings of the Technical Advisory Committee.	**M**	Yes	Standard	
	The vendor will be responsible for all activities related to planning for the meeting and for all costs associated with the meeting and activities, including reimbursements and payments made to TAC members.	**M**	Yes	Standard	
D-4	**Pricing**		**Yes**	**Standard**	
Topic 42 Pricing Model					
Topic 42 Pricing Model	The vendor shall provide a detailed description of the Pricing Model for the proposed solution that addresses the following components: • Fixed prices; • Recurring prices; • Price per transaction; • Price per batch; • Implementation pricing; • Operations pricing; • Transition services pricing; and • Other applicable prices.	**M**	Yes	Standard	

	The vendor shall include all components found in Appendix F. and D-5 ASSURANCE AND TRANSITION.	**M**	Yes	Standard	
D-5	**Assurance and Transition**		**Yes**	**Standard**	
Topic 43 Quality Control Sign-Offs					
Topic 43 Quality Control Sign-Offs	Reviews and signoffs for all deliverables shall be documented and available to NH DOE upon request. The vendor shall document the steps, timeline, and staff involved in the quality control procedures for each phase and deliverable of the project.	**M**	Yes	Standard	
Topic 44 Invoices					
Topic 44 Invoices	The vendor shall submit invoices according to the procedures and requirements set forth by NH DOE. It is expected that the payment schedule for this contract will be four quarterly and one final payment for the services performed and deliverables provided during each period. The fiscal year for the State of New Hampshire runs from July 1 to June 30. The last invoice for each fiscal year must be received by June 15. The final invoice for each assessment cycle must be provided by September 1.		Yes	Standard	
Topic 45 Transition					
Topic 45 Invoices	Proposals must include an end of service transition plan detailing the transfer of relevant assessment documents and materials. An organized transition that ensures the continuity of the state assessment program is of the essence. The Transition Plan must address the transfer of materials, both pre-existing and newly developed, from the vendor to NH DOE or another vendor upon termination or expiration of the contract.		Yes	Standard	
	The vendor shall assist NH DOE with all activities required to transfer all assessment documents and materials during the transition phase. Draft transition plans shall include procedures for the transition of documents and materials.		Yes	Standard	

	The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • Test development - all critical documents and materials used in the test development process; • Item and test specifications – all item format details, test map requirements, test blueprints, and technical reports; • Test books –all paper and electronic test booklets and electronic answer documents from previous test administrations; test maps for each form from the previous year's administration with keys and metadata;		Yes	Standard	
	The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • Passages and artwork – all photocopies of the original passages with source documentation, copies of contracts, original electronic art files and applicable permission information; • Item bank, item and test statistics – all item-level metadata and previous usage statistics, available test-level statistics, previous anchor range finding papers, rubrics, constructed-response materials such as training material protocols, previous operational and field test usage of each item year and form item position status; • Program administration - all critical documents and materials used with the test administration process;		Yes	Standard	

	The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • General program documentation –all critical documents and materials used for general program documentation and summary reports; • Reports –sample copies of all reports provided to districts and schools; • Manuals/guides –sample copies of all guides and manuals (hard copy and electronic versions) for the operational test administrations, and copies of all electronic materials posted on the state website during the operational test administration;		Yes	Standard	
	The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • Scoring information - all critical documents and materials used in the scoring process; • Scoring/reporting specifications – all documentation regarding scoring rules, aggregation rules, roll-up algorithms, and tables used to calculate student, school, district, and state results; • Psychometric and related assessment information required for the program - all critical documents and materials used for psychometric analyses and related procedures; • Professional development – all critical documents and materials used for professional development;		Yes	Standard	

	The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • Editing Specifications – all documentation that outlines how the state would like answer documents edited during the scanning process • Equating data files –all documentation that outlines layouts for files including item statistics, master file, pre-id, school/district score data and state-level score data; • Performance scoring specifications – all training papers, anchor sets, calibration papers, rubrics, and constructed-response scoring rules; previous year's score distributions for each item and historical reader agreement rates;		Yes	Standard	
	but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • Technical reports and other validity and reliability reports - all electronic copies of past technical reports produced by the previous vendor and electronic copies of any other reports that discuss the validity or reliability of the assessments; • Project plan - all documents that outline the tasks/deliverables and corresponding schedule for those tasks/deliverables;		Yes	Standard	
	The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • Schedules - all previous project schedules containing dates/durations for the following tasks: • Developing items, forms, and materials • Enrollment and pre-identification • Receiving and scanning • Scoring and reporting		Yes	Standard	

	The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s): • Packaging specifications - all documentation concerning packaging algorithms and shipping points; and • Print specifications - all spreadsheets detailing print specifications for test booklets, scannables, answer documents, labels, envelopes, and manuals.		Yes	Standard	
	Draft Transition Plans shall include procedures for the transition of documents and materials related to the following: • Program administration - The vendor shall ensure that all critical documents and materials used with the test administration process are transferred efficiently between NH DOE and/or vendors. • Test development - The vendor shall ensure that all critical documents and materials used in the test development process are transferred efficiently between NH DOE and/or vendors.		Yes	Standard	
	Draft Transition Plans shall include procedures for the transition of documents and materials related to the following: • Scoring information - The vendor shall ensure that all critical documents and materials used in the scoring process are transferred efficiently between NH DOE and/or vendors. • Psychometric and related assessment information required for the program - The Vendor shall ensure that all critical documents and materials used for psychometric analyses and related procedures are transferred efficiently between NH DOE and/or vendors.		Yes	Standard	

	Draft Transition Plans shall include procedures for the transition of documents and materials related to the following: • General program documentation – The vendor shall ensure that all critical documents and materials used for general program documentation and summary reports are transferred efficiently between NH DOE and/or vendors. • Professional development – The vendor shall ensure that all critical documents and materials used for professional development are transferred efficiently between NH DOE and/or vendors.		Yes	Standard	
	The vendor must describe the process for the safe handling of State data during the transition phase.		Yes	Standard	

Application Requirements

APPLICATION REQUIREMENTS					
State Requirements			Vendor		
Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
GENERAL SPECIFICATIONS					
A1.1	Ability to access data using open standards access protocol (please specify supported versions in the comments field).	M	Yes	Standard	Our system is APIP compliant and validates against the QTI/APIP 2.2 standard.
A1.2	Data is available in commonly used format over which no entity has exclusive control, with the exception of National or International standards. Data is not subject to any copyright, patent, trademark or orhter trade secret regulation.	M	Yes	Standard	
A1.3	Web-based compatible and in conformance with the following W3C standards: HTML5, CSS 2.1, XML 1.1	M	Yes	Standard	
APPLICATION SECURITY					
A2.1	Verify the **identity or authenticate** all of the system client applications before allowing use of the system to prevent access to inappropriate or confidential data or services.	M	Yes	Standard	
A2.2	Verify the **identity and authenticate** all of the system's human users before allowing them to use its capabilities to prevent access to inappropriate or confidential data or services. .	M	Yes	Standard	
A2.3	Enforce unique user names.	M	Yes	Standard	
A2.4	Enforce complex passwords for Administrator Accounts in accordance with DoIT's statewide *User Account and Password Policy*	M	Yes	Standard	
A2.5	Enforce the use of complex passwords for general users using capital letters, numbers and special characters in accordance with DoIT's statewide User Account and Password Policy.	M	Yes	Standard	
A2.6	Encrypt passwords in transmission and at rest within the database.	M	Yes	Standard	
A2.7	Establish ability to expire passwords after a definite period of time in accordance with DoIT's statewide User Account and Password Policy	M	Yes	Standard	
A2.8	Provide the ability to limit the number of people that can grant or change authorizations		Yes	Standard	
A2.9	Establish ability to enforce session timeouts during periods of inactivity.		Yes	Standard	
A2.10	The application shall not store authentication credentials or sensitive data in its code.	M	Yes	Standard	
A2.11	Log all attempted accesses that fail identification, authentication and authorization requirements.	M	Yes	Standard	
A2.12	The application shall log all activities to a central server to prevent parties to application transactions from denying that they have taken place.	M	Yes	Standard	

Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
A2.13	All logs must be kept for 120 days	M	Yes	Standard	
A2.14	The application must allow a human user to explicitly terminate a session. No remnants of the prior session should then remain.	M	Yes	Standard	
A2.15	Do not use Software and System Services for anything other than they are designed for.	M	Yes	Standard	
A2.16	The application Data shall be protected from unauthorized use when at rest	M	Yes	Standard	
A2.17	The application shall keep any sensitive Data or communications private from unauthorized individuals and programs.	M	Yes	Standard	
A2.18	Subsequent application enhancements or upgrades shall not remove or degrade security requirements	M	Yes	Standard	
A2.19	Utilize change management documentation and procedures	M	Yes	Standard	

Testing

TESTING					
State Requirements			Vendor		
Req #	**Requirement Description**	**Criticality**	**Vendor Response**	**Delivery Method**	**Comments**
APPLICATION SECURITY TESTING					
T1.1	All components of the Software shall be reviewed and tested to ensure they protect the State's web site and its related Data assets.	M	Yes	Standard	
T1.2	The Vendor shall be responsible for providing documentation of security testing, as appropriate. Tests shall focus on the technical, administrative and physical security controls that have been designed into the System architecture in order to provide the necessary confidentiality, integrity and availability.	M	Yes	Standard	
T1.3	Provide evidence that supports the fact that Identification and Authentication testing has been recently accomplished; supports obtaining information about those parties attempting to log onto a system or application for security purposes and the validation of users	M	Yes	Standard	
T1.4	Test for Access Control; supports the management of permissions for logging onto a computer or network	M	Yes	Standard	
T1.5	Test for encryption; supports the encoding of data for security purposes, and for the ability to access the data in a decrypted format from required tools.	M	Yes	Standard	
T1.6	Test the Intrusion Detection; supports the detection of illegal entrance into a computer system	M	Yes	Standard	
T1.7	Test the Verification feature; supports the confirmation of authority to enter a computer system, application or network	M	Yes	Standard	
T1.8	Test the User Management feature; supports the administration of computer, application and network accounts within an organization.	M	Yes	Standard	
T1.9	Test Role/Privilege Management; supports the granting of abilities to users or groups of users of a computer, application or network	M	Yes	Standard	
T1.10	Test Audit Trail Capture and Analysis; supports the identification and monitoring of activities within an application or system	M	Yes	Standard	
T1.11	Test Input Validation; ensures the application is protected from buffer overflow, cross-site scripting, SQL injection, and unauthorized access of files and/or directories on the server.	M	Yes	Standard	
T.1.12	For web applications, ensure the application has been tested and hardened to prevent critical application security flaws. (At a minimum, the application shall be tested against all flaws outlined in the Open Web Application Security Project (OWASP) Top Ten (http://www.owasp.org/index.php/OWASP_Top_Ten_Project))	M	Yes	Standard	
T1.13	Provide the State with validation of 3rd party security reviews - perform on the application and system environment. The review may include a combination of vulnerability scanning, penetration testing, static analysis of the source code, and expert code review (please specify proposed methodology in the comments field)	M	Yes	Standard	We employ a third party penetration testing firm to independently audit our testing platform to identify any security vulnerabilities or risks.
T1.14	Prior to the System being moved into production, the Vendor shall provide results of all security testing to the Department of Information Technology for review and acceptance.	M	Yes	Standard	
T1.15	Vendor shall provide documented procedure for migrating application modifications from the User Acceptance Test Environment to the Production Environment.	M	Yes	Standard	
STANDARD TESTING					

T2.1	The Vendor must test the software and the system using an industry standard and State approved testing methodology as more fully described in Appendix G.	M	Yes	Standard	
T2.2	The Vendor must perform application stress testing and tuning as more fully described in Appendix G.	M	Yes	Standard	
T2.3	The Vendor must provide documented procedure for how to sync Production with a specific testing environment.	M	Yes	Standard	
T2.4	The vendor must define and test disaster recovery procedures.	M	Yes	Standard	

Hosting-Cloud Requirements

HOSTING-CLOUD REQUIREMENTS

State Requirements			Vendor		
Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
OPERATIONS					
H1.1	Vendor shall provide an ANSI/TIA-942 Tier 3 Data Center or equivalent. A tier 3 data center requires 1) Multiple independent distribution paths serving the IT equipment, 2) All IT equipment must be dual-powered and fully compatible with the topology of a site's architecture and 3)Concurrently maintainable site infrastructure with expected availability of 99.982%	M	Yes	Standard	
H1.2	Vendor shall maintain a secure hosting environment providing all necessary hardware, software, and Internet bandwidth to manage the application and support users with permission based logins.	M	Yes	Standard	
H1.3	The Data Center must be physically secured – restricted access to the site to personnel wi	M		Standard	
H1.4	Vendor shall install and update all server patches, updates, and other utilities within 60 days of release from the manufacturer.	M	Yes	Standard	
H1.5	Vendor shall monitor System, security, and application logs.	M	Yes	Standard	
H1.6	Vendor shall manage the sharing of data resources.	M	Yes	Standard	
H1.7	Vendor shall manage daily backups, off-site data storage, and restore operations.	M	Yes	Standard	
H1.8	The Vendor shall monitor physical hardware.	M	Yes	Standard	
H1.9	Remote access shall be customized to the State's business application. In instances where the State requires access to the application or server resources not in the DMZ, the Vendor shall provide remote desktop connection to the server through secure protocols such as a Virtual Private Network (VPN).	M	Yes	Standard	
H1.10	The Vendor shall report any breach in security in conformance with State of NH RSA 359-C:20. Any person engaged in trade or commerce that is subject to RSA 358-A:3, I shall also notify the regulator which has primary regulatory authority over such trade or commerce. All other persons shall notify the New Hampshire attorney general's office.	M	Yes	Standard	
DISASTER RECOVERY					
H2.1	Vendor shall have documented disaster recovery plans that address the recovery of lost State data as well as their own. Systems shall be architected to meet the defined recovery needs.	M	Yes	Standard	
H2.2	The disaster recovery plan shall identify appropriate methods for procuring additional hardware in the event of a component failure. In most instances, systems shall offer a level of redundancy so the loss of a drive or power supply will not be sufficient to terminate services however, these failed components will have to be replaced.	M	Yes	Standard	
H2.3	Vendor shall adhere to a defined and documented back-up schedule and procedure.	M	Yes	Standard	
H2.4	Back-up copies of data are made for the purpose of facilitating a restore of the data in the event of data loss or System failure.	M	Yes	Standard	
H2.5	Scheduled backups of all servers must be completed regularly. The minimum acceptable frequency is differential backup daily, and complete backup weekly.	M	Yes	Standard	
H2.6	Tapes or other back-up media tapes must be securely transferred from the site to another secure location to avoid complete data loss with the loss of a facility.	M	Yes	Standard	

Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
H2.7	Data recovery – In the event that recovery back to the last backup is not sufficient to recover State Data, the Vendor shall employ the use of database logs in addition to backup media in the restoration of the database(s) to afford a much closer to real-time recovery. To do this, logs must be moved off the volume containing the database with a frequency to match the business needs.	M	Yes	Standard	
HOSTING SECURITY					
H3.1	The Vendor shall employ security measures ensure that the State's application and data is protected.	M	Yes	Standard	
H3.2	If State data is hosted on multiple servers, data exchanges between and among servers must be encrypted.	M	Yes	Standard	
H3.3	All servers and devices must have currently-supported and hardened operating systems, the latest anti-viral, anti-hacker, anti-spam, anti-spyware, and anti-malware utilities. The environment, as a whole, shall have aggressive intrusion-detection and firewall protection.	M	Yes	Standard	
H3.4	All components of the infrastructure shall be reviewed and tested to ensure they protect the State's hardware, software, and its related data assets. Tests shall focus on the technical, administrative and physical security controls that have been designed into the System architecture in order to provide confidentiality, integrity and availability.	M	Yes	Standard	
H3.5	The Vendor shall ensure its complete cooperation with the State's Chief Information Officer in the detection of any security vulnerability of the hosting infrastructure.	M	Yes	Standard	
H3.6	The Vendor shall authorize the State to perform scheduled and random security audits, including vulnerability assessments, of the Vendor' hosting infrastructure and/or the application upon request.	M			
H3.7	All servers and devices must have event logging enabled. Logs must be protected with access limited to only authorized administrators. Logs shall include System, Application, Web and Database logs.	M	Yes	Standard	
H3.8	Operating Systems (OS) and Databases (DB) shall be built and hardend in accordance with guidelines set forth by CIS, NIST or NSA	M	Yes	Standard	
H3.9	The Vendor shall notify the State's Project Manager of any security breaches within two (2) hours of the time that the Vendor learns of their occurrence.	M	Yes	Standard	
H3.10	The Vendor shall be solely liable for costs associated with any breach of State data housed at their location(s) including but not limited to notification and any damages assessed by the courts.	M			
SERVICE LEVEL AGREEMENT					
H4.1	The Vendor's System support and maintenance shall commence upon the Effective Date and extend through the end of the Contract term, and any extensions thereof.	M	Yes	Standard	
H4.2	The vendor shall maintain the hardware and Software in accordance with the specifications, terms, and requirements of the Contract, including providing, upgrades and fixes as required.	M	Yes	Standard	
H4.3	The vendor shall repair or replace the hardware or software, or any portion thereof, so that the System operates in accordance with the Specifications, terms, and requirements of the Contract.	M	Yes	Standard	
H4.4	All hardware and software components of the Vendor hosting infrastructure shall be fully supported by their respective manufacturers at all times. All critical patches for operating systems, databases, web services, etc, shall be applied within sixty (60) days of release by their respective manufacturers.		Yes	Standard	

Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
H4.5	The State shall have unlimited access, via phone or Email, to the Vendor technical support staff between the hours of 8:30am to 5:00pm- Monday thru Friday EST;	M	Yes	Standard	
H4.6	The Vendor shall conform to the specific deficiency class as described: o Class A Deficiency - Software - Critical, does not allow System to operate, no work around, demands immediate action; Written Documentation - missing significant portions of information or unintelligible to State; Non Software - Services were inadequate and require re-performance of the Service. o Class B Deficiency - Software - important, does not stop operation and/or there is a work around and user can perform tasks; Written Documentation - portions of information are missing but not enough to make the document unintelligible; Non Software - Services were deficient, require reworking, but do not require re-performance of the Service. o Class C Deficiency - Software - minimal, cosmetic in nature, minimal effect on System, low priority and/or user can use System; Written Documentation - minimal changes required and of minor editing nature; Non Software - Services require only	M	Yes	Standard	
H4.7	As part of the maintenance agreement, ongoing support issues shall be responded to according to the following: a. Class A Deficiencies - The Vendor shall have available to the State on-call telephone assistance, with issue tracking available to the State, eight (8) hours per day and five (5) days a week with an email / telephone response within two (2) hours of request; or the Vendor shall provide support on-site or with remote diagnostic Services, within four (4) business hours of a request; b. Class B & C Deficiencies –The State shall notify the Vendor of such Deficiencies during regular business hours and the Vendor shall respond back within four (4) hours of notification of planned corrective action; The Vendor shall repair or replace Software, and provide maintenance of the Software in accordance with the Specifications, Terms	M	Yes	Standard	
H4.8	The hosting server for the State shall be available twenty-four (24) hours a day, 7 days a week except for during scheduled maintenance.	M	Yes	Standard	
H4.9	A regularly scheduled maintenance window shall be identified (such as weekly, monthly, or quarterly) at which time all relevant server patches and application upgrades shall be applied.	M	Yes	Standard	
H4.10	If The Vendor is unable to meet the uptime requirement, The Vendor shall credit State's account in an amount based upon the following formula: (Total Contract Item Price/365) x Number of Days Contract Item Not Provided. The State must request this credit in writing.	M	Yes	Standard	
H4.11	The Vendor shall use a change management policy for notification and tracking of change requests as well as critical outages.	M	Yes	Standard	
H4.12	A critical outage will be designated when a business function cannot be met by a nonperforming application and there is no work around to the problem.	M	Yes	Standard	
H4.13	The Vendor shall maintain a record of the activities related to repair or maintenance activities performed for the State and shall report quarterly on the following: Server up-time; All change requests implemented, including operating system patches; All critical outages reported including actual issue and resolution; Number of deficiencies reported by class with initial response time as well as time to close.	M	Yes	Standard	

Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
H4.14	The Vendor will give two-business days prior notification to the State Project Manager of all changes/updates and provide the State with training due to the upgrades and changes.	M	Yes	Standard	

Support & Maintenance Requirements

SUPPORT & MAINTENANCE REQUIREMENTS					
State Requirements			Vendor		
Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
SUPPORT & MAINTENANCE REQUIREMENTS					
S1.1	The Vendor's System support and maintenance shall commence upon the Effective Date and extend through the end of the Contract term, and any extensions thereof.	M	Yes	Standard	
S1.2	Maintain the hardware and Software in accordance with the Specifications, terms, and requirements of the Contract, including providing, upgrades and fixes as required.	M	Yes	Standard	
S1.3	Repair Software, or any portion thereof, so that the System operates in accordance with the Specifications, terms, and requirements of the Contract.	M	Yes	Standard	
S1.4	The State shall have unlimited access, via phone or Email, to the Vendor technical support staff between the hours of 8:30am to 5:00pm- Monday thru Friday EST;	M	Yes	Standard	
S1.5	The Vendor response time for support shall conform to the specific deficiency class as described below or as agreed to by the parties: o Class A Deficiency - Software - Critical, does not allow System to operate, no work around, demands immediate action; Written Documentation - missing significant portions of information or unintelligible to State; Non Software - Services were inadequate and require re-performance of the Service. o Class B Deficiency - Software - important, does not stop operation and/or there is a work around and user can perform tasks; Written Documentation - portions of information are missing but not enough to make the document unintelligible; Non Software - Services were deficient, require reworking, but do not require re-performance of the Service. o Class C Deficiency - Software - minimal, cosmetic in nature, minimal effect on System, low priority and/or user can use System; Written Documentation - minimal changes required and of minor editing nature; Non Software - Services require only minor reworking and do not require re-performance of the Service.	M	Yes	Standard	
S1.6	The Vendor shall make available to the State the latest program updates, general maintenance releases, selected functionality releases, patches, and Documentation that are generally offered to its customers, at no additional cost.	M	Yes	Standard	
S1.9	For all maintenance Services calls, The Vendor shall ensure the following information will be collected and maintained: 1) nature of the Deficiency; 2) current status of the Deficiency; 3) action plans, dates, and times; 4) expected and actual completion time; 5) Deficiency resolution information, 6) Resolved by, 7) Identifying number i.e. work order number, 8) Issue identified by;	P	Yes	Standard	

Req #	Requirement Description	Criticality	Vendor Response	Delivery Method	Comments
S1.10	The Vendor must work with the State to identify and troubleshoot potentially large-scale System failures or Deficiencies by collecting the following information: 1) mean time between reported Deficiencies with the Software; 2) diagnosis of the root cause of the problem; and 3) identification of repeat calls or repeat Software problems.	P	Yes	Standard	
S1.11	As part of the Software maintenance agreement, ongoing software maintenance and support issues, shall be responded to according to the following or as asgreed to by the parties: a. Class A Deficiencies - The Vendor shall have available to the State on-call telephone assistance, with issue tracking available to the State, eight (8) hours per day and five (5) days a week with an email / telephone response within two (2) hours of request; or the Vendor shall provide support on-site or with remote diagnostic Services, within four (4) business hours of a request; b. Class B & C Deficiencies –The State shall notify the Vendor of such Deficiencies during regular business hours and the Vendor shall respond back within four (4) hours of notification of planned corrective action; The Vendor shall repair or replace Software, and provide maintenance of the Software in accordance with the Specifications, Terms and Requirements of the Contract; or as agreed between the parties	M	Yes	Standard	
S1.12	The Vendor shall use a change management policy for notification and tracking of change requests as well as critical outages.	M	Yes	Standard	
S1.13	A critical outage will be designated when a business function cannot be met by a nonperforming application and there is no work around to the problem.	M	Yes	Standard	
S1.14	The Vendor shall maintain a record of the activities related to repair or maintenance activities performed for the State and shall report quarterly on the following: All change requests implemented; All critical outages reported including actual issue and resolution; Number of deficiencies reported by class with initial response time as well as time to close.	M	Yes	Standard	

Project Management

PROJECT MANAGEMENT					
State Requirements			Vendor		
Req #	**Requirement Description**	**Criticality**	**Vendor Response**	**Delivery Method**	**Comments**
PROJECT MANAGEMENT					
P1.1	Vendor shall participate in an initial kick-off meeting to initiate the Project.	M	Yes	Standard	
P1.2	Vendor shall provide Project Staff as specified in the RFP.	M	Yes	Standard	
P1.3	Vendor shall submit a finalized Work Plan within ten (10) days after Contract award and approval by Governor and Council. The Work Plan shall include, without limitation, a detailed description of the Schedule, tasks, Deliverables, critical events, task dependencies, and payment Schedule. The plan shall be updated no less than every two weeks.	M	Yes	Standard	
P1.4	Vendor shall provide detailed monthly status reports on the progress of the Project, which will include expenses incurred year to date.	M	Yes	Standard	
P1.5	All user, technical, and System Documentation as well as Project Schedules, plans, status reports, and correspondence must be maintained as project documentation. (Define how- WORD format- on-Line, in a common library or on paper)	M	Yes	Standard	Delivered via KnowledeTree, our proposed online document management repository.

FERPA Requirements

	FERPA REQUIREMENTS		
State Requirements			**Vendor**
Req #	**Requirement Description**	**Criticality**	**Vendor Response / Acknowledgement**
FERPA REQUIREMENTS			
	Family Educational Rights and Privacy Act of 1974, as amended (FERPA, 20 U.S.C. 1232g; 34 CFR Part 99) – FERPA is a federal law that protects student and family privacy. It gives parents and students age 18 and over the rights to view the student's record and apply to amend it. It also requires education agencies to restrict access to the records by only authorized persons and prevent against inappropriate disclosure.	M	Yes
	Under FERPA, no private or confidential data will be released except under the following circumstances as stated in 34 CFR Part 99 Final Regulations for FERPA: 1. To teachers and officials of the district in which the student is enrolled when the determination has been made that there are legitimate educational interests, under Section 99.31(a)(1). 2. To school and district personnel when a student is seeking to enroll, under Section 99.31(a)(2). 3. To comply with a lawfully issued subpoena or court order, under Section 99.31(a)(9)(i). 4. To educational officials in connection with an audit or evaluation of a federal or state supported education program, under Section 99.32(c)(3). 5. To appropriate parties in connection with an emergency if such knowledge is necessary to protect the health and safety of the student or other individuals, under Section 99.36(a). In cases of health or safety emergency, the request for release must first be directed to the school district that owns the data. The data reporting/database manager, under Section	M	Yes

	Federal Education Rights to Privacy Act (FERPA) Access Procedures Student data will be disclosed only on the conditions that: (1) the party to whom the data are released does not disclose the information to any third party without the prior written consent of the authorized person within New Hampshire DOE, the company who provided the student assessment data (if assessment data are being disclosed), or the school district that owns the data; (2) the data are protected in a manner that does not permit the personal identification of an individual by anyone except the party referenced in the disclosure; and (3) the data are destroyed when no longer needed for the purposes under which the disclosure was granted.	M	Yes
	researchers, and the general public are fairly common in most school systems and state education agencies. In this regard, the FERPA statute provides that an education agency or institution may not have a policy of disclosing education records or personally identifiable information from education records, without prior consent from the parent or eligible student, unless it is considered directory information or falls under one of the other consent exceptions contained in the law [20 U.S.C. §1232(g)(b)(1)]. (For exceptions to consent guidelines, see Disclosure of Student Information.) Agencies should determine whether requests for data meet these exceptions on a case-by-case basis. Nothing in FERPA prohibits a school from disclosing information in aggregate, or in another form that is not personally identifiable. Personally identifiable information includes: • the student's name; • the name of the student's parent or other family member; • the address of the student or student's family; • a personal identifier, such as the student's social security number or student number; • a list of personal characteristics that would make the student's identity easily	M	Yes

	In circumstances that may lead to the identification of an individual, the disclosing education agency or institution must ensure that student-level information is not personally identifiable by removing the student’s name and ID number, as well as any “personal characteristics” and “other information that would make the student’s identity easily traceable.” This includes, but is not limited to, such factors as physical description (race, sex, appearance, etc.); date and place of birth; religion and national origin; participation in sports, clubs, and other activities; academic performance; employment; and disciplinary actions or criminal proceedings. “Other information that would make the student’s identity easily traceable” may also exist in the form of small cell sizes in aggregated or statistical information from education records.	M	Yes
	In cases where personal information cannot be removed, school officials must secure written parental consent before disclosing the data to outside organizations. The required consent form should specify: • the records that may be disclosed; • the purpose of the disclosure; and • the identity of the party or class of parties to whom the disclosure may be made. [34 CFR § 99.30(b)]	M	Yes

Section IV: Narrative Responses

Section IV: Narrative Responses

D-1 PROPOSED SOLUTION

D1.1 Test Construction

In this section, we propose our plan for designing New Hampshire's new statewide assessment system for operational test administration in spring 2018. Our state of the art test delivery system and our extensive Independent College and Career Ready (ICCR) item banks afford New Hampshire tremendous flexibility to design a statewide assessment system that validly, reliably, and fairly measures student achievement of New Hampshire College and Career Ready Standards. In Topic 1, we propose to design and implement a standards based computer adaptive assessment to measure student achievement of New Hampshire College and Career Ready Standards in ELA and mathematics. To ensure comprehensive coverage of all science standards at aggregate levels, while minimizing individual student testing time, we propose a matrix of psychometrically equivalent fixed-form tests to measure achievement of New Hampshire's academic standards in science. We note, however, that our test delivery system and ICCR item banks can be used to implement a variety of assessment designs, including fixed-form assessments, stage adaptive assessments, and fully adaptive standards-based assessments.

As we describe in Topic 2 Item Development, the ICCR banks are extensive and support adaptive test administrations in ELA and mathematics. These bank sizes allow participating ICCR states to implement educator and community review procedures to ensure that all items deployed for statewide assessments are aligned to state academic standards and meet community standards for appropriate content.

Quality assurance processes and quality control checks are the bedrock for all of AIR's assessment processes. In each section of this proposal, we describe the quality assurance processes that AIR implements to ensure the development and implementation of valid, reliable, and fair assessments of student academic achievement, as well as the quality control checks performed to verify the performance of those systems.

Topic 1 Test Design

Summative Assessments

ELA and Mathematics. For New Hampshire's summative assessments in ELA and mathematics, we propose to administer standards-based, computer adaptive tests. As we describe in detail in Topic 2, AIR's ICCR item banks have been developed to measure College and Career Ready Standards in statewide assessments.

The ICCR item pools have been constructed explicitly to support statewide assessment programs and have been administered as part of the Arizona, Florida, Ohio, Tennessee and Utah statewide assessments, and they have been embedded in Oregon's administration of the Smarter Balanced assessments. ICCR items were developed in conjunction with state departments of education, following a rigorous system of internal and external review procedures implemented by each of the participating states. Thus, all ICCR items have been reviewed by content review committees comprising educators in one or more states, as well as by bias and fairness review committees in those states. All ICCR items have been field tested in embedded slots within operational test administrations so that item parameter estimates, based on large numbers of students participating in state summative assessments, are highly precise and stable. Following field test administration, item statistics for all ICCR items are evaluated for discrimination, difficulty, and differential item functioning, and any items flagged for out of range statistics are forwarded for further review by AIR content staff, AIR psychometric staff, and state assessment staff, with items that performed poorly rejected from the pool.

Following each test administration, ICCR items are calibrated using multiple IRT models and linked back to each state's scale. In addition, a linking design was enacted to link item parameters between each of the state assessment systems to the common ICCR scale. Thus, the performance levels for each ICCR state can be represented on the ICCR scale. Moreover, benchmarks for other assessments, such as NAEP, TIMSS, or PISA identified for any participating state, can also be represented on the ICCR scale. The ICCR items are therefore not only very robust, with each item having passed through rigorous reviews in typically multiple statewide assessment systems, the validity of test score interpretations based on the ICCR scale are greatly enhanced by virtue of abundant benchmarking of scale score locations.

The ICCR item bank has been constructed to enact a model college and career ready blueprint. Blueprints are generally consistent across the participating states, but do allow flexibility for each state to craft and implement a custom blueprint that meets the unique requirements and needs of each state. AIR will work with the Department to finalize a blueprint that aligns to the New Hampshire College and Career Ready Standards; yields test scores that are valid and reliable, both overall and for all domain reporting categories; and ensures that test administration times remain within desired limits.

An important focus of the New Hampshire Statewide Assessments is on reporting, and adaptive test administration provides important advantages for interpreting test scores for all test users. With fixed-form tests, test forms are necessarily constructed using only a small sample of items comprising the content domain. At the level of individual standards, fixed-form tests may contain only 1–2 items and are thus not representative of all items comprising the standard. Conversely, adaptive test administrations proceed from the full pool of available items, providing a far better representation of the intended content domain to be assessed. As with a fixed-form test, any one student may see only 1–2 items sampled from a standard, but in the aggregate students are seeing different samples of items measuring those standards; at the level of classrooms and schools, each standard is assessed across a much broader sample of items representing the content domain, allowing for a much more robust measure of classroom and school performance at a finer grained level of analysis. Thus, AIR is able to provide educators with standard-level analyses of class and school performance that are highly reliable and lead to more valid interpretations of student performance since those indicators are based on a more representative range of the knowledge and skills subsumed within those standards.

Science Assessments. As the Department is aware, the ICCR item bank in science is being developed in collaboration with a group of states developing common item and item cluster specifications to measure three dimensional science standard. As with the ICCR item pools in ELA and mathematics, the ICCR item pools in science will allow the Department to identify the performance standards for any consortium state on the common ICCR scale. ICCR science items will be available for administration in spring 2018. For spring 2018, we propose to administer science item clusters and items in an operational field test design that will administer test forms that meet all blueprint specifications, allow for calibration and equating of science items to establish the ICCR science scale, and support identification and adoption of performance standards for New Hampshire's statewide assessments in science. Because parameter estimation must follow test administration in spring 2018, immediate scoring and reporting of test scores will not be possible as part of the first administration of the science assessments. However, since standard-setting workshops necessary to recommend performance standards cannot commence until after the first operational test administration in order to obtain impact data, the post-equating approach would not further delay reporting of student test scores.

Interim Assessments

Interim Assessments

To support more truly formative uses of interim test results, we propose an interim test design around AIRWays, our innovative reporting tool for interim benchmark assessments. AIRWays is designed to leverage testing events to drive students and teachers to interact. The system reports only non-secure test results and shows the teacher both the item and each student's actual response. This provides a platform and opportunity for the teacher and student to begin exploring gaps in knowledge to support instruction. AIRWays fosters more truly formative assessments by allowing teachers and students to explore their responses to test items and to focus instruction in more meaningful ways. Truly formative testing requires teachers and students to work together to understand how and why students respond to test items, which precludes administration of secure item content.

Eventually, however, the NH DOE may wish to aggregate student interim results to replace a summative test administration for accountability purposes. Since any accountability use of test items requires rigorous test security, an assessment regime in which interim assessments replace a summative assessment for accountability would have to be administered using the same ICCR item pool and test administration procedures as the summative assessments, precluding reporting of test results in AIRWays.

Since aggregating interim assessments to replace a summative test administration is not yet an available option for New Hampshire, we propose a two-stage solution to the NH DOE's interim assessment requirement. Until USED determines that summative assessments can be replaced by aggregating a series of interim assessments, we propose to deliver New Hampshire's interim assessments using one of several non-secure formative assessment banks.

One option is to deliver interim assessments from our own Learning Point Navigator item bank, as well as items from banks provided by Key Data System (KDS). Items in Navigator were developed by the same AIR test developers who develop items for summative tests, including the ICCR items. Items in the Navigator bank thus have the same look and feel as our ICCR items and are delivered using the same test engine as the summative assessments we are proposing. The KDS pools, which offer quality mathematics items aligned to state College and Career Ready Standards, provide a variety of item types that will allow us to deliver interim math assessments consistent with the summative assessments in both content and functionality. AIR will work with the Department to embed interim items in the summative assessments to calibrate and equate the Navigator and KDS items pools to the ICCR scale so that interim assessment results based on these item pools can be reported on the New Hampshire reporting scale.

New Hampshire may also be able to access Utah's formative item pool populated with items already calibrated and equated to the ICCR scale. The item pools making up Utah's formative assessments have several very important strengths. Utah's formative item pool is composed of the same kind of items and item types used to administer Utah's Student Assessment of Growth and Excellence (SAGE) accountability assessments. In fact, the formative item pools were drawn from the SAGE item pools. Thus, all formative items were developed to align with Utah's Core Standards, which are consistent with New Hampshire College and Career Ready Standards. Moreover, because these items were originally part of Utah's SAGE accountability assessment item pool, the items were developed following the same rigorous procedures described in Topic 2 Item Development. All items currently in the formative pool passed through all levels of content, fairness, and field-test data review and were promoted to the SAGE operational item pool. All items in the formative pool have item calibrations estimated from student responses in operational test administrations. As one of several states collaborating in the development of AIR's ICCR item pool, Utah's formative item bank includes the same high-quality items and versatile machine-scored item types as the ICCR item banks that we propose for New Hampshire's summative assessments. Thus, the look and feel of the Utah formative items will be consistent with the ICCR summative assessments. Importantly, IRT parameters for items in the formative pool are already linked to the ICCR scale to support consistent reporting of interim and summative assessments results.

Moreover, because Utah's formative item banks were designed to support multiple-opportunity adaptive test administrations, they can be flexibly deployed to meet a range of formative assessment goals. For example, the NH DOE could elect to deploy a system of multiple-opportunity comprehensive interim assessments configured to administer adaptive test administrations meeting a proportional length summative blueprint. These interim assessments can be used by educators to track student progress toward achievement of the grade-level standards. Alternatively, the formative pool can be configured to administer a series of fixed-form benchmark assessments that educators can administer to evaluate student achievement of discrete instructional modules.

Utah's SAGE and ICCR scales have been linked via a common-item design so that interim assessment results, including scale scores and performance levels, can be reported on the same scale used to report

New Hampshire's summative test results. This allows educators to monitor student progress toward achieving New Hampshire College and Career Ready Standards.

Because the items making up the Utah formative assessments are not secure, the results of benchmark assessments can be reported back to educators using AIRWays, allowing teachers to observe how students respond to individual items and to identify gaps in student understanding.

Should the US Education Department (USED) determine that the Every Student Succeeds Act (ESSA) allows states to report aggregate interim assessment results in lieu of a summative assessment for accountability purposes, the NH DOE may wish to offer this option to New Hampshire schools. Of course, any interim assessment system designed to substitute for summative test scores would need to satisfy the same peer review elements required of summative assessment systems, including requirements for item and test security. Thus, should the Department wish to transition to an interim assessment system that supports requirements for an accountability system, we propose to administer both the interim and summative assessments from the ICCR item pools. Test administrations can be configured for comprehensive or block interim assessments in addition to grade-level summative assessments. The ICCR item pools are sufficiently large to support both interim and summative test administrations, and AIR's adaptive algorithm is configured to ensure that students are not administered the same item across test administrations. (this can be relaxed to allow a previously administered item if necessary to meet blueprint.) Because the accountability-based interim assessments would need to be secure, reporting for both the interim and summative assessments would be through our online reporting system (ORS), which we describe fully in Topic 22 Reports, and which provides educators with a highly intuitive and powerful tool for navigating assessment results.

Blueprints

The ICCR item banks have been constructed to enact model college and career ready blueprints that are generally consistent across participating states, but do allow flexibility for each state to craft and implement a custom blueprint that meets the unique requirements and needs of each state. AIR will work with the Department to finalize a blueprint that aligns to the New Hampshire College and Career Ready Standards and that yields test scores that are valid and reliable, both overall and for all domain reporting categories. In addition, as New Hampshire works to revise their academic standards, AIR will work with the Department to revise the blueprints as necessary to align with the new standards.

ELA. Although multiple configurations can be supported, the ICCR item banks have been constructed to support model ELA blueprints that are designed to report student achievement in Reading Literature, Reading Informational Text, and Writing/Language. Exhibit D1.1-1 provides a sample ELA blueprint for grades 3–5. This blueprint supports a test design in which students would be administered four reading passages (two Reading Literature and two Reading Informational Text), two editing tasks comprised of 3–4 language items each, and one writing prompt. AIR proposes to work with Department to evaluate whether this blueprint meets the needs of New Hampshire's statewide assessment system and to make modifications to it as needed.

Exhibit D1.1-1: Sample ELA Blueprint for Grades 3–5

Reading	Literary: 50% Informational: 50%		42 items
Reading Literature			**16–18**
	Key Ideas and Details		**4–6**
	Craft and Structure		**4–6**
	Integration of Knowledge and Ideas		**1–4**
	Language/Vocabulary Acquisition and Use		**1–2**
	Listening		**0–4**
Reading Informational			**16–18**
	Key Ideas and Details		**6–8**
	Craft and Structure		**6–8**
	Integration of Knowledge and Ideas		**2–6**
	Language/Vocabulary Acquisition and Use		**1–2**
	Listening		**0–4**
Language/Editing			**6–8**
Writing			**1**
Task 1: Expository Essay			**0–1**
Task 2: Argumentative Essay			**0–1**

Mathematics. The model blueprints on which the ICCR mathematics item banks have been constructed are designed to support content domain reporting of student performance, combining domains when necessary to support reporting requirements. The blueprints call for 50 items for the grades 3–8 tests.

The blueprints contain item ranges for each reporting category, content cluster, and content standard, ensuring comprehensive content coverage while still allowing flexibility to adapt item selection to student ability. The blueprints also contain ranges for Depth of Knowledge (DoK), ensuring that test administrations probe student achievement across the range of cognitive demand specified in the academic standards. AIR proposes to work with the Department to ensure that this blueprint meets the needs of New Hampshire's statewide assessment system, and to make modifications to it as needed. Exhibit D1.1-2 below provides a sample from the grade 7 blueprint showing how the minimum and maximum number of items are nested in reporting category/cluster/standard.

Exhibit D1.1-2: Sample Mathematics Blueprint for Grade 7

Domain			Total	
Cluster			50	50
Standard			MIN	MAX
RP	**Ratios and Proportional Relationships**		11	13
A	Analyze proportional relationships and use them to solve real-world and mathematical problems.		11	13
7.RP.1			0	5
7.RP.2(abcd)			0	5
7.RP.3			0	5

Exhibit D1.1-2: Sample Mathematics Blueprint for Grade 7 (continued)

Domain			Total	
Cluster			50	50
Standard			MIN	MAX
NS	**The Number System**		9	11
B	Apply and extend previous understandings of operations with fractions to add, subract, multiply, and divide rational numbers.		9	11
7.NS.1(abcd)			0	4
7.NS.2(abcd)			0	4
7.NS.3			0	4
EE	**Expressions and Equations**		8	10
C	Use properties of operations to generate equivalent expressions.		2	6
7.EE.1			0	3
7.EE.2			0	3
D	Solve real-life and mathematical problems using numerical and algebraic expressions and equations.		2	6
7.EE.3			0	3
7.EE.4(ab)			0	3
G	**Geometry**		9	11
E	Draw, construct, and describe geometrical figures and describe the relationships between them.		2	6
7.G.1			0	2
7.G.2			0	2
7.G.3			0	2
F	Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.		2	6
7.G.4			0	2
7.G.5			0	2
7.G.6			0	2
SP	**Statistics and Probability**		9	11
G	Use random sampling to draw inferences about a population.		0	3
7.SP.1			0	2
7.SP.2			0	2
H	Draw informal comparative inferences about two populations.		0	3
7.SP.3			0	2
7.SP.4			0	2
I	Investigate chance processes and develop, use, and evaluate probability models.		0	6
7.SP.5			0	2
7.SP.6			0	2
7.SP.7(ab)			0	2
7.SP.8(abc)			0	2
		TOTAL ITEMS (for affinity groups)	46	56

Science. Construction of test blueprints for New Hampshire college and career ready standard science presents a special challenge due to the cluster design of test items and the very large number of standards to be assessed. The goal of blueprint construction for these assessments is to ensure that students are administered psychometrically equivalent test forms, with respect both to coverage of discipline core ideas and distribution of test information. And at the aggregate level, to ensure that full range of standards are assessed and that group means for standards are as precise as possible.

Exhibit D1.1-3 presents a sample grade 5 ICCR science blueprint. In this blueprint, each student is administered a summative test form comprising 7 clusters and 12 stand-alone items, plus an embedded field test slot for administering one field test cluster or six stand-alone items. Three clusters and approximate 16 stand-alone items measure standards in the Physical Science DCI, while the Life Science and Earth/Space Science DCIs are each measured by 2 clusters and approximately 12 stand-alone items.

Employing a matrix design, multiple test forms are constructed, each conforming to blueprint specifications but measuring different standards within the DCIs. Thus, while each student is administered equivalent numbers of clusters and items for each DCI, students in a classroom will be assessed across the range of standards defining each DCI and resulting in aggregate measures of science achievement that measure the entire content domain defined by the New Hampshire College and Career Ready Standards. Moreover, for larger aggregate units, student achievement of science can be meaningfully evaluated at the standard level, helping educators to identify areas of strength and weakness in the science curriculum.

Exhibit D1.1-3: Sample Grade 5 Science Blueprint

	Item Ranges - Operational						Item Ranges - Embedded Field Test						Total Items	
	Min Clusters	Max Clusters	Min Stand Alone Items	Max Stand Alone Items	Min Items	Max Items	Min Clusters*	Max Clusters*	Min Stand Alone Items	Max Stand Alone Items	Min Items	Max Items	Min Items	Max Items
Physical Science	**3**	**3**	**4**	**4**	**15**	**17**	**0**	**1**	**0**	**6**	**0**	**6**	**21**	**23**
3-PS2-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-PS2-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-PS2-3	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-PS2-4	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-PS3-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-PS3-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-PS3-3	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-PS3-4	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-ESS3-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-PS4-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-PS4-3	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-PS1-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-PS1-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-PS1-3	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-PS1-4	0	1	0	1	0	5	0	1	0	1	0	6	0	6
Life Science	**2**	**2**	**4**	**4**	**11**	**13**	**0**	**1**	**0**	**6**	**0**	**6**	**17**	**19**
3-LS2-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-LS4-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-LS4-3	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-LS4-4	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-LS1-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-LS3-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-LS3-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-LS4-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-PS4-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-LS1-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-LS1-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-PS3-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-LS1-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-LS2-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
Earth/Space Science	**2**	**2**	**4**	**4**	**11**	**13**	**0**	**1**	**0**	**6**	**0**	**6**	**17**	**19**
3-ESS2-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-ESS2-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
3-ESS3-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-ESS1-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-ESS2-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-ESS2-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
4-ESS3-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-ESS2-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-ESS2-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-ESS3-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-PS2-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-ESS1-1	0	1	0	1	0	5	0	1	0	1	0	6	0	6
5-ESS1-2	0	1	0	1	0	5	0	1	0	1	0	6	0	6
Total	**7**	**7**	**12**	**12**	**39**	**41**	**0**	**1**	**0**	**6**	**6**	**6**	**45**	**47**

Constructing Test Forms

ICCR items are currently being used to support a range of assessment systems, including those administering fixed-form tests (Florida) and computer adaptive tests (Utah). We propose to offer the New Hampshire statewide assessments in ELA and mathematics as computer adaptive test administrations in order to provide the strongest base to support valid and reliable reporting of test results for classrooms and schools. Because the ICCR science pools are currently more limited and because the number of standards that must be assessed is very large, we propose to construct a matrix of fixed-form tests at each grade level that conform to a common blueprint for score reporting but allow for comprehensive coverage of all standards at aggregate levels of reporting. Thus, while clusters and stand-alone items will be developed for all standards, each student will be administered clusters and stand-alone items measuring only a subset, but at least 51% of, the standards. Each test form would, however, conform to specifications for number of clusters/items per reporting category and other blueprint requirements.

Both the adaptive and matrix test form designs result in more comprehensive coverage of the content domain defined by New Hampshire College and Career Ready Standards. Topic 20.2 details our procedures for configuring the adaptive algorithm to administer standards-based assessments that meet blueprint, as well as procedures for constructing fixed-form tests that are psychometrically equivalent across forms within the matrix design and across test administrations.

Topic 1.1 Standards Revision

Because the Department is embarking on a revision of the New Hampshire College and Career Ready Standards, revisions to the adopted blueprints may be necessary for ELA and mathematics following the spring 2018 test administration and for science in 2021. AIR test developers will work with the Department to modify the test blueprints to reflect revisions to the academic standards. In Topic 1, we provide sample blueprints for measuring New Hampshire College and Career Ready Standards. AIR test developers will work closely with the Department to customize blueprints to meet New Hampshire's unique assessment goals. Following revision of the state's academic standards, AIR will work with the Department to review all standards revisions and to identify necessary changes to the blueprints.

Of concern with any significant revision of blueprints is that the blueprint modifications may alter the construct being measured. This is fairly obvious in cases where, for example, writing items are removed from an ELA assessment or probability standards are shifted between grade-level assessments. The removal or addition of standards may have important consequences for the validity of inferences that can be made from test scores. In particular, standard-setting panelists will have been guided by a set of performance level descriptors (PLDs) and an ordered-item book of test items constructed to align with the New Hampshire blueprints enacted in spring 2018. The PLDs represent the valid inferences that can be made from student test scores and performance level classifications. Changes to the test blueprints may impact the validity of some of the assertions in the PLDs.

To evaluate whether PLDs continue to be valid descriptors of achievement at each performance level, AIR will convene a panel of educators to review modified blueprints, PLDs, and OIBs to determine whether blueprint revisions warrant either modifications to the PLDs or reestablishment of performance standards.

Review of performance standards will be conducted by a panel of educators consistent with those recruited for the summer 2018 standard-setting workshops. We propose subject area panels with three representatives each from grade bands 3–5 and 6–8.

Panelists will be provided with the current PLDs, the standard setting OIBs, and an item map highlighting the location of items with respect to adopted performance standards. Panelists will also be provided with the original and revised blueprints, and the OIBs and item maps will highlight items that would no longer be administered given deletions in the revised blueprint or a range of additional items representing additions to the revised blueprint.

Panelists will review the PLDs, the notated OIBs, and the item maps in light of blueprint revisions to determine whether the PLDs as constructed for the standard-setting workshops continue to be valid descriptors of student achievement or whether the PLDs must be revised. Where revisions to PLDs are required, panelists will work to revise the PLDs in ways that support valid interpretations of performance level classifications.

AIR psychometricians will work with the Department and in consultation with the TAC to determine whether results indicate that PLDs can stand with no or minor revisions, indicating that the assessed construct is reasonably similar following blueprint revision, or whether PLD revisions based on the revised blueprint indicate substantial alteration of the assessed construct and signify a need to adopt new performance standards.

Topic 1.2 Test Administration

Testing Time. Test blueprints are designed to achieve multiple objectives. As a criterion test of student achievement, the New Hampshire blueprints must yield test score interpretations indicating the degree to which students have achieved the knowledge and skills prescribed in the New Hampshire College and Career Ready Standards following instruction in each grade level. Test blueprints must therefore align with New Hampshire College and Career Ready Standards, ensuring that students are administered items sampling the full range of academic content standards. The more comprehensively that the test blueprints sample from the content domain, the more confidence we have that individual test scores validly index student achievement of New Hampshire's standards.

While comprehensive coverage of the standards places some constraints on how short the assessments can be, reliable measurements of overall subject area performance can be made with fewer items than are currently administered. Test lengths in the proposed blueprints are importantly determined by the desire of state assessment programs to report subscale performance for individual students.

The Department seeks a test design that limits test administration times to about two hours per subject area assessment. AIR will work with the Department to design and implement test blueprints that allow the Department to meet test administration time limits while still accomplishing necessary measurement and reporting goals. Exhibit D1.1-4 shows the 85th percentile test administration times for the spring 2016 administration of Utah's adaptively-administered SAGE assessments. The SAGE assessments enact blueprints that are similar to the blueprints proposed for the New Hampshire statewide assessments, and the SAGE assessments are not timed. Because SAGE administers two writing tasks at each grade, the table in Exhibit D1.1-4 divides the SAGE writing times by half to account for the single writing task proposed for New Hampshire. As the table indicates, total test administration times at the 85th percentile are uniformly at or below eight hours at the elementary and middle school grades, and even lower at the high school grades. Test administration times vary across grades, with elementary school students taking more time to complete reading tests than middle and high school students, and both elementary and middle school students required more test administration time to complete the mathematics tests than did high school students. AIR will work with the Department to finalize blueprints that yield reliable domain reporting while keeping test administrations times to a minimum.

Exhibit D1.1-4 Spring 2016 Utah SAGE 85th Percentile Testing Times.

	Utah SAGE 85th Percentile Testing Times				
Grade	Reading	Writing*	Math	Science	Total
3	2.30	1.51	2.28	1.29	7.38
4	2.20	1.56	2.35	1.53	7.63
5	2.23	1.68	2.91	1.23	8.04
6	2.30	1.61	2.71	1.36	7.97
7	1.98	1.38	2.69	1.59	7.63
8	1.87	1.28	2.51	1.68	7.34
9	1.82	1.26	1.97	1.80	6.84
10	1.68	1.12	1.77	1.35	5.91
11	1.57	1.03	1.97	1.28	5.84

*Because SAGE administers two writing tasks, 85th percentile SAGE writing times were divided by two.

Test Segmentation. Standardization of test administrations is essential for valid interpretation of test scores and comparison of results between students, schools, and districts. Standardization does not, however, imply rigidity in the administration of achievement tests. To yield valid interpretations of test scores, tests must measure the intended achievement construct while minimizing the influence of construct-irrelevant factors that may influence student performance. This is plain in the case of accommodations provided to students with disabilities, but it also applies for students and schools more generally.

Test administrations that require sustained focus of attention may be unduly influenced by individual differences in attentional capacity. AIR's test delivery system enforces standardization where it counts, in the presentation of test content and collection of student responses, but it allows great flexibility for teachers to administer tests over multiple sessions and pause the tests when needed. Segmentation of test administrations and pause rules provide tremendous flexibility to accommodate shorter or longer class periods.

Test administrations are designed to allow pausing at any point and to return students to where they left off. The pause function is always available. For pauses lasting longer than a Department-defined grace period, typically set at 20 minutes, TDS allows students to pick up where they left off but prevents them from viewing or changing responses to items administered before the pause. Pausing allows educators to fit test sessions into any bell period configuration. Pausing also recognizes that students do not work at the same pace or share the same attention span; educators can use pausing to ensure that all students are able to demonstrate what they know and are able to do.

Topic 1.3 Student Registration

We propose to use AIR's Test Information Distribution Engine (TIDE) to support New Hampshire's new statewide assessments. TIDE is the same system that New Hampshire Department staff, district test coordinators, school test coordinators, school teachers, and school test administrators currently use as they prepare and engage in administering Smarter Balanced interim and summative assessments. TIDE provides an integrated system for

- gathering and managing student enrollment and pre-ID labels;
- adding, editing, and deleting users (e.g., district or school test coordinators) and granting them certain pre-defined authorizations based on a role hierarchy;
- monitoring testing progress; and

- role-based access to add, edit, or view student accommodations, interim assessment assignments and the creation or modification of class rosters.

TIDE will be available during the entire testing window. Throughout the testing window, demographic and other data can be reviewed and corrected in TIDE, speeding the process of data cleanup and improving reporting timelines.

Currently, AIR supports the enrollment of New Hampshire students through a nightly file exchange using i4see. This nightly exchange allows AIR to automatically import student demographic and accommodation information into TIDE or make updates to existing information in TIDE. AIR is prepared to continue this method to import student information into TIDE and is open to exploring other options to upload student information. The familiarity New Hampshire users have with the TIDE interface and the student registration process will allow for a seamless transition for student registration for the new assessments.

Import between AIR and the Department Data Systems

TIDE supports a variety of file import formats and interfaces that can be customized to integrate with the Department's student information systems.

These data systems can send AIR flat files that are either delimited (e.g., CSV), fixed-width, SIF, IMS, or custom XML formats. TIDE imports have built in a level of efficiency that allows for the import of incremental or complete data files every day or at any other desired import frequency.

Files are typically deposited by the state systems to a monitored secure file transfer protocol (SFTP) location. In addition to the typical authentication requirement, access to this location is restricted to a limited number of IP addresses.

Every import can be configured with its own set of validation and processing rules. After each import, import summaries, validation messages, and processing errors can be sent via e-mail to a specified set of users. Before any file is processed, it is sent through a data validator that identifies any errors present in the records. A fail-safe can be configured to reject a file if the number of errors, updates, or deletions exceeds a predetermined threshold set by the Department.

TIDE supports importing information about institutions, students, users, test settings, class groups and rosters, and test assignments. The import can take place either through automated data exchanges between the Department and AIR or by allowing state, district, or school users to upload data directly into TIDE. TIDE's user interface also allows for ad hoc updates to student information.

Setting Up a Network of Schools and Districts

TIDE supports adding and editing school and district information in a variety of file formats such as CSV, Excel, or fixed format. TIDE can be configured to support custom school-district hierarchies and can also support different school type designations such as public, private, and charter schools. TIDE supports customized attributes for schools and districts, including name, institution identifier, the National Center for Education Statistics (NCES) ID, local ID, and type. The Department may provide a list of schools and districts to TIDE at a predetermined frequency. TIDE will load this information, which subsequently forms the basis for associating every user and student added to TIDE with a specific institution and school-system hierarchy. AIR's suite of online testing systems interacts with the same list of schools and districts and establishes boundaries in the system to allow or limit access to data based on the association of a specific user with specific schools or districts.

AIR will work with New Hampshire to establish a user role hierarchy. Based on this hierarchy, TIDE can be configured so that users at higher levels can delegate tasks of adding users to other roles. A typical setup follows:

- State-level users can upload district-level users.
- District-level users can add school test coordinators.
- School test coordinators can add teachers and test administrators.

User enrollment is discussed in greater detail in Topic 4 Technology Requirements.

Student Registration

TIDE supports uploading student information using either automated data exchanges between the Department and AIR or allowing the state, district, or school users to upload student information directly into TIDE. Students can also be manually added by designated user role in TIDE's user interface.

In each case, TIDE can work with the Department to establish validation rules and configurable error messages. Students added in TIDE are reflected in real time, along with their test eligibilities, in the other online systems like the test delivery system (TDS) and online reporting system (ORS). This ensures that any students who are enrolling at the last minute can take tests immediately following registration.

Automated Upload from New Hampshire's Data System

TIDE can acquire an initial list of students from New Hampshire's student data management system. AIR will establish an SFTP location where the Department can place a student file at the start of the testing administration or at predetermined intervals.

TIDE services will validate the file present in the folder in the SFTP location and import records from the file. TIDE can also be configured to provide an import summary with validation and error logs, and the frequency distribution of the imported data. This information can be e-mailed to a selected set of recipients or can be placed in a secure location that is accessible to the Department. The file layout used for the student data upload is configurable and can be updated to account for changes in student data attributes from year to year.

Currently, AIR supports the enrollment of New Hampshire students through a nightly file exchange using i4see. This nightly exchange allows AIR to import student demographic and accommodation information into TIDE automatically. AIR is prepared to continue this method to import student information into TIDE and is open to exploring other options to upload student information. Students are associated with districts and schools in TIDE so that associated users can access student information in their jurisdiction. The existing file exchange between AIR and the Department allows for updates to be made to student enrollment information overnight if necessary to account for students who enroll in other schools or move to a different school or district after the initial registration window has closed. This plan allows the Department to review and amend registration information.

We propose the collection of the institution file to be uploaded into TIDE approximately one month before the TIDE system go-live date and the collection of the initial student file approximately two weeks before the TIDE go-live date. The institution file will contain all the district and school information, whereas the student file will contain all the student demographic and enrollment data. AIR will work with the Department to ensure that accurate information is captured and appropriate quality checks occur and to provide the Department with sufficient time to review and amend the registration information as needed.

AIR will work with the Department to include and accommodate the registration of any student in private out-of-school district placements, home education, or private school environments that wish to participate in the statewide assessments. AIR will work with the Department to determine a method to identify these students' registration separate from public school registrations for accountability purposes.

Web-Based Student Upload

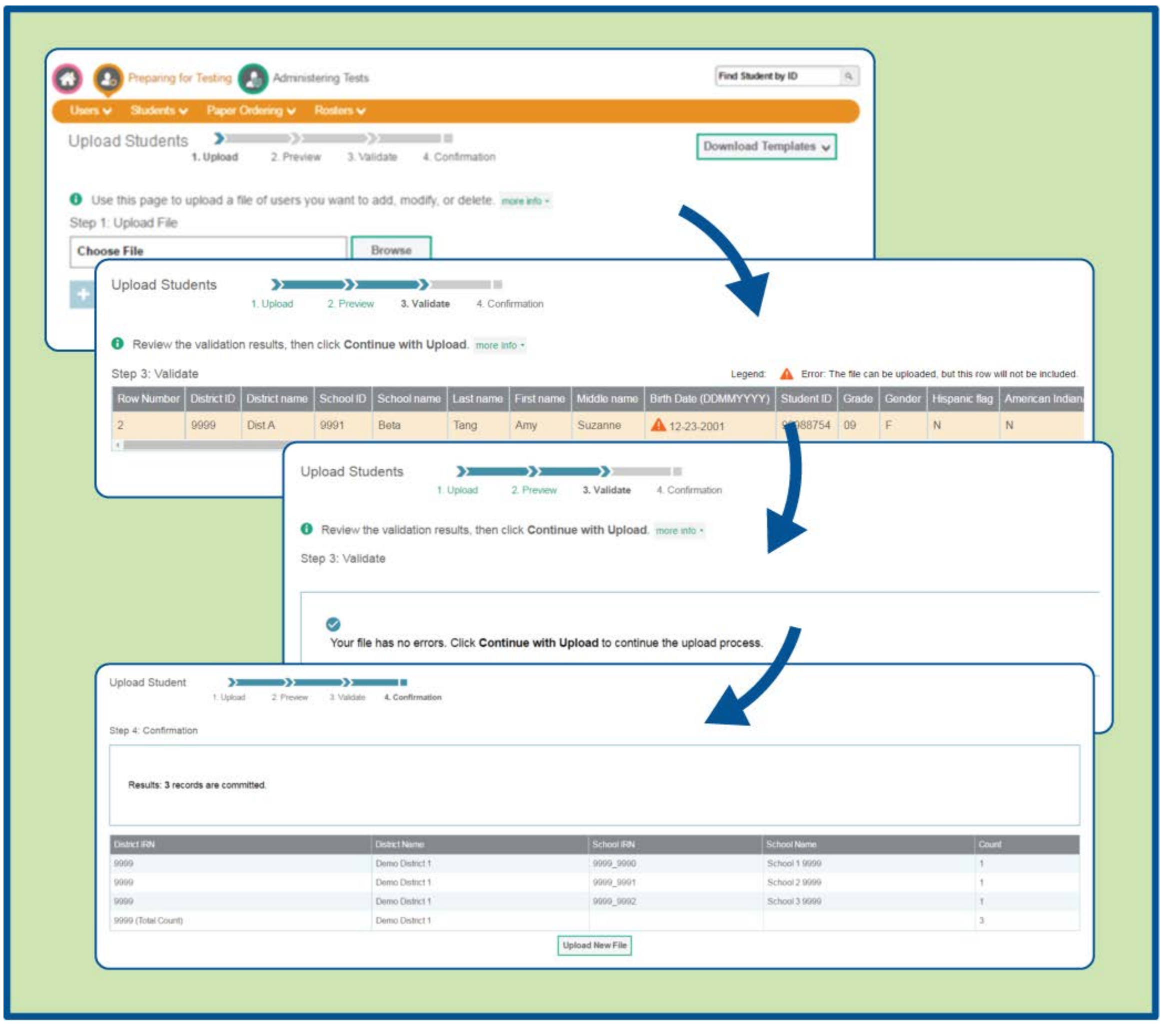
Once the initial file of student records is loaded, TIDE can be configured to support the upload of student information by state-, district-, and school-level users. The TIDE website has a step-by-step interface that guides the user through the process of securely uploading student information. TIDE allows users to identify a file on their computer or network that is in the agreed-on format. Once the file is identified, the system scans it to ensure that the data match the format and to allow the user to preview a few records of the file being uploaded. The file then goes through validation to ensure that its data conform to the business rules set in TIDE. Any data format or validation errors are reported to the user in real time.

Validations of the files are accomplished through configurable business rules. Project staff will collaborate with the Department to define the validation business rules for New Hampshire. These rules can trigger one of three outcomes:

1. Rejection of the entire file
2. Rejection of only the offending records
3. Warning message (the file may be committed, with noted issues)

In the third step, the user commits the file to the database and TIDE provides any errors that occurred during the upload in real time. The process of uploading student information is illustrated in Exhibit D1.1-5.

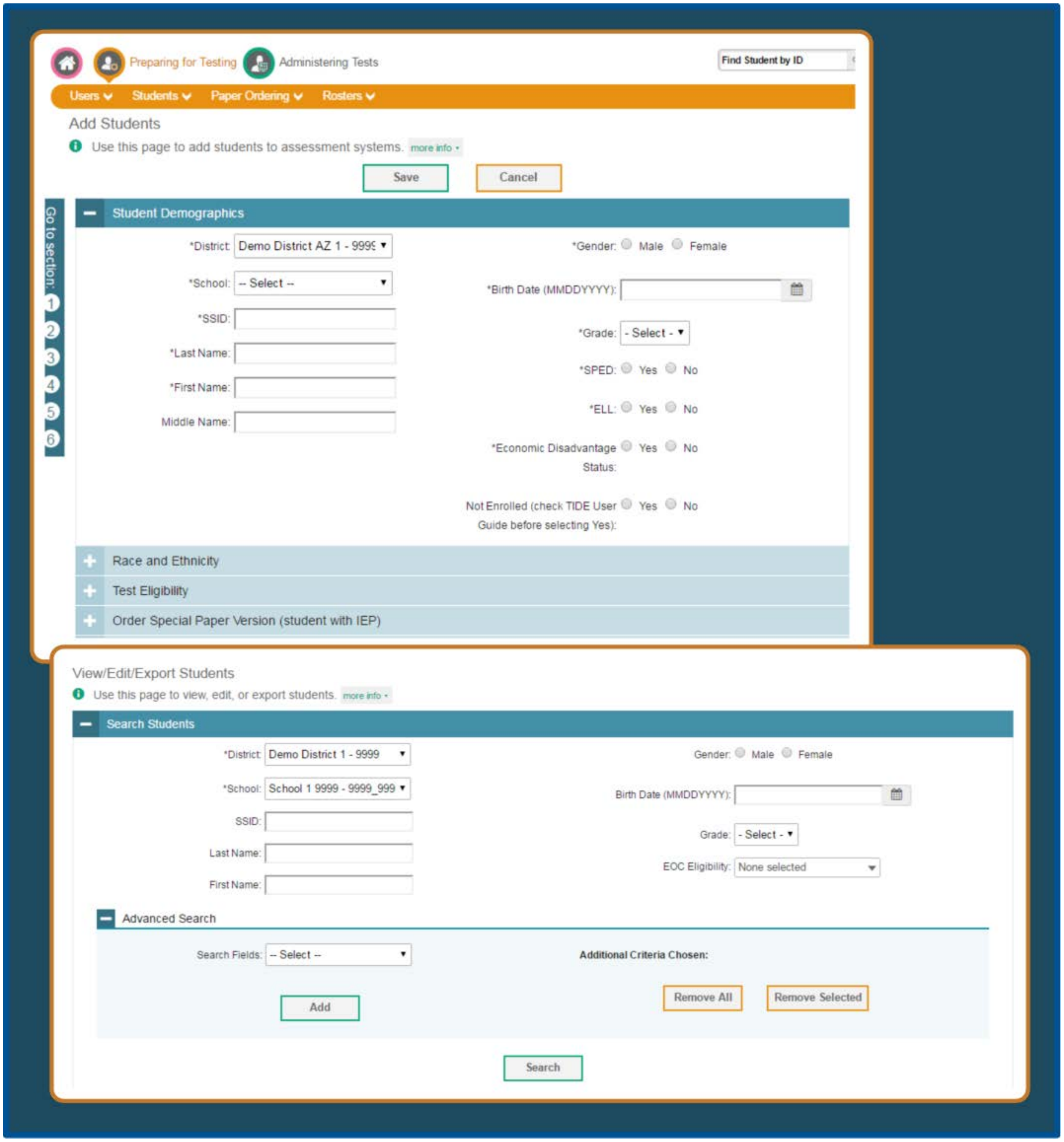
Exhibit D1.1-5: Student Upload Process



Manually Adding and Editing Students

TIDE's record change system allows users to review student information and update it reliably and efficiently before, during, and after an assessment is taken. TIDE's manual addition and editing features are shown in Exhibit D1.1-6.

Exhibit D1.1-6: Manually Adding and Editing Student Information



Changes made in TIDE are reflected immediately in each integrated online system component (the TDS and the ORS).

AIR's TIDE system is completely configurable in terms of what attributes it can support. During the specifications and requirement gathering process, AIR will work with the Department to identify all relevant attributes and demographic information that need to be collected for students. TIDE can collect any additional attributes other than the ones listed:

- Gender
- Race/ethnicity
- English language learner
- Economically disadvantaged (free/reduced meals)
- Students with disabilities/individual education program (IEP)
- Specific testing accommodations (flags for which accommodation supports were used)
- Foster students
- Military dependents
- Homeless students
- Migrant students

TIDE can be configured with business rules to validate the data being imported. If errors occur during import, TIDE can notify users and can also provide a detailed validation/error report. The Department can provide additional subgroups to further disaggregate the data. Reports can be generated in the ORS based on these subgroups.

Should the Department's system not support certain attributes, they can be collected on the TIDE user interface as well. TIDE can be configured to allow only users with certain access rights or roles to edit any information.

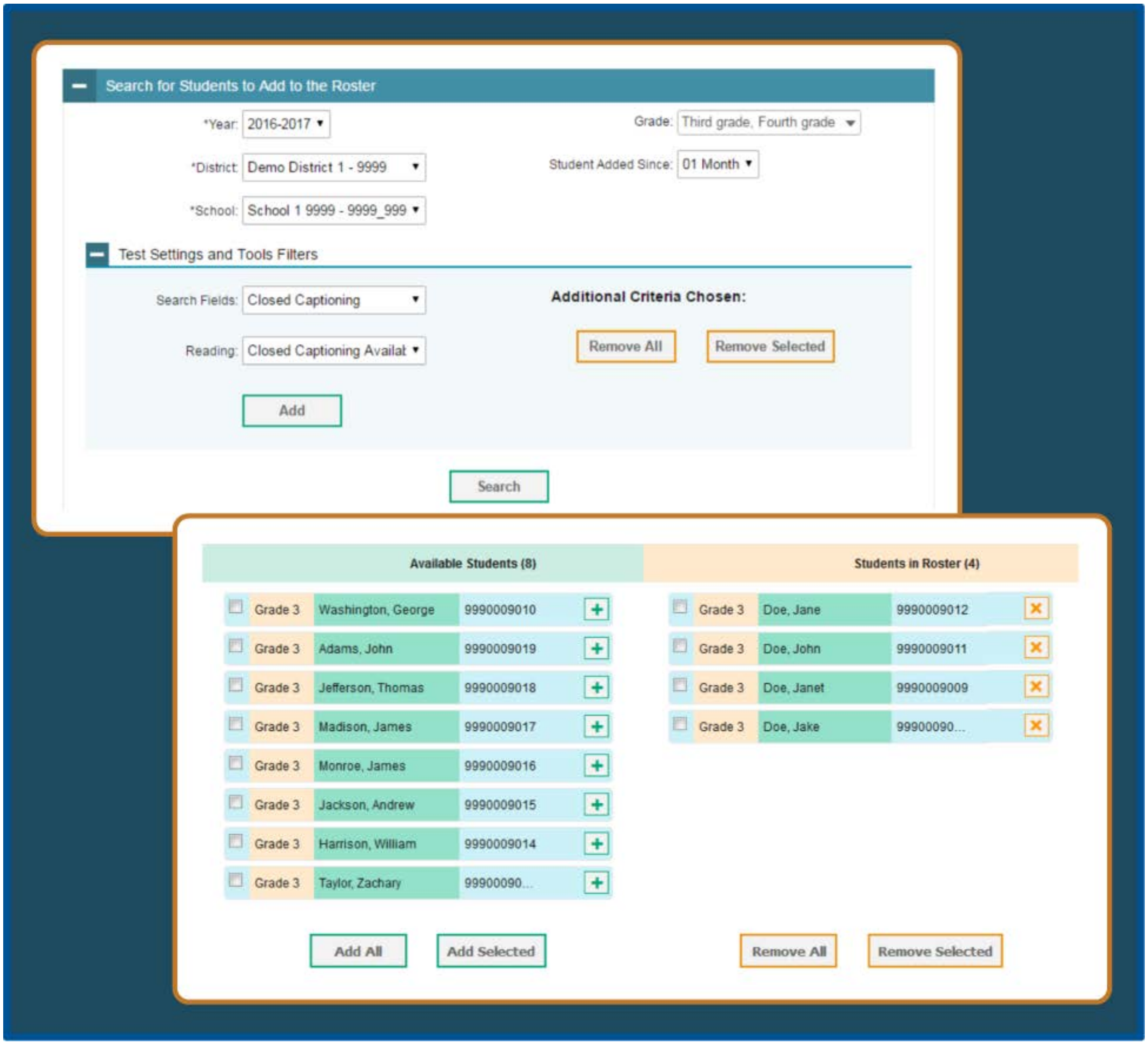
- Validation rules are also present in the manual interface and provide helpful validation messages to the user. TIDE can be configured to perform multiple validations on each data element being added or uploaded.
- Validation rules prevent district users or test coordinators from editing information for students who are not present in their district or school.
- TIDE role-based access limits which roles may view or edit student demographic information.
- During the data upload process, multiple districts may attempt to claim the same student. TIDE can enforce business rules that prevent a district from claiming a student until the student is released by his or her current district. TIDE can also allow the most recent claim to hold. More specialized business rules can be implemented, as needed.

In addition to the ability to add and edit student records, TIDE also provides robust search functionality that allows state, district, and school users to search for students based on different attributes, including demographic information, test settings, and test eligibilities. Business rules exist to limit the search function to certain user roles and to limit the search results in the user's specific jurisdiction (a particular school or district). Users have the option of easily exporting the full list or a subset of student record search results to a CSV or Excel file. Exported data can have a configurable set of columns that can include but is not limited to a student's ID, grade, school, district, and name.

Electronic Grouping of Students

TIDE can import the class assignment/course enrollment (electronic grouping) information for students from the Department files or allow users to create groups of students through file uploads or through the user interface (UI). The feature can be limited to certain user roles. Roster changes made in TIDE are reflected in the other online systems like the ORS, as shown in Exhibit D1.1-7. Based on these groups, the ORS can provide aggregate information. The ORS can also provide aggregate information for a school, a district, or the entire state.

Exhibit D1.1-7: Manage Rosters



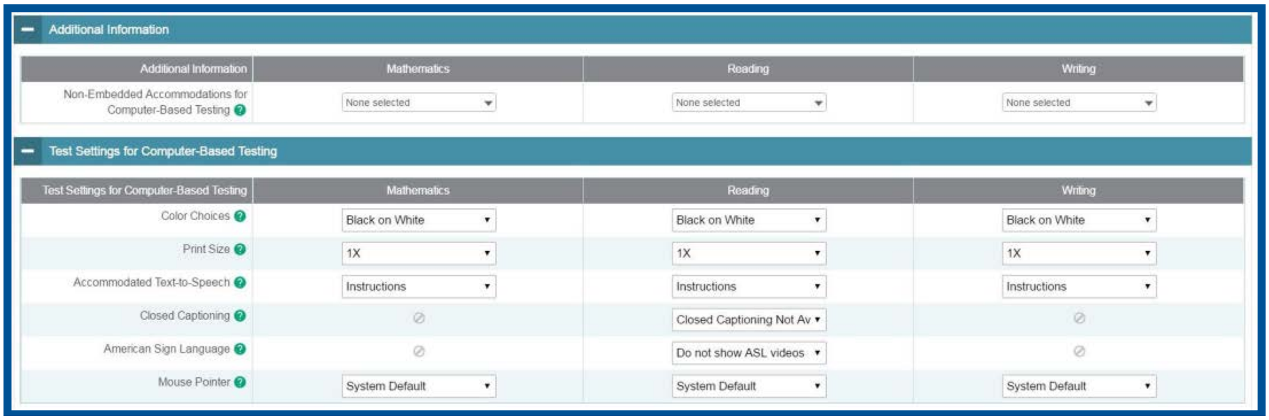
Student Test Settings

Users can either choose to upload student test setting information into TIDE or use the interface shown in Exhibit D1.1-8 to edit the test setting information directly in TIDE. The test setting features shown on this screen are linked directly with the testing features of the online testing system and are configurable based on the selections made by the Department.

The user interface can be configured to allow only a certain set of user roles to have read/write access to specific test settings. The robust search function in TIDE can be configured to search for students based on different attributes, including test settings, accommodations, and test eligibilities, and it allows easy export of search results so that users can verify student test settings before testing. The search feature also has business rules to limit the search function to certain user roles and to limit the search results in the user's jurisdiction (school or district). Search results can be readily exported to CSV or Excel, and users have the option to export the entire set or just a selected set of student records. The online testing system can be configured to override some test settings as needed on the day of testing. Because the TIDE system is integrated with all other online systems, including the TDS and the ORS, edits made in TIDE are reflected immediately across all systems.

Student test settings and accommodations are addressed in greater detail in Topic 1.4, Accessibility and Fairness.

Exhibit D1.1-8: Editing Test Setting Information



Additional Information

Additional Information	Mathematics	Reading	Writing
Non-Embedded Accommodations for Computer-Based Testing	None selected	None selected	None selected

Test Settings for Computer-Based Testing

Test Settings for Computer-Based Testing	Mathematics	Reading	Writing
Color Choices	Black on White	Black on White	Black on White
Print Size	1X	1X	1X
Accommodated Text-to-Speech	Instructions	Instructions	Instructions
Closed Captioning		Closed Captioning Not Av	
American Sign Language		Do not show ASL videos	
Mouse Pointer	System Default	System Default	System Default

Student Test Assignment

TIDE can use student demographic information such as grade and/or test settings to assign tests to a student automatically. Business rules can be configured to account for cases like accelerated testing, parent test exemptions, and special test codes to override the default eligibilities. In addition to the automated test assignments, TIDE can import explicit test assignments from state files, user file uploads, or individual test assignments through the TIDE user interface. Such access can be limited to certain user roles. Test assignments can be searched, exported, or printed as needed. Test assignments made in TIDE are reflected in real time in the online testing system.

Similarly, TIDE can import form assignments for particular tests, or forms can be spiraled in real time to ensure even distribution. TIDE can also assign forms based on the test settings of the student. For example, if the student has a Braille accommodation, TIDE will ensure that the Braille form is assigned for this student and other students will have their forms spiraled.

Paper-Pencil Tests

TIDE can also identify which students need to be administered paper-pencil tests based on demographics or through information received from the state or users. This feature can be used to automatically calculate material counts and generate paper-pencil test orders. TIDE also takes into account the test settings, such as large print, for the student and orders the appropriate form for that student, allowing users to provide explicit material counts for each school or district. Material orders can include configurable overage rules, rounding rules, and approval rules. This feature can be limited to certain user roles.

TIDE supports multiple ordering windows (e.g., initial and additional orders) and can be configured with different business rules and materials for each window. Orders generated from TIDE can be tracked for outbound and inbound shipment.

Ordering Materials

TIDE supports the traditional ordering model where users can provide material counts during the order window. TIDE can also preload quantities for its order window based on the number of paper tests taken the prior year. In addition to these methods, data collected in TIDE are used to determine student test eligibility for online and paper tests. TIDE can identify which students need to be administered paper-pencil tests based on demographics or through information received from the Department or users. This feature can be used to calculate material counts and generate paper orders automatically. TIDE also takes into account the test settings, such as large print or Braille, for the student and orders the appropriate form for that student, allowing users to provide explicit material counts for each school or district.

Material orders can include configurable overage rules, rounding rules, and approval rules. This feature can be limited to certain user roles. TIDE supports multiple ordering windows (e.g., initial and additional orders) and can be configured with different business rules and materials for each window. Orders generated from TIDE can be tracked for outbound and inbound shipment.

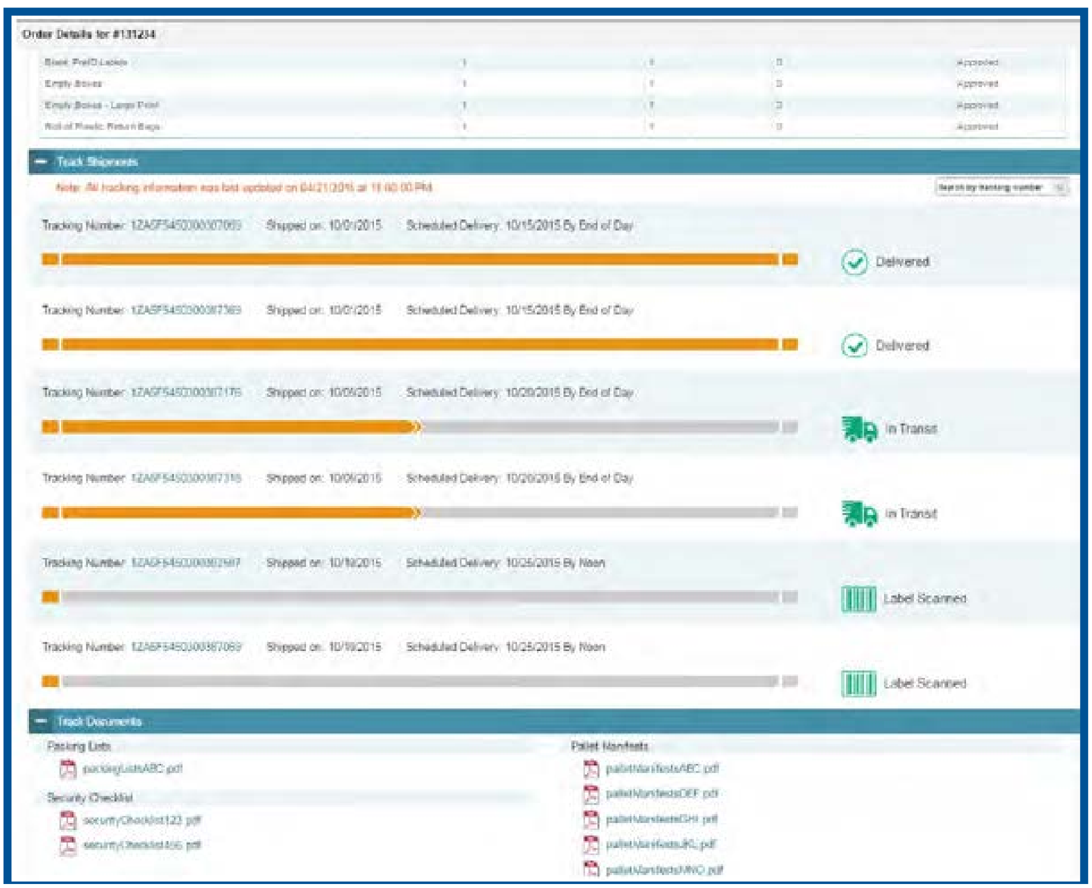
The Orders module has a check to identify whether a district is participating in paper testing. With this check, even if materials are preloaded, districts have the option of not receiving these materials. Users can enter at a later point and confirm participation in the paper testing window. They can do this even one day before testing begins and then order materials during the additional order window.

The Orders module in TIDE has a provision where test coordinators can provide the shipping addresses for materials and the contact information of school staff who will be receiving these materials. While this information can serve as the default contact information for shipment, the Orders module also supports overrides for individual orders so that they can have a different shipping address.

TIDE provides end-to-end integration of the Orders module with the shipping companies so that users placing orders can verify the status of the order, its tracking information, and any associated packing and security checklists.

Nightly, TIDE even checks the status of individual tracking numbers with the shipping companies and thereby provides the status of all tracking numbers on a single page, as shown in Exhibit D1.1-9.

Exhibit D1.1-9: Order Details



Real-Time ID

AIR is introducing a new service: Real-Time ID. New Hampshire schools and districts will be able to register students only one or two days before testing for both online and paper-pencil testing to accommodate newly entering students. Real-Time ID speeds reporting by eliminating most of the post-testing cleanup that occurs when schools must hand-bubble information for late-arriving students.

Real-Time ID, a process introduced two years ago and proven successful in Arizona, Florida, and Ohio, extends some of the efficiencies of testing to paper testing. Rather than the tedious and error-prone process of marking bubbles on unassigned booklets, TIDE allows the printing of scannable labels using any laser printer. For students who enter the school the day before testing, a test administrator can enter the student in TIDE, print a label immediately, and affix that label to the test booklet instead of gridding the data by hand. Subsequent updates and corrections to the student information are performed directly in TIDE, making the corrections process much more automatic and reliable.

Real-Time ID will supplement pre-ID, not replace it. Pre-ID or enrollment uploads may come directly from the Department or from individual districts at the Department’s discretion. If New Hampshire can collect near-real-time data from districts, central delivery will help ensure the integrity of New Hampshire’s statewide student information system and drive data improvement. If the mechanisms are not in place, we can accept uploads directly from districts.

Monitoring Test Progress

An effective tool that keeps educators informed about the progress of student testing is TIDE’s Monitor Test Progress feature. This feature allows users to generate various reports, including the following:

- *Plan and Manage Testing Report* details all of a student’s test opportunities and the status of those test opportunities.
- *Test Completion Rates Report* summarizes the number and percentage of students who have started or completed a test.
- *State Participation Counts Report* shows, at the state level, how many students started or completed a test on a certain day, and provides cumulative counts for the current test administration.

Topic 1.4 Accessibility and Fairness

AIR has a long-standing commitment to administering fair and valid assessments for students with disabilities and to students who are English language learners (ELLs). We have been committed to accessibility for all students ever since we conceptualized and designed our Test Delivery System (TDS) and the item types administered in TDS. We continue to make accessibility a top priority as we build more innovations around student testing and learning. We are committed to 100% accessibility for all students, and no existing system is more accessible than ours, and only our system has been proven in real statewide accountability tests.

Each tool or embedded support provided by AIR’s system, by design, may be configured to be available to all students; available to students by one or more user roles in the system (e.g., state-level NH DOE staff, district test coordinator, school test coordinator, teacher, or test administrator [TA]); and offered to all students but configured for each student by adult users. The system allows personal needs and accommodation information to be entered in advance of testing through TIDE, which we generally recommend, or by authorized individuals in the Test Administrator Interface at testing time (at the discretion of the Department). Exhibit D1.1-10 shows many of the configurable accessibility features of our system, including the configurations of New Hampshire’s Smarter Balanced assessments in ELA and mathematics for the 2016–2017 test administration. As noted in Topic 14 Practice Test and Student Materials, accessibility features are available in the practice and training test environments to increase student familiarity with the available accessibility features prior to interim and summative testing.

However, students who have taken assessments that are delivered on AIR's testing platform in past years already have familiarity with the Student Interface and available accessibility features.

Exhibit D1.1-10: Configurable Accessibility features of AIR's Systems

	AZ	CA, ID	MT, SD, USVI	NH, VT	CT	DE	ELPA21	FL	HI	IA	ND	OH	OR	UT	WA	WV	WY
American Sign Language	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Audio Playback Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Braille Type		3	3	3	3	3			3		3		3	2	3	3	
Calculator	2	3	3	3	3	6		1	5		3	2	4	4	4	3	
Closed Captioning	✓	✓	✓	✓	✓	✓			✓		✓		✓		✓	✓	
Color Choices	6	5	5	5	5	14	6	8	7	6	5	10	10	6	7	5	
Dictionary	3	3	3	3	3	3			3		3		3	3	3	3	
Emboss		✓	✓	✓	✓	✓			✓		✓		✓	✓	✓	✓	
Emboss Request Type		2	2	2	2	2			2		2		2	2	2	2	
Expandable Passages	✓	✓	✓	✓	✓	✓	✓	✓	2	✓	✓	✓	✓	✓	✓	✓	
Font Type	✓	2	2	2	3	3	2	✓	4	2	2	3	5	✓	3	2	✓
Formula						2		8	✓			6		2		✓	
Global Notes	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Hardware Checks	✓	✓	✓	✓	✓	✓	✓	✓	2	✓	✓	✓	2	✓	2	✓	
Highlight	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
IntraTest TTS Controls	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	
Item Font Size	2	✓	✓	✓	2	2	3	✓	2	3	✓	4	3	✓	3	✓	✓
Item Tools Menu	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Language	✓	3	3	3	3	3	2	2	4	2	3	3	3	2	3	3	✓
Line Reader	✓				✓		✓	✓		✓		✓	✓		✓		
Mark for Review	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Masking		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Mute System Volume		3	3	3	3	3			3		3		2		3	3	
Paginate Item Groups	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Passage Font Size	2	✓	✓	✓	2	2	3	✓	2	3	✓	4	4	✓	3	✓	✓
Periodic Table					✓	✓						✓	✓			✓	
Permissive Mode		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	
Print on Request		3	3	3	3	3	3		3	3	3	3	3	✓	3	3	
Print Size	5	4	4	9	9	9	9	5	4	5	4	9	5	5	4	4	
Review Screen Layout	✓	✓	✓	✓	2	✓	✓	✓	✓	✓	✓	2	✓	✓	✓	2	✓
Streamlined Mode		✓	✓	✓	✓	✓			✓		✓	✓	✓	✓	✓	✓	
Strikethrough	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Student Comments	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
System Volume Control	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Test Notifications	✓	✓			✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Test Opportunity Indicator	✓						✓			✓		✓	✓		✓		
Test Progress Indicator		✓									✓	✓					
Test Survey		✓							✓						✓		
Thesaurus	✓	✓	✓	✓	✓	✓			✓		✓		✓	✓	✓	✓	
TTS	2	4	5	5	3	4		✓	7		5	✓	4	2	5	5	
TTS Audio Adjustments	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	
TTS Tracking	✓				✓			✓				✓		✓			
TTX Business Rules	✓	2	2	2	2	2		✓	3		3	✓	3	2	2	2	
Tutorial	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Video Playback Controls		✓					✓			✓		✓	✓		✓		
Word List		22	22	22	22	22			22		22		22		22	22	

Exhibit D1.1-11 shows examples of test settings that may introduce security risks and AIR's recommendations on mitigating the risk.

Exhibit D1.1-11: Mitigating Security Risks of Test Settings

Test	Description	Mitigation of Security Risk
Permissive Mode	■ Allows a student to access accessibility applications (e.g., Point N See on OSX, Onscreen Keyboard on Windows) that he or she uses on a regular basis. ■ Permissive Mode is a Windows and OSX feature only.	■ This is provided as an accommodation only (not a universal tool).
Print-on-Demand	■ Allows a student to request an item or passage to be printed. That request is sent over a secure channel to the TA Interface, at which point the TA can approve or deny such request.	■ This is provided as an accommodation only (not a universal tool). ■ It requires that the student's request to print be approved (or denied) by the TA overseeing that testing session. ■ Only a TA can print the request out. ■ Since TAs typically are not allowed to view test items, this feature restricts what they see through the print preview and the print output. ■ All print requests have a cover page with a security warning, and each corresponding page has a header containing information about the student, the test, and the item. None of the test content is displayed on print preview. ■ Ensuring that secure materials are properly handled and discarded can be done through strict test administration procedures.

Tools, Supports, and Accommodations

AIR's TDS offers the industry's most robust set of operational features. In this section, we describe the tools and accessibility features available.

AIR's system offers a broad array of tools, including calculators, formula sheets and other resources, notepads, and measurement tools. Exhibit D1.1-12 shows our straightedge, ruler, compass, and protractor. As with the rest of our system, these tools are entirely built in HTML5 and can be completely controlled by the keyboard. We believe these HTML5- and keyboard-controlled tools are unique within the industry.

Exhibit D1.1-12: Test Delivery System Measurement Tools

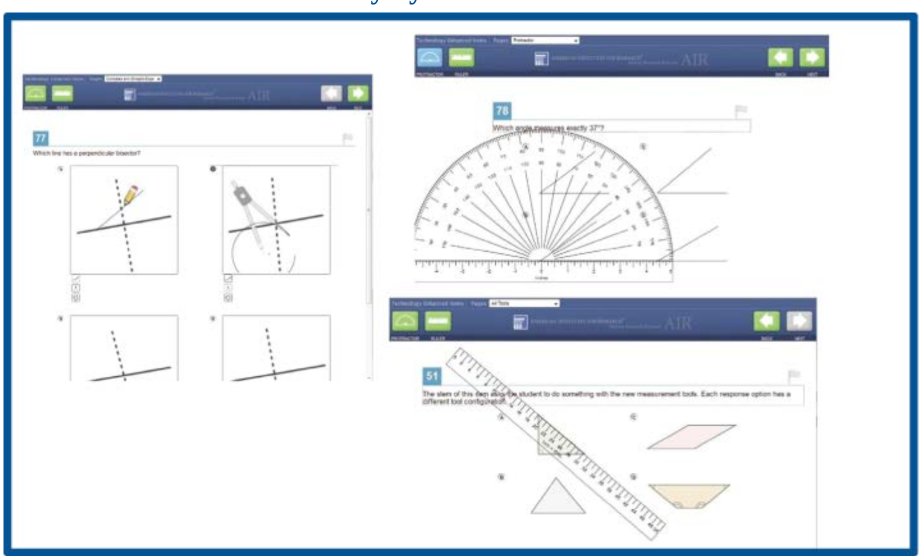


Exhibit D1.1-13 offers a more complete listing of our available tools.

Exhibit D1.1-13: Summary of Available Tools

Tools	Description	AIR System Capability and Comments
Calculator	Available calculators include basic, scientific, graphing, regression, and matrices.	Currently Available
Reference Sheet	Students can access a formula sheet and periodic table for mathematics or science tests.	Currently Available
Ruler	Students can access a pop-up ruler.	Currently Available
Straight Edge	Students can access a pop-up straight edge.	Currently Available
Option Eliminator	Students can eliminate any multiple-choice option, whether it is in text or a graphic. This capability persists throughout the test.	Currently Available
Bookmark	Bookmark capability allows students to flag an item so they can review it.	Currently Available
Highlighter	Students can select any text to highlight. This capability persists throughout the test.	Currently Available
Notes	Item notes allow students to jot down ideas about items or passages.	Currently Available
Footnote Pop-up	Our word list feature supports footnote pop-ups.	Currently Available
Student Tutorials/ Practice Sessions	A reference feature, practice tests, and tutorials familiarize students with the online testing system. Our tutorials can vary by item, so items of each type share a tutorial.	Currently Available
Spell Check	The spell check tool identifies words in the response field that may be misspelled, and provides suggestions for each.	Currently Available
Response Field Tools	The response field tools allow students to apply styling to text (e.g., **bold**, *italics*, underline) and use standard word-processing features such as moving and indenting text, cut, redo, and paste.	Currently Available
Embedded Dictionary and Thesaurus	Students can open the Merriam-Webster dictionary and thesaurus within the test. This tool is available during the second segment of ELA performance task tests.	Currently Available
Global Notes	Students can access a notepad throughout the test. This notepad allows students to enter notes for themselves and is not item-specific.	Currently Available
American Sign Language (ASL)	This capability consists of recorded videos on sign language.	Currently Available
Refreshable Braille/Tactile With External Embosser Printer	Items can be rendered to desktop embossers that can integrate Braille and tactile graphics. The items will simultaneously render on a reader-accessible screen, and the student will be able to navigate to response spaces to provide answers.	Currently Available

It makes sense to configure some tools by item, such as the measurement tools in Exhibit D1.1-12. Other tools, such as calculators, must be configured in separate test "segments." Our segmentation capability allows students to complete items in one segment, review them, and indicate completion before moving on to the next segment. Once a segment is completed, the student cannot return to it; this capability would prevent the student from, for example, accessing a calculator when its use is prohibited.

Accommodations and Embedded Supports

AIR has a long-standing commitment to administering fair and valid assessments for students with disabilities and to students who are English language learners (ELLs). We have been committed to accessibility for all students ever since we conceptualized and designed our TDS and its items. We continue to make accessibility a top priority as we build more innovations around student testing and learning.

AIR's TDS offers a robust array of accessibility features, as illustrated in Exhibit D1.1-14. AIR is proud not only that we offer these features but that virtually all are in operational use right now in states for which we are providing online testing.

Exhibit D1.1-14: Accommodations and Accessibility Supports

Accessibility Feature	Description	AIR System Capability and Comments
Text-to-Speech—Directions, Passages Items	Computer reads text and graphics aloud on directions, passages, and items. What is read and how it is read is configurable. The Department can offer a variety of options.	Currently Available
Text-to-Speech—Graphic Description	Computer reads descriptions of graphics and content of tables aloud.	Currently Available
Magnification Interface	Student can zoom in and zoom out on the entire page. This capability persists throughout the test.	Currently Available
Magnifier	Student can magnify a selected portion of an item.	Available Through Third-Party Software
Variable Font Size	The number of levels (generally, five levels) and rate of increase (generally, 1.25x the previous level) are configurable.	Currently Available
Refreshable Braille/ Tactile With External Embosser Printer	Items can be rendered to desktop embossers that can integrate Braille and tactile graphics. The items will simultaneously render on a reader-accessible screen, and the student will be able to navigate to response spaces to provide answers.	Currently Available
Reverse Contrast	Display colors are altered so that the background is black and the text is white.	Currently Available
Administrator-Selectable Variable Font and Background Colors	Any foreground and background color can be supported.	Currently Available
Color Overlay	Any color can be laid on the screen. This persists throughout the test.	Currently Available
Increased White Space	This is the streamlined interface.	Currently Available
Sign Language—Directions, Passages, Items	This capability consists of recorded videos of test directions and content using sign language. Avatars are not recommended by experts on hearing-impairment because they do not translate well to American Sign Language.	Currently Available
Translations	Versions are available in alternate languages.	Currently Available
Keyword Translation	This enables translators to associate keyword translations.	Currently Available
Glossaries and Dictionaries	These enable content developers to associate additional content with words or phrases. The content can be of multiple types, and the content shown to a student can be controlled by his or her personal profile.	Currently Available
Alternate Language Glossaries and Dictionaries	These enable content developers to associate alternate-language content with words or phrases. The content can comprise multiple types, and the content shown to a student can be controlled by his or her personal profile.	Currently Available

Exhibit D1.1-14: Accommodations and Accessibility Supports (continued)

Accessibility Feature	Description	AIR System Capability and Comments
Administrator-Selectable Assistive Devices Integration	Our system has a standard interface and a streamlined interface. Most assistive devices can work with the former, and an even wider group works with the latter. If the use of the device requires relaxation of certain security features (e.g., if suppression of pop-up windows interferes with on-screen keyboards), the administrator can select our more "permissive mode" if the Department chooses to offer this feature.	Currently Available
Line Reader	This feature will allow a student to track the line that he or she is reading.	Currently Available
Masking	Students can mask extraneous information on the screen.	Currently Available
Speech-to-Text	Speech will be converted to text and then saved in the database.	Available Through Compatibility With Third-Party Assistive Technology
Auditory Calming	A tool that plays music or white noise in the background.	Available Through Third-Party Software
Administrator-Selectable Zoom	Default font size can be set in advance through a file upload or user interface or at the time of testing by the test administrator. Student can zoom in or zoom out at any time.	Currently Available
Administrator-Selectable Large Print Font	Default font size can be set in advance through a file upload or user interface or at the time of testing by the test administrator. Student can zoom in or zoom out at any time.	Currently Available
Administrator-Selectable Screen-Reader	The system supports an integrated screen reader that can be configured to provide a variety of support levels, each selectable by the administrator.	Currently Available
Additional Time	AIR's system currently does not impose a time limit on the test. It is up to the proctor to stop a student's test or stop the entire session. However, if there are unforeseen events, such as a fire alarm, that trigger additional testing time, AIR's system can enable a grace period extension (GPE) for a single test opportunity or for multiple test opportunities.	Currently Available
Segment Breaks	AIR's system has the capability of adding test segments within a test. A test segment is made up of multiple item groups and creates a logical break between segments within a test. For example, a segment break might separate a calculator segment and a non-calculator segment on a math test.	Currently Available
Recorded Audio	Computer efficiently delivers recorded audio. We are able to deliver voice-audio using only about 10 Kbps of bandwidth.	Currently Available
Secure Print Facility	A visual accessibility feature, the secure print facility allows the secure printing of items or passages. A student requests that a passage or item be printed, the request is then encrypted and sent securely to the proctor, and the proctor needs to approve the request before it is sent to the printer. In addition, this feature also allows for the delivery of real-time paper tests, including large-print tests.	Currently Available

Exhibit D1.1-14: Accommodations and Accessibility Supports (continued)

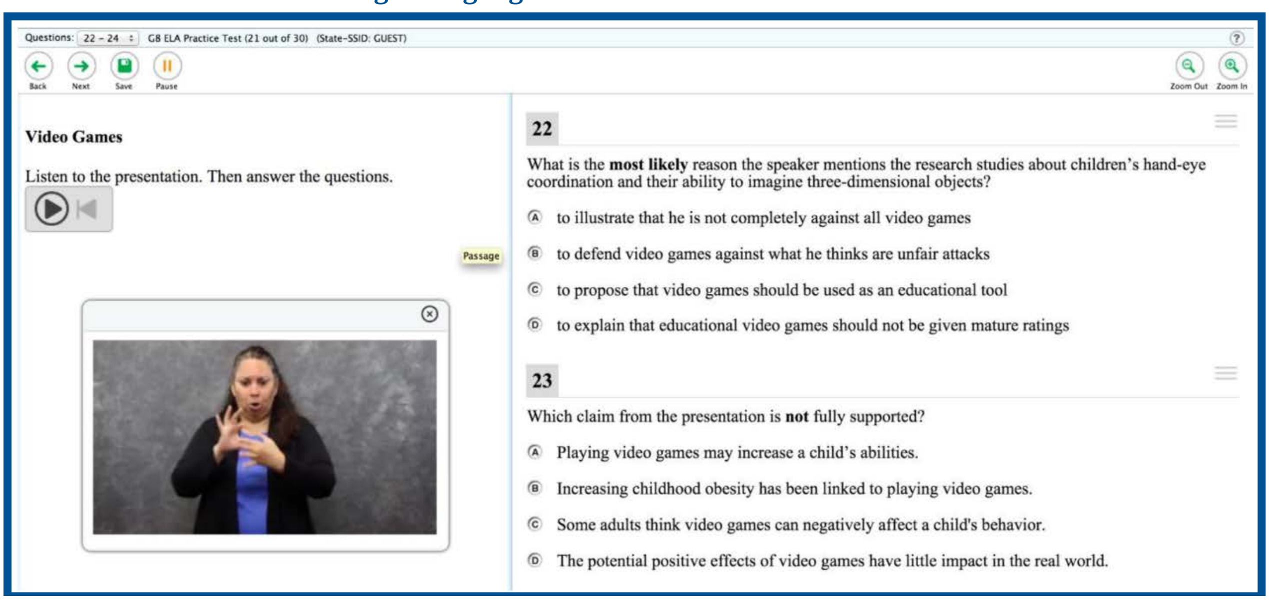
Accessibility Feature	Description	AIR System Capability and Comments
Test Pauses and Restarts	An attention accessibility feature, test pauses and restarts, allows the test to be paused at any time and restarted and taken over many days. So that security is not compromised, visibility on past items is not allowed when the test has been paused longer than a specified period of time.	Currently Available
Writing Checklists	An attention accessibility feature generally for essay items, the writing checklist enables a student to check off writing guidelines from a checklist.	Currently Available
Review Test	Students can review the test before ending it.	Currently Available
Area Boundaries	An agility accessibility feature, area boundaries for mouse-clicking multiple-choice options allow students to click anywhere on the selected response text or button.	Currently Available
Language	Any language that is necessary can be supported.	Currently Available
Help Section	A reference feature, the Help Section explains how the system and its tools work.	Currently Available
Performance Report	A reference feature, a performance report is available at the end of the test for the student.	Currently Available

Accessibility Features

In this section, we would like to highlight a few accessibility features that we have deployed recently. In 2013, AIR introduced individualized item pools that take each student's access restrictions and produce a custom filter on the pool in real time when items are being selected. This feature is most relevant to adaptive tests and helps demonstrate the depth of our commitment and capability to serving all of New Hampshire's students.

We note that we support American Sign Language (ASL) through recorded video rather than avatars. Working with representatives of the deaf community, we learned that ASL is not simply a representation of English. Rather, it is a unique language. Avatars are not capable of translating from English to ASL and instead provide a word-for-word translation. To speakers of ASL, this seems as unnatural as a word-for-word translation from English to any other language. Furthermore, ASL is more syntactically dense than the major gestures. Cues such as facial expression and lip movement form part of the language. Avatars neglect this aspect of the communication. Therefore, we present ASL through video, as shown in Exhibit D1.1-15.

Exhibit D1.1-15: American Sign Language

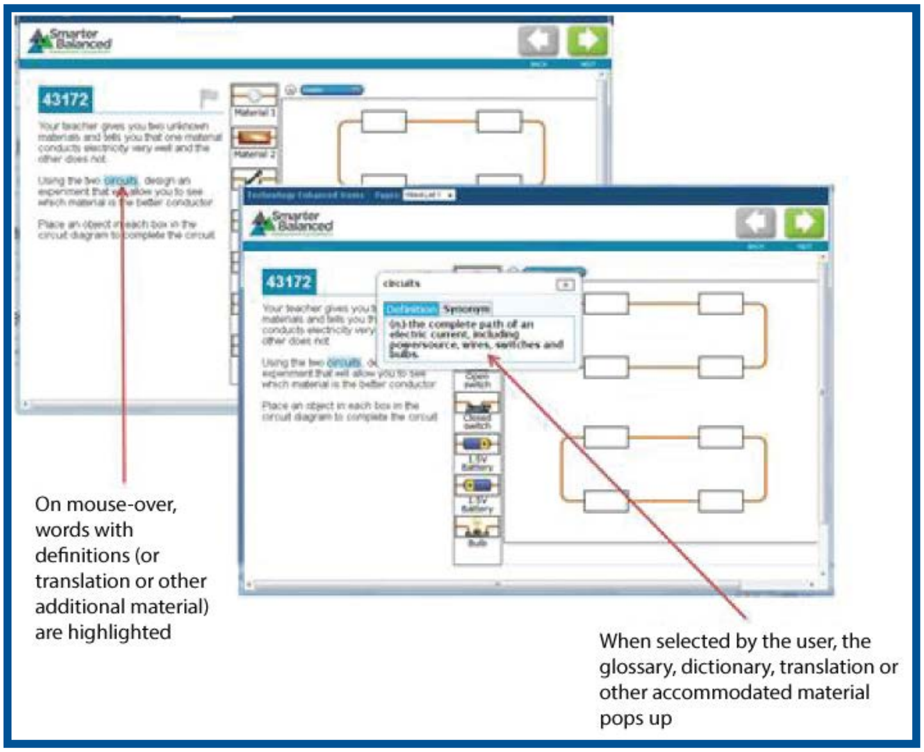


We have also developed a feature that we call “word lists.” Word lists support a wide variety of accommodations and embedded supports through a common, simple interface.

Some students require a contextualized dictionary, glossary, or second-language dictionary for translation. Word lists support all of these features. Any word in any item text can be associated with word lists of each type. Individual student profiles determine which types of word lists are active on each student’s tests.

The interface is simple, intuitive, and not distracting. Students mouse over words or move to them using the keyboard, and words associated with entries in the word list are highlighted. This “mouse over” approach avoids the distraction of many highlighted words (which research has shown decreases comprehension). Once a word is highlighted, clicking or selecting with the keyboard brings up the word list entries indicated by the student profile. This feature is illustrated in Exhibit D1.1-16.

Exhibit D1.1-16: Word Lists



Assignment of Tools and Supports

Each accommodation can be one of three different configurations or combinations of the following choices:

1. Available to all students
2. Assigned to students in advance by the designated state, district, or school test coordinators or through data upload (either manually in TIDE or from New Hampshire’s student information system)
3. Assigned to students at testing time by the proctor

Delivery Choices

We propose to offer a broad array of features. Furthermore, as we continue to innovate to extend the accessibility of the test, improve measurement, and make testing more engaging, features will proliferate. This presents a bit of a conundrum: How will students know what features are available? Students only test once or a few times a year. Therefore, they will generally not become “expert” users of the software. Making the appropriate tools available is best accomplished by offering each student the set of tools he or she needs, without the distraction created by tools that are not needed.

Each tool or embedded support provided by AIR’s system, by design, may be configured to be available to all students; available to students by one or more adult roles in the system (e.g., state, district, or school test coordinator or test administrator); and offered to all students but configured for each student by adult users.

Exhibit D1.1-17 illustrates some of the configurations other clients have chosen for some of the offered accommodations. Exhibit D1.1-18 illustrates the current accommodations New Hampshire has chosen to offer students.

Exhibit D1.1-17: Accessibility Configurations for a Sample of Embedded Supports and Accommodations

Accessibility Feature	Example of Configuration Choice	State’s Rationale Behind Configuration Choice
Color Choice	Assigned to students at testing time by the proctor	Most of our clients recognize that the proctor is the best person to determine the appropriate color since the student is with him or her in the room. If this tool is made available to all students, it would provide a distraction.
Increased Font Size	Default font size (assigned to students in advance)	A student’s need for this feature is part of his or her school record and therefore can be determined in advance.
Text-to-Speech—Directions, Passages, Items	Assigned to students in advance OR Assigned to students at testing time by the proctor	In some states, a student’s need for this feature is part of his or her school record and therefore can be determined in advance. In other states, it is available to all students but requires a proctor to turn it on to ensure that the student has a headset and so will not disturb anyone else.
Magnification	Available to all students	This feature does not provide a distraction, and the student may need it on different screens.
Answer Eliminator	Available to all students	This feature does not provide a distraction, and the student may need it on different screens.
Refreshable Braille	Assigned to students in advance	A student’s need for this feature is part of his or her school record and therefore can be determined in advance.

Exhibit D1.1-18: Current New Hampshire Accommodation Options

Field	Description
Text-to-Speech (TTS)	Reads test questions and answers.
Masking	Allows students to temporarily cover a distracting area of the test page.
Permissive Mode	Allows students to use accessibility software in addition to the secure browser.
Print Size	Displays tests in the selected magnification.
Translation (Glossary)	Allows students to display a glossary containing translations of specific words into the selected language.
Mouse Pointer	Modifies the size and color of the mouse.
Streamline Mode (Accommodation)	Displays tests in a simplified layout.
Presentation	Displays tests in the selected language.
American Sign Language	Allows students to watch an American Sign Language video for the corresponding test item.
Closed Captioning	Automatically displays closed captions for test items with audio elements.
Print on Demand	Allows a student to print selected components of test items.

Braille Type	Specifies the type of Braille in which tests are printed.
Audio Transcriptions	Transcribes output from a screen reader to a Braille display.

Some of the accessibility features require special attention and preparation to make them effective and meaningful for the item and the student taking the test. Each feature is designed to serve its primary purpose in as meaningful a way as possible by pulling in the appropriate resources. However, each feature is effective only if it interacts seamlessly with the items with which it is presented, as well as with other available features. Exhibit D1.1-19 lists these features and their respective requirements for preparation.

Exhibit D1.1-19: Accessibility Features and Preparation*

Accessibility Feature	Preparation
Color Choice	Colors that will be available on the system require that they do not conflict with the item graphics and text. For example, if a blue background is one of the background colors, the same blue cannot be used on item graphics. Thus, items need to be reviewed and may need tweaks if color conflicts arise.
Text-to-Braille and Text-to-Speech	Though the computer can automatically speak text rendered on the screen, special tags need to be created on graphics, directions, passages, and items. These tags would need to comply with the read-aloud principles determined by the Agency.
Large Print and Magnification	Graphics need to be legible in large print and all available magnification levels. Ideally, graphics would be vector-based so they magnify in a clear and consistent way. However, if vector-based graphics are not possible, some graphics may need to be enhanced depending on their legibility at specific zoom levels.
Sign Language	Recorded sign language translations need to be created in advance and undergo review levels (ASL or Exact English).
Audio Files	Audio files need to be created or recorded in advance.
Language	Content needs to be properly translated in supported languages other than English.

*We note that the availability of system features does not imply content annotated or prepared for delivery using the feature.

Setting Student Accommodations in TIDE

At the discretion of the Department, AIR can carry over student accommodations from year to year. If the Department elects to allow student accommodations to persist across years, when students are registered in TIDE each year, their previously assigned accommodations will pre-populate in the system. Otherwise, accommodations may be uploaded using the upload process described below.

TIDE will allow designated users to review student accommodations for each test administration and edit them as needed. Designated users will have the ability to download the accommodations template for a school, district, or selected set of students; edit the settings in the spreadsheet template; and upload the accommodations into TIDE. Users can edit the accommodations in the same file that they downloaded (the upload and download file formats will be same). As always, TIDE will restrict edits to accommodations by roles. Based on security settings discussed with the Department, only users with certain roles will have access to setting and uploading certain accommodations. New Hampshire users are familiar with the processes associated with assigning student accommodations in our TIDE interface, eliminating the learning curve often associated with adapting to a new system.

Please see Topic 1.3 Student Registration for more information about TIDE.

Assistive Technology

AIR's TDS currently supports a wide array of assistive technologies. AIR is working to expand the classes of assistive technologies supported. Our streamlined interface adheres to Web Content Accessibility Guidelines (WCAG) 2.0. We have achieved WCAG 2.0 AA certification, which means that the system is accessible and compatible with standards-compliant assistive technology. In addition, we know that our security mechanisms can interfere with some assistive technologies, and we have built in a permissive mode that relaxes these security restrictions for individual students who need to use such technologies.

AIR provides text-to-speech support with configurable text-to-speech tracking. The computer reads text and graphics aloud on directions, passages, and items. What is read and how it is read are configurable, and the Department can offer a variety of options. AIR's approach to text-to-speech uses well-specified guidelines to annotate the items for speech and generate the speech on the student machines. We use the voice packs on the machine (and can support any compliant voice pack supported). This does result in some variability in intonation and pronunciation from machine to machine; however, we have taken steps to mitigate the impact of these differences. In particular, the one popular operating system that offers a stock voice pack that is below our expectations is Windows XP.

To mitigate this problem, we provide a high-quality Julie voice pack from NeoSpeech that can be downloaded and installed on Windows machines at no cost to the Department. Windows users in New Hampshire who utilize this tool should have the voice pack installed on their Windows devices from the 2016–2017 test administration for use in the secure browser. Therefore, this should impose no additional work on the behalf of users.

The system currently works with a wide range of refreshable Braille devices, as well as screen readers, on-screen keyboards, and a wide array of input devices. While it is impossible for any organization to guarantee support for unknown hardware and software, AIR is committed to accessibility for all students and has demonstrated that commitment through our leadership in this area.

Print-on-Demand

AIR's system delivers real-time adaptive Braille, large print (online and paper), and paper tests. All of these features use our secure print-on-demand feature. This feature prints an item or item group to a designated printer (large print and other paper) or embosser (Braille). The student or scribe presses a request button, and a print request appears on the test administrator's screen. The test administrator (TA) approves the print request and retrieves the document from the printer or embosser.

For students who cannot interact with the computer, we offer our data entry interface (DEI). The student can respond on paper, and a designated scribe can enter the data at a later date through the DEI, which essentially allows the TA to go through the test for the student. To enter student responses into the scoring system, whether during test administration or at a later time, TAs will enter through the DEI, which does not require the TA Interface (explained in further detail in Topic 4 Assessment Delivery Platform) or the creation of a test session. TAs can pause data entry and return to complete the entry at a later time. TAs may review their entries, revising any data entry errors, prior to submitting the test record for scoring. Once the TA has submitted the test record for scoring, the system will perform a quality assurance (QA) review of the test record before scoring and sending the scored test information to the Online Reporting System (ORS). This process occurs almost instantaneously.

In most cases, students using Braille and large print are able to navigate the test themselves (the system is fully keyboard navigable) and enter their own responses.

The print-on-demand function is protected with security controls at three levels:

- Embedded security in the print-on-demand function
- Authentication, ensuring that only authorized users access information
- Policy and test administration procedures, ensuring that secure materials are properly handled, retrieved, and tracked

Embedded Security

The print-on-demand function has the following security features:

- It requires that the student's request to print is approved (or denied) by the TA overseeing that testing session.
- Only a TA can print the request.
- Since TAs typically are not allowed to view test items, this feature restricts what they see through the print preview and the print output. All print requests have a cover page with a security warning, and each corresponding page has a header containing information about the student, the test, and the item. None of the test content is displayed on print preview.

In addition, the system collects and stores all information about the print request so data can be used for further tracking, analysis, and auditing purposes. Data collected on each page include a unique identifier, identifiers on the item or passage, student information, test information, and time stamps. In other words, the system has the ability to track every document printed and conduct reconciliation processes between what was requested by the student and what was collected. The printed requests will be linked to each student and test administrator.

Last, this function is configurable; therefore, at the Department's specification, it can be restricted in different ways or a combination thereof.

Authentication

The print-on-demand function can only be activated by authorized users for the student test. At the Department's request, authorized users can be set up in the system based on roles, so that they only have access to information they can view. Authorized users can predetermine and set the print-on-demand function through TIDE. In addition, if the Department chooses to empower TAs to activate this function, AIR can arrange for certified TAs to have authorization. Certified TAs complete an online TA certification/training program, which includes signing a TA confidentiality form.

Topic 2 Item Development

New Hampshire seeks an off-the-shelf assessment that may include customized item development and the creation of specific test items. AIR recognizes the importance of both state and stakeholder involvement in the design and use of that assessment system. AIR's solution provides New Hampshire with summative and interim assessments that are aligned to New Hampshire College and Career Ready Standards. For the summative assessment, we are pleased to offer New Hampshire a standards-based, adaptively administered test based on our ICCR item pools. The ICCR item banks have been developed to measure College and Career Ready Standards in statewide assessments. As mentioned in Topic 1, these items have been administered as part of the Arizona, Florida, Ohio, Tennessee and Utah statewide assessments, and they have been embedded in Oregon's administration of the Smarter Balanced assessments. ICCR items were developed in close collaboration with these state partners. All items were reviewed by state staff, teachers, and experts in bias and sensitivity. All ICCR items have been field tested in embedded slots within operational test administrations. Data was collected and reviewed by AIR content and psychometric staff, as well as state department staff and their data review committee; poorly performing items were rejected from the pool.

As mentioned in Topic 1 Test Design, we propose to deliver New Hampshire's interim assessments using one of several formative item pools. AIR's own Navigator item pool includes items developed by the same AIR item development staff who produce test items for statewide assessments, including the ICCR item pools. Navigator items were also developed to be administered by the same test delivery system used to deliver statewide assessments, so Navigator items have the same look and feel as the ICCR items we propose for the Department's summative assessments. We also have available item pools developed by Key Data Systems (KDS), which offer quality math items to measure College and Career Ready Standards, and provide a variety of item types that will allow us to deliver interim math assessments consistent with the summative assessments in both content and functionality. New Hampshire may also be able to access Utah's formative item pool populated with items already calibrated and equated to the

ICCR scale. These items were developed using item specifications that are nearly identical to those developed for ICCR, and include the same types of interactions, and have the same look and feel as ICCR items.

The ICCR item pools were developed to ensure that statewide assessments cover the full range and depth of the content standards at the aggregate level for each test administration. The bank grows larger each year as we continue to field test new items in multiple state assessment systems. This section provides a description of the procedures used to develop the ICCR items for use in multiple state assessments. We will address the ways in which the Department can participate in that review process and adopt only the ICCR items that best fit the needs of New Hampshire. In addition, we describe our item banking tool, which is used to manage continued development of the ICCR item pool and which includes all attributes of test items, including item content, rubrics, item statistics, the history of item reviews, and more.

Throughout this section, we will refer to the ways in which our response

- reflects knowledge of the New Hampshire College and Career Ready Standards;
- offers the industry's most cutting-edge technology to measure higher-order thinking skills;
- emphasizes accessibility and the ways in which technology can enhance accessibility;
- ensures strict security protocols, especially as security relates to the review of items by multiple states and committees; and
- offers a robust pool of rigorous, text-based writing prompts.

Topic 2.1 Item Development Team

AIR's test development structure supports highly effective units organized around each content area (math, English language arts, and science). Unit directors oversee team leaders who work with team members to ensure item quality and adherence to best practices. In addition to team leaders, each unit is staffed with senior content consultants who provide further support to the teams by leading item reviews and designing ongoing training when needed. All of the items we develop, including ICCR items, are created within this unit structure with a wide variety of content experts reviewing and approving items/passages. Because we work as a unit rather than small project-specific teams, we are able to offer the experience and expertise of our entire staff to the development of items across projects. This obviates the need for individual grade-level developers and allows us to provide our clients with a single point of contact in each content area.

The team that we assemble at AIR to work with the NH Instructional Support Team will consist of a lead developer in each content area and a test development manager, who will work across content areas to plan and organize the workflow. The test development manager will develop a plan for New Hampshire's review of ICCR items and passages and will ensure that a logical and reasonable schedule is in place.

Our Item Tracking System (ITS), which is described further in Section 2.2 Item Development and Review, will facilitate the communication between AIR's lead developer and the NH Instructional Support Team. All items and passages will be delivered to the Department through ITS so that the NH Instructional Support Team can review them directly in the database where they are housed, using the same rendering software that our test delivery system uses. The AIR lead developer will work with his or her counterpart to assist the NH Instructional Support Team in the review and approval of ICCR items. Although most items have been field tested and can no longer be edited, the ICCR banks are robust enough to allow the Department to choose only those that best align with New Hampshire's vision for its new assessment. The online delivery of these items will allow for easy, efficient communication between AIR and the NH Instructional Support Team, reducing the number of in-person meetings needed. AIR staff will attend an initial in-person meeting as called for in Topic 2.2 of the RFP. After this initial meeting, we believe that the review of ICCR items can occur through virtual meetings between lead developers and the NH Instructional Support Team.

Topic 2.2 Item Development and Review

AIR looks forward to an initial in-person meeting with the NH Instructional Support Team so that we may lay the foundation for a successful partnership. Here, we hope to have the opportunity to go over the blueprints and item specifications that were used to develop the ICCR bank. We look forward to conversations about how these align with the Department's vision for its new assessment and where modifications to the blueprint, if any, would be made. As mentioned in Topic 1, we have worked with a variety of states who have used the model ICCR blueprint as a springboard to create their own customized assessments. We will be prepared with bank analyses that we can share with the Department so that we can have detailed discussions about the kinds of blueprint modifications the ICCR bank will support. At this initial meeting, we will also discuss communication protocols and plans for future meetings.

The Department may wish to be solely responsible for reviewing items in the ICCR item pools, or to additionally conduct a stakeholder committee review of the item pools as some ICCR states have done. Should the Department choose this option, we would make the items available immediately and allow the NH Instructional Support Team to conduct a review at their convenience by viewing the items directly in ITS. AIR would be happy to facilitate a brief training to show the NH Instructional Support Team how to use this system. The NH Instructional Support Team could flow their feedback to AIR on a regular schedule that would be designed to conclude prior to committee review. AIR would then bring to committee only the ICCR items that the NH Instructional Support Team had approved.

As mentioned in Section 2.1, AIR will deliver items and passages to the NH Instructional Support Team via ITS. Among the many advantages of reviewing items directly in our system are convenience and efficiency. The NH Instructional Support Team can easily and securely access the system from any location with Internet access. Reviewers can search for items by any attribute that is tracked in ITS (e.g., by grade, standard, item type, passage association, etc.). AIR will provide clear instructions on how to access the items. We are also happy to plan a virtual meeting with the NH Instructional Support Team to show them how to navigate the system. Once the NH Instructional Support Team has completed their review, AIR staff will carefully review their feedback. Many ICCR items will be accept/reject only, but for any items that are still in development, we are happy to receive the NH Instructional Support Team's feedback and consider those suggestions as we complete the items. The AIR test development manager will create a schedule that provides a sufficient amount of time for the NH Instructional Support Team to conduct this review.

Topic 2.3 Item Review Committees

Soliciting input on items and passages from educators within the state is a key part of ensuring the validity of each assessment as well as gaining the confidence and support of stakeholders in the community. AIR will work with the Department to plan and facilitate item review committees in New Hampshire. Our Item Tracking System (ITS) will support this process by allowing us to display items online for educator committees using the same rendering engine that displays the items in the test delivery system to students. Because our test developers create technology-enhanced items themselves, the items are fully functional, complete with scoring rubrics, from the earliest stages of development. This allows us to show educator committees exactly how the item will function and score in the real test environment. We note, however, that most ICCR items are already operational and will be in a finished state. They will be presented as accept/reject only, but because the ICCR items have already been through, in many cases, multiple committee reviews, we anticipate a successful review.

AIR will support grade-level item review committees in grades 3–8 for both math and ELA and in grades 5, 8, and 11 for science. We will defer to the Department on the recruitment of educators and the composition of those committees. AIR understands that summer is the best time to schedule these meetings due to the availability of educators and the timing of the test development cycle. We will be prepared to support these meetings in the summer months. We will work with the Department to determine the most efficient way to organize educator groups, most likely by grade level or grade band. AIR facilitators will work with the Department to create an agenda for the meeting and prepare a brief introduction and training. AIR staff will take responsibility for test security, making sure that any secure materials distributed at the meeting are collected and accounted for afterward. We will provide the committee members with non-disclosure forms and collect them before item review begins. We will work closely with the Department and the NH Instructional Support Team prior to the meeting to provide sufficient time for reviewing all materials.

Per the RFP, AIR will support item review meetings with the following specifications:

- Each grade-level committee will consist of 3–6 members.
- Committee members will be paid a stipend of $150 per day for participation in the summer meetings.
- Representatives from New Hampshire will also attend the meeting, including the NH Instructional Support Team (up to a maximum of three [3] people total).
- AIR will support and arrange lodging for committee members and New Hampshire representatives. We will assume that all participants will require lodging for three nights for the meeting (beginning one night prior to the meeting).
- AIR will also be responsible for travel expenses (e.g., mileage, airfare) for all participants to attend any meetings out of state.
- AIR will provide breakfast and lunch each day of the meeting and be responsible for dinner expenses ($25 per day) on days that require an overnight stay.

AIR will plan meetings at a hotel, conference center, or other approved suitable location in New Hampshire. We will work with the Department well in advance of these meetings to ensure that all meeting logistics, including any needed equipment (e.g., laptops, projectors, etc.), are accounted for.

Once the meeting has concluded, AIR staff will schedule one additional day to meet with the NH Instructional Support Team to reconcile feedback from the item review committee. Within two weeks of the meeting, AIR will deliver to the Department a report documenting the meeting. Documented final outcomes of items and passages will be available in ITS or can be delivered as a summary report to the NH Instructional Support Team for their records.

Topic 2.4 Bias/Sensitivity Review Committee

AIR will work with the Department to plan bias/sensitivity review committees. These committees, consisting of external educators and experts recruited and selected by the Department, will review the content of passages, other stimuli, and test items for potential violations of bias and sensitivity guidelines. The first bias/sensitivity meeting will consist of ICCR items/passages that are slated for use on the new operational assessments, as well as ICCR items that will be placed in embedded field-test slots on those assessments. In 2018 and beyond, bias/sensitivity meetings will also include ICCR field-tested items that were flagged for differential item functioning. Should the Department engage AIR in the development of customized items for New Hampshire, we will ensure that those items are also reviewed by a bias/sensitivity committee both prior to field testing and after field testing when differential item functioning statistics become available.

Once the Department has determined the composition of the bias/sensitivity committee, AIR will support communication with committee members and meeting logistics. Per the RFP, we will plan for one committee in each content area (ELA, math, and science), consisting of five members each. Trained in coordination with the Department, AIR staff will facilitate these meetings by providing the committee with an overview of New Hampshire-approved bias/sensitivity guidelines. The meetings will occur annually in the summer.

During the development of ICCR items each year, AIR test development staff receive extensive training on bias/sensitivity and language accessibility. As a result, we are typically able to address any potential issues early on in the item development process and avoid the loss of items later on. A review of bias/sensitivity is built into our item review process and items go through a systematic check for these issues at each step of the review. As a result, we find that our item loss at bias/sensitivity review is typically minimal. The greatest potential loss occurs when a passage or stimulus is rejected for bias/sensitivity, which results in the loss of all its associated items. Should the Department engage AIR in customized item development, AIR would propose a separate committee review of passages/stimuli in ELA and science prior to item development. These meetings would occur early in the year and function as a combined content/bias/sensitivity review to ensure that passages and stimuli are approved for item development. The Department staff are welcome to participate. Our proposal includes travel costs for one Department staff in each subject area.

AIR will support the following specifications for bias/sensitivity review meetings:

- Committee members will be paid a stipend of $150 per day plus travel expenses for participation in the summer meeting.
- The meeting will also be attended by a representative of the Department (one person).
- AIR will support and arrange for lodging for committee members and the state's representative attending the meeting.
- AIR will provide breakfast and lunch each day of the meeting and be responsible for dinner expenses ($25 per day) on days that require an overnight stay.

Within two weeks of these meetings, AIR will provide the Department with a written report documenting the meeting.

Topic 2.5 Content Review

AIR's internal review process is among the most rigorous in the industry. It includes multiple stages of review by content experts who are extremely knowledgeable of the latest thinking in their field. In this section, we describe that process in further detail and explain how it is supported by our Item Tracking System (ITS).

AIR's test development staff and structure ensure the quality of items by enforcing a structured review process and formalizing feedback to our staff and item writers to continuously improve content and share ideas. We equip our people with cutting-edge tools, empowering them to create innovative items that challenge students to apply what they have learned.

Our sound processes, backed by a team of highly qualified test developers, result in valid, reliable, innovative items that elicit meaningful performances from students. AIR's structured item development process, managed by our ITS, ensures that all items flow through each step of the review process. ITS maintains records of each review and draft of an item from inception to operational usage.

Our test development group, organized by content area, is closely managed by unit directors who ensure that each test developer receives ongoing training in assessment best practices and AIR's systems and processes. This organizational structure, which promotes effective communication and continuous learning, ensures institutionalization of knowledge around standards and specifications, even when those specifications may change as a result of the adoption of new content standards.

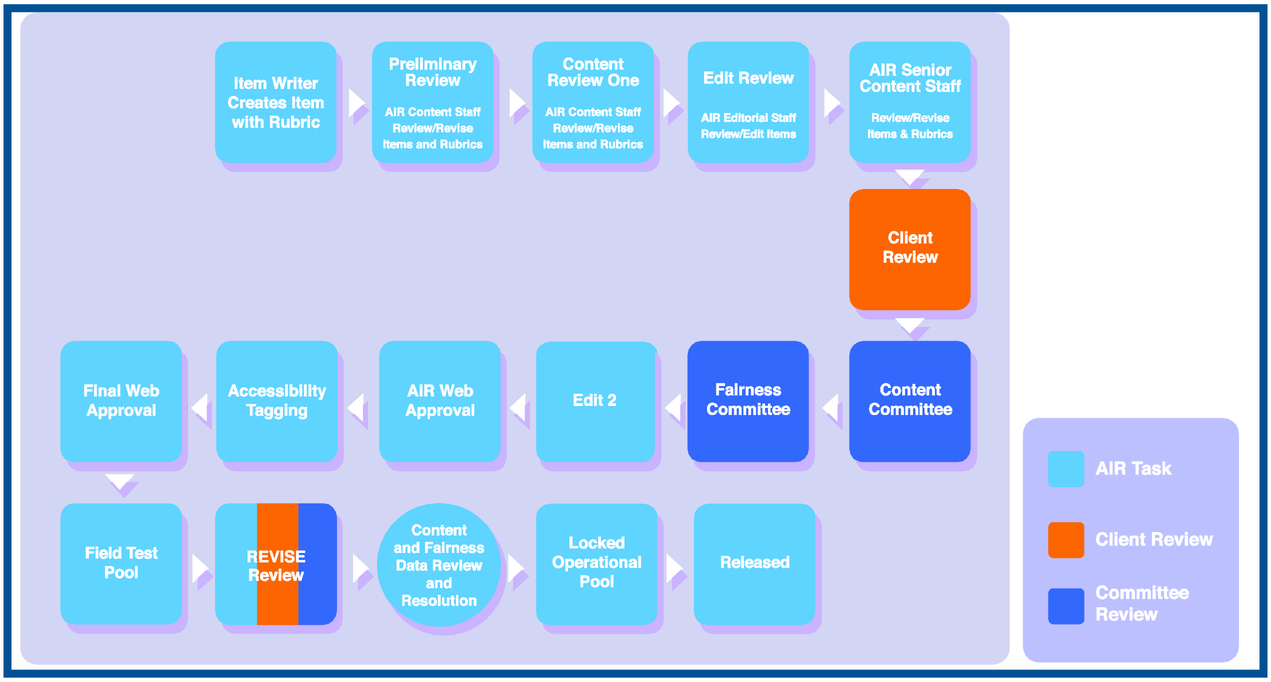
ICCR items went through AIR's standard review process, which is among the most rigorous in the industry. As part of our standard item review process, items go through multiple levels of internal review before they are sent to a client. Exhibit D1.1-20 summarizes the item development sequence in ITS for the ICCR item bank. At each stage of the review process, AIR test developers check items to ensure that they are aligned to the standards, are clearly and succinctly worded, are grade appropriate, contain appropriate graphics, have accurate and fact-checked content, are free of bias and sensitivity issues, follow the style guide, and follow the principles of universal design. Each step in the review process corresponds to a level in ITS. The levels are as follows:

- **Preliminary Review—**typically conducted as a group made up of both senior and junior level staff.
- **Content Review One—**generally conducted by an individual test developer who examines each item based on all the criteria above.
- **Edit Review—**conducted by an editor who looks at every item for clarity, correctness, and appropriateness of language for the grade level assessed.
- **Senior Review—**generally conducted by team leaders and/or senior content consultants. The senior reviewer looks back at the item's entire review history, making sure that all the issues identified in that item have been fully addressed before items are delivered to the Department.

In addition to supporting this extensive internal review process, ITS also ensures item quality by tracking the review of clients and committees. Items must move through the following levels and have documented acceptance outcomes before becoming eligible for field testing:

- **Client Review and Resolution—**Here, AIR maintained a record of client feedback on all ICCR items that went through a state client's review. We recorded any revisions that may have occurred as a result of that review process. We note that ICCR items are edited by consensus between AIR and the client(s) reviewing them.
- **Content Committee Reviews—**Each ICCR item and passage was reviewed by educator committees. As mentioned in Section 2.3 Item Review Committees, the items were fully functional and included scoring rubrics during that review. Educators had the ability to see and interact with the items, review the scoring rubrics, and provide feedback to AIR test developers who recorded those comments in ITS.
- **Bias/Sensitivity Committee Review—**All ICCR items also passed through a bias/sensitivity committee review. Again, the AIR facilitator recorded the outcome of that review and noted any revisions that were made as a result of it.
- **Accessibility Tagging—**AIR's team of accessibility specialists put the items through a series of review steps to ensure the annotations and translations were clear and precise. A full description of the accessibility features we offer can be found in Topic 1.4 Accessibility and Fairness.
- **Final Web Approval—**At this stage, AIR test developers reviewed the items a final time. Here we saw the final product displayed as it will render to the students. We checked to ensure that the online appearance was appropriate, the content was correct, the scoring was accurate, and any final edits were appropriately applied.

Exhibit D1.1-20: Item Development Sequence in ITS



This item development workflow offers the ability to manage items from inception, through a series of content, fairness, data, and other reviews to final publication. The system captures the outcomes and rationales at each review step and maintains previous drafts of each item. The workflow management ensures that each item receives each review in the designated sequence, and that the review is conducted (or recorded in the case of committee review) by an authorized person. Every version of every item is archived, along with each comment received in any review. Reviewers have immediate access to all older versions providing version control throughout development. All attributes of the item, including accommodation and accessibility tags, item statistics, rubrics, etc., are associated with the item and travel with it through each review step. This helps us ensure that the version of each item and passage moving into the field test is accurate and has been vetted through the full review process. Exhibit D1.1-21 provides a sample of one item's review history at the end of development. In Exhibit D1.1-21, we show the documented outcome of committee reviews, a final edit check, and a final review by an AIR test developer to confirm the content.

Exhibit D1.1-21: Review History in ITS

Action: View
Date: 7/13/2011 6:52:33 PM
Review Type: Content/Fairness Committee Review and Resolution
Review Comments:
kmc 07.13.11:
- per Content Fairness Item Review Mtg, in the stem, changed 'revise his' to 'test a new'
- in task statement A, changed 'revised' to 'new'
- in the background requested APG change 'A. Revised' to 'A. New'
- reviewed graphics; AAR

Review Outcome: Accept as Revised
Review by: Kevin_Chandler

Action: View
Date: 7/14/2011 4:46:09 PM
Review Type: Edit Check II
Review Comments:
CR: OK
Review Outcome: Accept as Appears
Review by: Christine_Ryerson

Action: View
Date: 7/29/2011 3:30:08 PM
Review Type: AIR Edit Resolution/Final Review
Review Comments:
TG 7.29.11:
-updated SA rationales to match the revisions made to the item at CFCR
Review Outcome: Pend
Review by: Thomas_Goblet

AIR understands that New Hampshire has begun a process to review and revise the standards in ELA, math, and science. We stand ready to work with the state to assist in this endeavor. We have supported many other states in the transition to new standards and will be fully prepared to help with the revision of test/item specifications, blueprints, style guides, and any other guiding documents that may need to change as a result of this shift. Additionally, new items may need to be developed. We will work with New Hampshire to develop a scope of work and timeline for that process.

Topic 2.6 Item Types and Number of Items

AIR offers a wide array of interaction types that go above and beyond those listed in the RFP. We provide sample items in this section that not only demonstrate our capability to provide a variety of machine/AI-scorable items but that also show the functionality of our cutting-edge item authoring tool, AIRCraft. AIRCraft removes constraints that have historically impeded item developers from developing items that engage students in meaningful activities. AIRCraft also allows us to challenge students to apply the skills and knowledge that they have acquired in class.

To understand AIRCraft, set aside the notion of an item type. Think instead about interactions, that is, different ways that students can interact with the item to demonstrate what they know and can do. A multiple-choice item has a single interaction and is a selection interaction. Another item might ask a student to draw a graph (graphic-response interaction) or write an equation (equation interaction). Our revolution in test development breaks down the barriers between these types of interactions and enables us to incorporate all of these interactions in a single item. Students can be scored not only on whether they get the right answer, but also on whether they are able to appropriately incorporate choices in earlier interactions with choices made in other interactions. This capability allows for multiple solution paths and can challenge students to apply knowledge in meaningful settings to craft multi-step solutions.

The interaction types currently available are shown in Exhibit D1.2-22. Please see Appendix Q for screen captures of each of the interaction types we offer.

At the same time that AIRCraft facilitates more meaningful, challenging items, it also integrates accessibility into each item, automatically inserting Accessible Rich Internet Applications (ARIA) tags and other accessibility markup needed to have the items comply with Web Content Accessibility Guidelines (WCAG) 2.0 accessibility standards and promote compatibility with assistive technologies. Doing so makes our online tests accessible to virtually all students.

AIRCraft enables test developers to create machine-scored items that challenge students to actually construct something and demonstrate their understanding. Perhaps AIRCraft's most important innovation is the ability to base scoring on specific features of a student's response, rather than simply matching correct answers to a student's selection. As illustrated in the item samples that follow, AIRCraft items embody evidence-centered design; for each item, the scoring is based on an explicit set of assertions that link features of student responses to the specific skill or knowledge that the response demonstrates. AIRCraft is designed for both efficiency and power. Common item interactions are established using templates, and scoring keys or scoring rules can be specified with a few selections from a graphical interface. More complex item interactions require test developers to introduce (sometimes complex) scoring rules. AIRCraft includes an easy-to-use graphical user interface to author complex rubrics for items with minimal effort. The system has built-in tools to unit test the rubric by subjecting it to a variety of inputs. Validation checks are integrated into each step to guide the item-authoring process.

Sample Mathematics Task

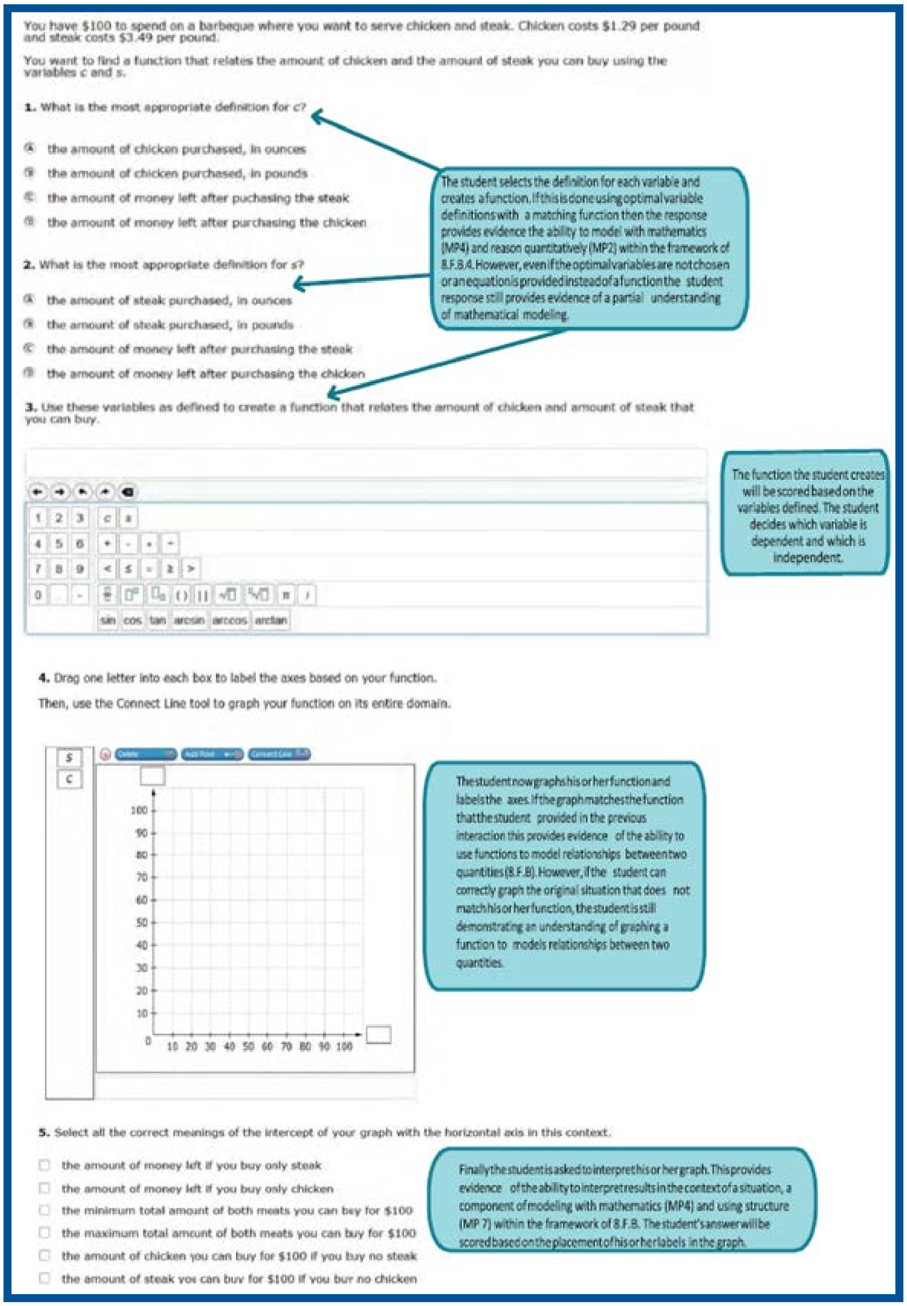
Exhibit D1.2-23 presents an example of the type of machine-scored, multiple-interaction item that test developers can develop using AIRCraft. This sample item is adapted from a classroom activity recommended by *Illustrative Mathematics*, an organization founded by Bill McCallum.

This item consists of five interactions, and the scoring of the interactions is dependent. The choices the student makes in one interaction are recognized and taken into account in other interactions. Using this technology, students have great latitude to make choices and demonstrate how they are thinking.

Exhibit D1.2-22 Interaction Types

Interaction Types	Description
associateInteraction	A student creates a one-to-one or many-to-one relationship between two pieces of data. Machine-scored.
ChoiceInteraction:MultipleChoice	A student selects from traditional multiple-choice options using radio buttons. Machine-scored.
ChoiceInteraction:MultipleSelect	This is a traditional multi-select (checkboxes) with flexibility on minimum and maximum number of selections. Machine-scored.
ChoiceInteraction:Scaffolding	This is a graphic-rich choiceInteraction variant often used with alternate assessment student populations which removes an incorrect response and gives the student a second try. Machine-scored.
ChoiceInteraction:ScoreInput	This is a variation of choiceInteraction that allows a data entry clerk to directly enter the score a student achieved instead of the response.
customInteraction:BarChart	The student uses simple single-click actions to draw bar charts. Machine-scored.
customInteraction:EditTask	A student clicks a word and replaces it with another word to revise a sentence. Machine-scored.
customInteraction:EquationEditor	A student uses flexible entry of one or more mathematical expressions. Machine-scored.
customInteraction:Graphing	A student graphs simple concepts like rays, line segments, and closed shapes. Machine-scored.
customInteraction:GraphicResponse	This is a flexible platform to create a wide variety of items from drag-and-drop, to drawing on a coordinate plane, to hot spots. Machine-scored.
customInteraction:Scratchpad	This allows freeform drawing, text, or mathematics entry onto a canvas. Hand-scored.
customInteraction:Simulation	A student conducts an experiment by manipulating inputs until the desired output is reached. Machine-scored.
customInteraction:TableInput	This solicits a student to complete tablular data. Machine-scored.
customInteraction:VerbalResponse	This records the student's voice using a microphone. Hand-scored.
customInteraction:WordBuilder (Numeric and Qwerty)	This is a fill-in-the-blank style response with an onscreen keyboard that shows the allowable keys. Machine-scored.
textEntryInteraction	This is a short answer text response. Hand-scored or machine-scored using natural language scoring engines.
extendedTextInteraction	This is an extended response for essays. Hand-scored or machine-scored using essay scoring engines.
gapMatchInteraction	A student drags pre-written text responses into target boxes to respond. Machine-scored.
graphicGapMatchInteraction	A student drags prepared graphic responses into target boxes to respond. Machine-scored.
hotSpotInteraction	A student clicks on clickable regions of a graphic to respond. Machine-scored.
hottextInteraction	A student clicks on a word or sentence to respond. Machine-scored.
html:GenericInstruction	This is a simple text display usually used for instructions. Not scored.
html:Slideshow	This is an audio presentation with static images used as slides. Not scored.
html:Tutorial	This is animation that describes how to use that type of interaction. Not scored.
inlineChoiceInteraction	A student selects responses from drop-down menus throughout the text. Machine-scored.
matchInteraction	A student creates a one-to-one relationship between two elements. Can be represented graphically or as a table. Machine-scored.
orderInteraction	A student reorders a list of elements to create the correct sequence. Machine-scored.

Exhibit D1.2-23: Sample Mathematics Item



The way the items are scored creates a direct linkage between *what the student does* and how features of the student's response provide evidence about the *knowledge and skills* the student has achieved. This approach provides a physical embodiment of *evidence-centered design*, Mislevy's well-regarded approach to cognitive measurement. This approach provides a structure for ensuring and reviewing alignment during test development, and a clear explanation not only of what was measured but also how and why to support tests scoring and reporting.

Exhibit D1.2-24 provides a real example of the scoring assertions (evidence statements) that are used to score the sample item above. This sample item assesses student understanding of the concept of functions (8.F.A.1) and graphs of functions via the use of ordered pairs (also 8.F.A.1). It requires students to construct a model of a function given key information (8.F.B.4) and interpret key parts of the function (8.F.B.4). The item draws upon students' knowledge of different representations of functions, in particular, linear functions (all part of cluster A: define, evaluate, and compare functions.) Further, students attend to the Standards for Mathematical Practice by interpreting parts of the model with respect to the context (MP4), reasoning abstractly about a function and generating an algebraic model (MP2), and attending to the structure of a function by focusing on its key parts (MP7). This item assesses knowledge of multiple complex mathematical concepts in one authentic, coherent flow. Students implement several steps of the modeling cycle and make choices in how they construct their models (with both an algebraic function and a graph modeling the function). Our scoring system recognizes the choices the student has made and machine scores subsequent responses based on those choices.

Exhibit D1.2-24: Sample Scoring Assertions Relating Specific Features from the Student Response to the Skills and Knowledge Being Tested of Which They Provide Evidence

Scoring Assertion	Standard/Practice
The student selected an appropriate definition for *s*, providing some evidence of the ability to identify relevant entities in a real-world phenomenon, an important part of modeling with mathematics.	MP4
The student selected an appropriate definition for *c*, providing some evidence of the ability to identify relevant entities in a real-world phenomenon, an important part of modeling with mathematics.	MP4
The student created an equation accurately expressing the relationship between *c* and *s*, given their definition, providing evidence of the ability to formalize relationships among real-world entities, an important part of modeling with mathematics and reasoning quantitatively.	8.F.B.4 MP2, MP4
The student isolated either *c* or *s* on one side of the equation, formulating an equation that most conveniently serves the assigned purpose, an important part of modeling with mathematics.	8.F.A.1
The student correctly graphed either the correct function or the function that he or she provided, providing evidence of the ability to model relationships between two quantities.	8.F First Cluster (Define, evaluate, and compare functions)
The student correctly interpreted the literal meaning of the intercept and did not select any incorrect options, providing evidence of the ability to interpret results in the context of a situation.	8.F.B.4 MP4, MP7
The student correctly made the inference that the intercept also marks the most (if the student graphed chicken on the horizontal axis) or least amount of meat (if the horizontal axis represented steak) that could be purchased and did not select any incorrect options, providing evidence of the ability to interpret results in the context of the situation.	8.F.B.4 MP4, MP7

Please note that the item asks the student for five interactions, but the scoring assertions extract seven pieces of information, covering four different standards and three mathematics practices. By inspecting the student response for every meaningful piece of student input, we are able to harvest more information about what the student knows and can do. This provides greater measurement efficiency, supporting shorter tests with more detailed reporting.

Sample English Language Arts Task

Many of AIR's most innovative English language arts items use scoring assertions, which are indicators of whether a student's response provides evidence of understanding tied to content standards or elements of those standards. In Exhibit D1.2-25, we present a sample ELA grade 10 item that demonstrates how scoring assertions explicitly tie features of a student response to the standards for which those features provide evidence. In this example, five interactions yield six scoring assertions (see Exhibit D1.2-24), providing evidence of four unique standards across two strands. The passage pair used for the item is a retelling of the Prometheus myth paired with an excerpt from Mary Shelley's *Frankenstein.* The item aligns to Reading Literature standards 2, 6, and 9 and Writing standard 9 in the Common Core State Standards.

The multiple interactions in this item provide opportunities for students to peel away the layers of a complex text in order to provide evidence of deeper understanding. Beginning with an analysis of ideas in the text, the interactions bring in more information that allows students to analyze the texts in the context of "ideas beyond the text." The scoring assertions in this item allow us to measure multiple standards by eliciting different types of evidence from the student. While some students may only provide evidence that demonstrates an understanding of the surface-level meaning in the texts, others will provide evidence that they have a full understanding of the texts' subtleties and relationship to ideas outside the of text.

Exhibit D1.2-25: Sample ELA Item

- As evidenced from the title of her novel, Shelley drew on the myth of Prometheus when she created the characters of Dr. Frankenstein and the creature. Just as Prometheus gave fire to humankind, Dr. Frankenstein gave life to the creature.

Which theme does Shelley develop in the excerpt? Select one sentence in Passage 2 that develops it.

Ⓐ the indifference of nature

Ⓑ the benefits of human progress

Ⓒ the negative impact of civilization

Ⓓ the importance of scientific research

- Select a sentence in Passage 2 that develops the theme.

"Every conversation of the cottagers now opened new wonders to me. While I listened to the instructions which Felix bestowed upon the Arabian, the strange system of human society was explained to me. I heard of the division of property, of immense wealth and squalid poverty, of rank, descent, and noble blood.

"The words induced me to turn towards myself. I learned that the possessions most esteemed by your fellow creatures were high and unsullied descent united with riches. A man might be respected with only one of these advantages, but without either he was considered, except in very rare instances, as a vagabond and a slave, doomed to waste his powers for the profits of the chosen few! And what was I? Of my creation and creator I was absolutely ignorant, but I knew that I possessed no

First, the student must determine the theme or central idea of the Shelley excerpt. A student who recognizes that Shelley focuses on the negative impact of civilization provides evidence of the ability to accurately determine the theme of a text. Next, the student must select a sentence that develops the theme, providing evidence of the ability to determine how a theme is shaped by specific details. Taken together, the two interactions provide multiple pieces of evidence of a student's ability to perform different parts of RL.2.

- Explain how the theme and sentence you selected supports the connection between the story of the creature and the myth of Prometheus.

The first two interactions scaffold to a short-answer item asking students to explain how the theme and sentence they selected support the connection between the myth of Prometheus and the excerpt from *Frankenstein*. The interaction collects information about a student's ability to analyze how an author draws on and transforms the source material in a specific work (RL.9).

- The Enlightenment was a period during which great thinkers believed that advances in science and technology promised human improvement and equality. Select **two** ways in which Shelley's characters in Passage 2 suggest her views on Enlightenment thought.

☐ The creature becomes physically superior to humans.

☐ The cottagers believe that divisions in society are just.

☐ The creature's search for knowledge only leads to misery.

☐ The creature is rejected by both his creator and the cottagers.

☐ The cottager's plight leads the creature to feel sympathy toward them.

- *Frankenstein* was published in 1818 during a literary movement known as Romanticism. Read the following quotation from Lord Byron, a famous writer of the Romantic movement.
    - "Knowledge is not happiness, and science but an exchange of ignorance for that which is another kind of ignorance."

Both *Frankenstein* and the myth of Prometheus suggest ideas about enlightenment and scientific advancement. Write an essay that analyzes how the quote supports the ideas in the myth and the excerpt. Be sure to use evidence from both texts in developing your response.

B I U Tx Ω

Next, the student must select two ways that Shelley's characters suggest her views on Enlightenment thought, providing evidence of a student's ability to analyze a point of view or cultural experience. Finally, the student must complete a literary analysis task that asks him or her to analyze both texts in relation to a quotation from an Enlightenment thinker, providing further evidence of his or her reading comprehension (RL.6) and providing information about W.9, as students must use evidence from literary texts to support analysis, reflection, and research (W.9).

Exhibit D1.2-26: Scoring Assertions

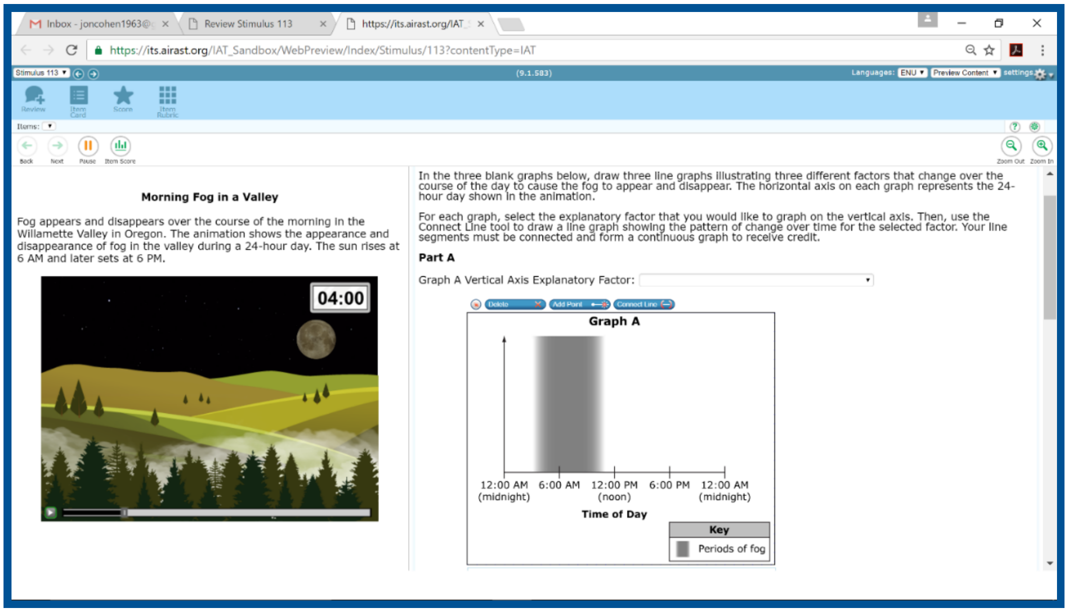
Scoring Assertion	Strand/Standard
The student selects only "the negative impact of civilization" as the theme, providing evidence of the ability to determine a theme or central idea of a text.	RL.2
The student selects one or more of the following options, and no others, providing evidence of the ability to determine how a theme is shaped by specific details: • **Supports "the negative impact of civilization" with "I heard of the division of property, of immense wealth and squalid poverty, of rank, descent, and noble blood."** • **Supports "the negative impact of civilization" with "I learned that the possessions most esteemed by your fellow creatures were high and unsullied descent united with riches."** • **Supports "the negative impact of civilization" with "A man might be respected with only one of these advantages, but without either he was considered, except in very rare instances, as a vagabond and a slave, doomed to waste his powers for the profits of the chosen few!"**	RL.2
The student correctly supports how the selected theme and sentence support the connection between the creature and the myth, providing evidence that he or she can analyze how an author draws on and transforms source material in a specific work.	RL.9
The student selects only "The creature's search for knowledge only leads to misery" and "The creature is rejected by both his creator and the cottagers," providing evidence of the ability to relate a work of fiction to the seminal ideas of its time.	RL.8.A
The student correctly analyzes how the quote supports the ideas in the myth and the excerpt, providing evidence of the ability to draw evidence from literary texts to support analysis, reflection, and research.	W.9
The student correctly analyzes how the quote supports the ideas in the myth and the excerpt, providing evidence of the ability to analyze a particular point of view or cultural experience reflected in a work of literature from outside the United States and to draw evidence from literary texts to support analysis, reflection, and research.	W.9 and RL.6

Sample Science Task

We offer a sample item cluster as an example of the clusters that we are building in other states and the capabilities that would be available to the Department under this contract.

Here we present a cluster measuring a middle school level performance expectation related to the cycling of matter and energy in the water cycle. The student will develop a model to explain that solar energy is driving the cycling of water. We begin with a phenomenon: fog regularly forms and then dissipates over the course of a morning in an Oregon valley. The phenomenon is communicated verbally and with an animation, as shown in Exhibit D1.2-27. The introduction and animation appear on the left side of the screen, and the items appear on the right.

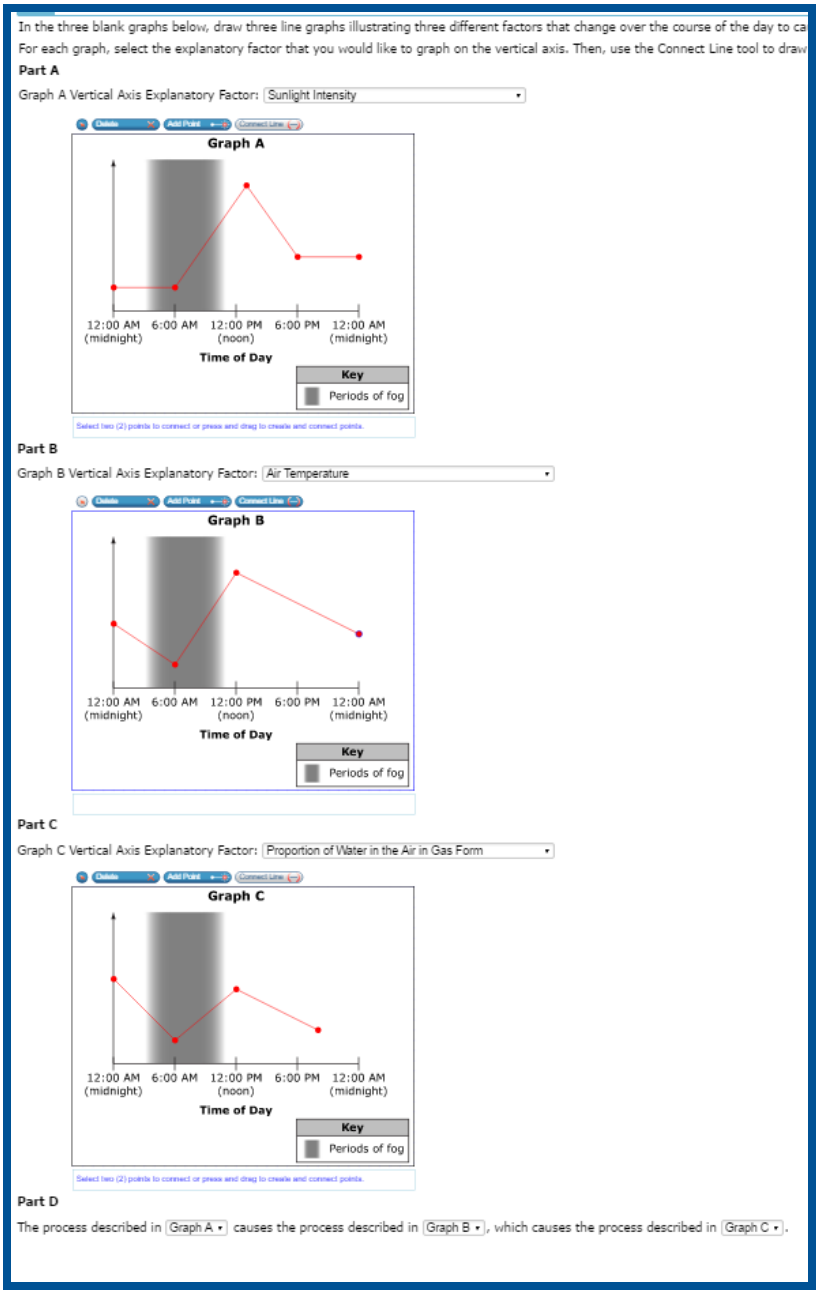
Exhibit D1.2-27: Sample Science Item



In this cluster, the student is asked to develop a mathematical model by identifying and graphing three factors that combine to create the phenomenon. Each empty graph has a 24-hour period on the horizontal axis. The period during which the fog is visible is marked on the graph. Using the drop-down menu, the student selects which factors to graph from a list containing distractors. Each graph is heuristic, rather than requiring specific quantities. Even though the student is asked to graph, the scoring rubric is looking for patterns reflecting conceptual understanding rather than mathematical understanding of the phenomenon.

Exhibit D1.2-28 illustrates one of many (virtually infinite) correct answers. The student should graph the amount of sunlight, the temperature, and the proportion of water in the air that is in a gas form. The final item asks the student to indicate the causal sequence of the fog's formation and dissipation. Note that students can graph the factors in any order, as long as the graphs have the right characteristics (for example, solar energy increasing over the course of the morning as the fog dissipates).

Exhibit D1.2-28: Items and Sample Answer



These interactions are actually a single item, and the scoring depends on the collection of responses rather than any single interaction. Our technology enables a scoring rubric to look across multiple interactions.

Using this approach, we engage students in actual scientific activities—in this case, modeling for the purpose of explanation. The performance expectation calls on students to actually employ a model, and they do that in this cluster. Moreover, they use a model that explains energy and matter transfers within part of the water cycle, thereby weaving in elements of all three dimensions of the performance expectation.

The questions are truly open-ended constructed-response items. *These items are also immediately and accurately machine scored.* Our tools allow our test developers to develop these sophisticated item clusters without requiring the assistance of software developers.

Finally, the features of the student responses that receive credit *and* the inference that the test developer would like to make from that evidence are explicitly captured as part of the item in the scoring assertions. Exhibit D1.2-29 presents the scoring assertions for this response to this item. These scoring assertions embody evidence-centered design as a physical part of the item.

Exhibit D1.2-29: Scoring Assertions for Fog Cluster

Score Result

Scoring Criteria	Your answer
The student chose sunlight intensity as one of the causal factors, thereby indicating an awareness of solar energy's role in the water cycle	✔
Graph of sunlight intensity shows increasing sunlight as fog ends, offering some evidence of an understanding that sunlight is providing the energy that ends the fog.	✔
The student chose temperature as a causal factor, thereby indicating an understanding that the heat energy is transferred to the atomosphere	✔
Student drew a graph showing decreasing temperature when the fog began to form, and rising temperatures when the fog dissipated providing some evidence of an understanding of that falling temperatures cause condensation, which appears as fog, and that rising temperatures cause vaporization ending the visible fog.	✔
The student graphed the proportion of water in vapor form, thereby providing evidence of recognizing that fog is condensation and the phase change to gas accounts for its disappearance	✔

OK

Interaction Types

The reader can see that AIR offers all of the interaction types called for in the RFP, plus many more. The mathematics task alone demonstrates a variety of selected-response interactions, along with an equation interaction, which allows for far more response types than the traditional "gridded" response. In English language arts, our sample task provides selected-response items, including "hot text," which requires the student to interact with the passage in order to link an inference to textual evidence. Short constructed-response items similar to the one in this sample task can be machine scored using our natural language tool, which matches the propositions in a student's response to an explicit rubric. Extended constructed-response items requiring a multi-paragraph response can be machine scored, using our AI scoring engine, which is further described in Topic 16 Machine Scored Items. The science task demonstrates our sophisticated, simulation-based items.

Topic 2.7 Number of Items

As we describe in Topic 1 Test Design, AIR proposes to offer our internally developed, ICCR item bank to assess student achievement of New Hampshire College and Career Ready Standards in ELA, mathematics and science. The ICCR bank provides comprehensive, robust item pools at each grade. The item pools were developed to ensure that ICCR assessments cover the full range and depth of the content standards at the aggregate level for each test administration. The bank grows larger each year as we continue to field test new items across multiple states. The multi-state participation in ICCR item development and test administration further strengthens the state-to-state comparisons afforded by the common ICCR scale.

Exhibit D1.2-30 provides the number of items that we anticipate to have in the operational pools at each grade in ELA and mathematics following field testing in spring 2017. (Note that these item counts do not reflect the attrition that may occur following item data review after field testing in spring 2017).

In addition, AIR has been developing science item clusters and standalone items to assess the three-dimensional science standards. Exhibit D1.2-30 also documents the number of clusters and standalone items that we anticipate will be eligible for administration in spring 2018. As we describe in Topic 18 Calibration and Scaling and Topic 19 Equating, we propose to administer the science items as an operational field-test design, in which a matrix of fixed forms, each conforming to blueprint specifications, enacts a balanced incomplete block matrix linking design to allow for concurrent calibration of all science items on a common IRT scale.

Exhibit D1.2-30: Item Counts at Each Grade in ICCR Bank Following Spring 2017 Field Testing

Pre-Equated Items Available for Adaptive Testing				Science Items Available for Operational Field Testing	
Grade	**Mathematics**	**ELA**	**Grade Band**	**Clusters**	**Stand-alone**
3	417	253	3−5	40	21
4	401	288			
5	381	279			
6	379	306	6−8	55	21
7	333	328			
8	381	255			
9	N/A	322	High School	67	21
10	N/A	341			
11	N/A	317			
HS	1,017	N/A			

The ELA numbers in Exhibit D1.2-30 show items in the ICCR item bank aligned to Reading Literature, Reading Informational Text, Language, and Speaking and Listening. In addition to these items, the ICCR bank also offers an extensive pool of writing prompts developed to support both informative and explanatory writing as well as opinion and argumentative writing. Exhibit D1.2-31 provides the number of writing prompts at each grade that have been field tested and are part of the operational pool.

Exhibit D1.2-31: Number of Writing Prompts in ICCR at Each Grade

Grade/Type	Number of Field Tested Writing Prompts
3	**6**
informative/explanatory	3
opinion/argumentative	3
4	**14**
informative/explanatory	7
opinion/argumentative	7
5	**15**
informative/explanatory	10
opinion/argumentative	5
6	**16**
informative/explanatory	9
opinion/argumentative	7
7	**16**
informative/explanatory	8
opinion/argumentative	8
8	**17**
informative/explanatory	10
opinion/argumentative	7
9	**14**
informative/explanatory	9
opinion/argumentative	5
10	**15**
informative/explanatory	9
opinion/argumentative	6
11	**14**
informative/explanatory	7
opinion/argumentative	7

Many of these writing prompts are accompanied by operationalized AI scoring models that are already in use in several states. Exhibit D1.2-32 shows the number of prompts at each grade in the ICCR bank with operationalized scoring models. Because we would not be using immediate scoring in year one, the Department could elect to administer some prompts that do not have scoring models. We would handscore these and build scoring models for future administrations.

Exhibit D1.2-32: Number of Writing Prompts with Operationalized Scoring Models

Grade	Number of Writing Prompts
Grade 3	6
Grade 4	6
Grade 5	6
Grade 6	6
Grade 7	6
Grade 8	9
Grade 9	8
Grade 10	6
Grade 11	6
Total: 59	

In spring of 2018, we anticipate field testing more than 1,200 new items in ELA, more than 400 new items in mathematics, approximately 165 science clusters, and about 65 stand-alone science items. As we mention in previous sections, we believe the Department will find the reading and mathematics item pools more than sufficient to support either our existing blueprint or a similar blueprint based on Department modifications. Our proposed spring 2018 operational field-test design will ensure that the full bank of science items is calibrated and equated to a common scale to support construction of psychometrically equivalent science matrix forms for future test administrations, as well as immediate reporting of test results.

Based on the ICCR model blueprints exhibited in Topic 1, Exhibit D1.2-33 summarizes the number of items that would appear on each assessment component. Please note that in ELA one item would be a text-based writing task.

We understand the state's desire for meaningful assessments that balance the need for a comprehensive measurement design while mitigating the sometimes onerous burden on schools in terms of testing time. In Topic 1.2 Test Administration we present 85th percentile testing times for Utah's SAGE summative assessments, which deliver test administrations based on blueprints similar to those we propose for New Hampshire's statewide assessments. We will work with the Department to finalize a set of blueprints that meet reporting requirements while limiting test administration times.

Exhibit D1.2-33: Number of Items on Each Assessment Component

Assessment Component	Grade	Number of Items on Assessment Component
English Language Arts	Grades 3−11	42
Mathematics	Grades 3−8	50
	Grades 9 and 10	40
Science	Grade 3	32
	Grade 5	37
	High School	46

Topic 2.8 Item Release

ELA and Math. AIR offers a half-length online practice test constructed with items from the ICCR bank. Many of these items were field tested in multiple states. They come complete with scoring rubrics and all the same metadata that our operational ICCR items and passages contain, including text complexity worksheets, standard alignments, distractor rationales, etc. Exhibit D1.2-34 and Exhibit D1.2-35 show the total number of items available at each grade in the practice test pool by item type.

AIR will work with the Department to identify more items each year from the operational ICCR bank for release. This activity will occur after the spring field test is complete and data review has occurred. AIR will first analyze the bank to ensure that items marked for release do not leave gaps in the item pools. We will then work with each participating state to agree on the set of items to be released each year to ensure that no states have included targeted items for future form construction.

Exhibit D1.2-34: Summary of ELA Items in ICCR Practice Test Pool

Grade	Multiple-Choice Items	Multiple Select Items	EBSR Items	Hot Text Items	Matching Items	Natural Language/Open Response Items	Editing Task Items	Extended Response (Writing) Items	Total Practice Test Pool
3	18	2	2	1	2	1	4		40
4	18	1	2	5	1	1	4	1	33
5	10	2	4	4		1	4	1	26
6	15	3	4	2		1	4	1	30
7	17	3	1	3	1	1	4	1	31
8	14	1	4	4	1	1	4	1	30
9	15	2	4	1	2		4	1	29
10	15	6	2	1	1		4	1	30
11	15	1	4	1	1		4	1	27

Exhibit D1.2-35: Summary of Math Items in ICCR Practice Test Pool

Grade	Multiple-Choice Items	Multiple Select Items	Equation Items	Grid Items	Matching Items	Table Match Items	Hot Text Items	Total Practice Test Pool
3	4	3	22	2		1		32
4	5	6	15	3	1	3		33
5	10	5	17	1				33
6	9	3	13	3		3		31
7	7	3	19	2	1	1		33
8	12	4	11	5	1			33
Algebra	9	2	10	3	1			25
Functions	9	1	10	2				22
Geometry	9	4	12	1		1	2	29
Number & Quantity		0	3					3
Statistics & Probability	5	2	4	1		1		13

In an effort to ensure that the practice test has the same look and feel as the summative test, we have provided the following accommodations:

- English TTS for all ELA and math items
- Spanish TTS for all math items
- Spanish translations for all math items
- Braille for all ELA and math items that are deemed brailleable
- ASL videos for ELA audio stimuli
- Glossaries for ELA and math items

Science. AIR also offers a training test in science that demonstrates some of our most innovative item clusters designed to measure three dimensional science standards. We are working toward building a more extensive practice test that will be available later this year. We will continue field testing science items in a variety of states to ensure that a sufficient number of items can be released each year from the operational pool to support updates to the practice test.

Topic 2.9 Rotation of Common Items

As we describe in Topic 19 Equating, AIR proposes to use pre-equated item parameters to score and report results of New Hampshire's new statewide assessments. The adaptive test administrations that we propose for ELA and mathematics require the use of pre-equated item parameters to enable adaptive item selection. But even beyond adaptive test administrations, when tests are administered online and scoring of test items is automated, there is an expectation in the field, among students, educators, and parents, that assessment results will be available immediately after testing or very quickly thereafter. When processing of answer documents and hand-scoring of written responses precluded immediate scoring and reporting of test results, post-equating allowed for increased precision of item parameter estimates without further delaying the reporting of assessment results. However, with the advent of online test administrations and automated scoring, the slight advantages accrued by post-equating may be outweighed by delays in score reporting.

We do note, however, that because the science item clusters and items have yet to be calibrated and placed on a common scale, we also plan to implement a balanced incomplete block matrix common-item linking design to allow items to be calibrated on a common scale. Because reporting of test results for all subjects will not be possible until after completion of standard-setting activities in summer 2018, the calibration and equating activities will not further delay reporting of science test results.

The ICCR item pools are relatively large and growing, reducing the exposure of individual items in the population. Moreover, when psychometricians configure the adaptive algorithm, in addition to evaluating blueprint match and score precision, they also evaluate item exposure to alert test developers to gaps in the item pool that may lead some items to be administered more frequently than desirable. Generally, we expect each item to be administered to fewer than 20% of students. Exposure rates greater than this represent targets for future item development.

Similarly, the matrix design we propose for administering the science assessments limits the exposure of individual clusters and stand-alone items. Each disciplinary core includes many more standards than can be realistically assessed for each student.

The matrix design has three major benefits: it ensures that students within a classroom are administered the same blueprint specifying the number of clusters/items per DCI but the clusters/items administered in each form will measure different standards within the DCI; it assesses the full range of standards across the classroom; and it limits the exposure of clusters/items.

As we describe in Topic 19 Equating, before we calibrate the field test items following summative test administrations each spring, operational items are evaluated for parameter drift. Items exhibiting drift are recalibrated and equated back to the reference scale, and the refreshed linked item parameters are used to score subsequent test administrations.

As we discuss in Topic 2.8 Item Release, AIR coordinates with all ICCR participating states to identify items for release each year so that items targeted for release are mutually agreed upon. Criteria for releasing items include exposure of test items, availability of items in the pool to meet blueprint requirements, and changes in the performance of test items over time, including item fit, that may indicate a compromise in item security.

Topic 2.9.1 Item Bank

AIR's content development process is managed using our Item Tracking System (ITS), which serves the following critical roles:

- A customizable content development and management tool
- An item bank
- A publication system that supports both paper and online publications

ITS relies on a series of distinct and specific review levels throughout the content development process. The system shepherds the content entities that AIR develops through every workflow step, from inception to final publication; it does so by forcing those entities to pass through targeted review, editing, and management status levels in which specific actions are taken or reviews are conducted. (Exhibit D1.2-20 in Topic 2.5 Content Review Item Review Committees illustrates this review sequence.) ITS enforces a standardization of process in content areas on all entities developed in those areas; in doing so it facilitates version control and management oversight and preserves the development history of the entity.

ITS manages our item development and review workflow, and in the process compiles a comprehensive history of the reviewer comments, changes, and review outcomes.

ITS shepherds each item through a series of required reviews. It programmatically selects a random sample of the work of each team member and inserts it into a queue to be reviewed by a senior test developer. These reviews provide the catalyst for discussions between senior staff and each team member to ensure that item quality remains high, items adhere to the item specifications, and end products adhere to AIR's vision for quality.

ITS also helps us track item usage history. Each time an item is published for use in a form or in an adaptive pool, ITS records the usage, either on operational tests or on released versions. Released or retired items are moved to special review levels that prevent their inclusion on operational forms. See Exhibit D1.2-36 for an example of the forms history that ITS maintains. Each use of this item was tracked with the role, form ID, and date on which it was administered. In this case of this item, it was developed for a practice test and went directly into operational usage.

Exhibit D1.2-36: Sample Forms History

10 **Associated Forms**

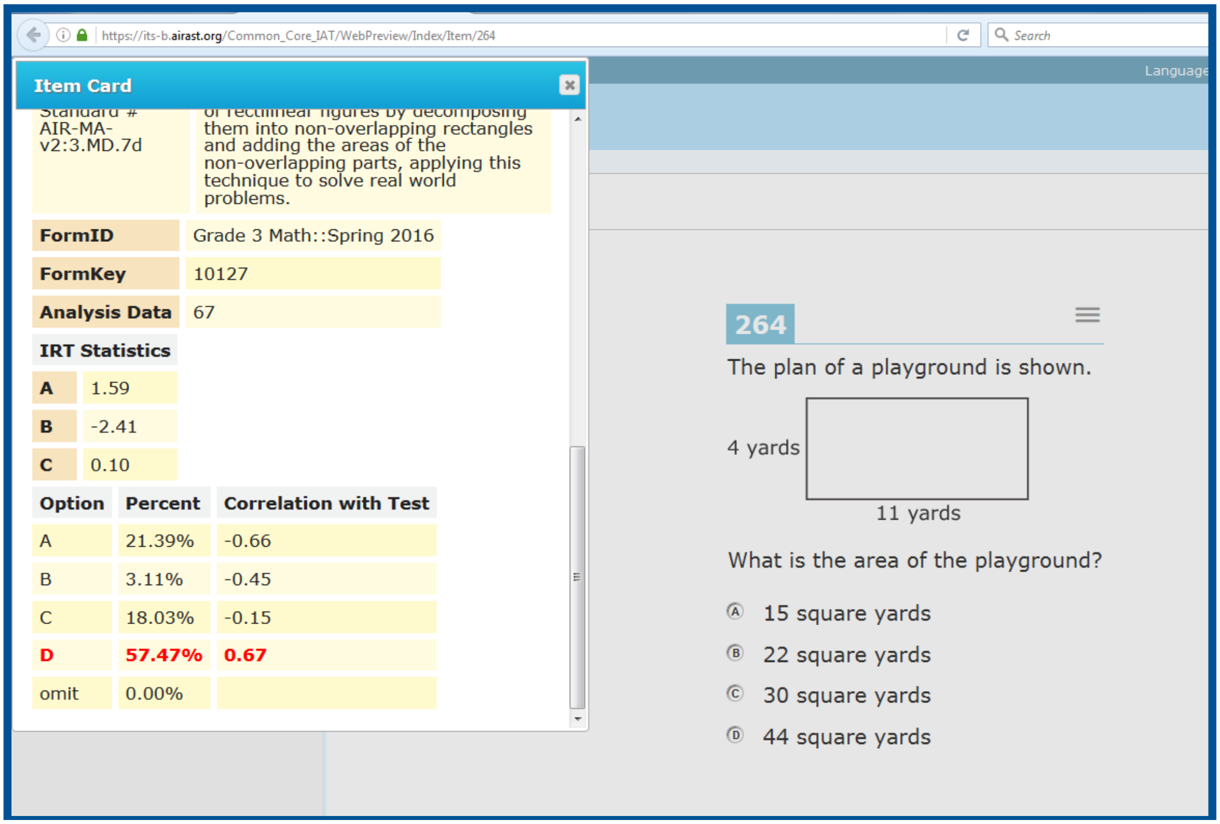
Item Role	Form ID	Updated By	Admin Start	Active
Operational	Math7and8TrainingFall2013	Paul_Maxon	11/1/2013	Yes
Operational	MG7and8 Fall 2015	Susan_Sherwood	11/1/2013	Yes
Operational	MG7and8 Winter 2017	Joseph_Kuhn	1/30/2017	Yes

Using ITS, our test developers maintain electronic versions of every item, including its many iterations, as it moves through the development cycle. ITS archives each prior version of every item, requires documentation of item edits, and enforces agreed-on item review levels to track an item's movement through the development cycle. This item development workflow offers the ability to manage items from inception, through the series of content, fairness, data, and other reviews (as described in Topic 2.4 Bias/Sensitivity Review Committee), to final publication. The system captures the outcomes and rationales at each review step and maintains previous drafts of each item. The workflow management ensures that each item receives each review in the designated sequence, and that the review is conducted (or recorded, in the case of committee review) by an authorized person. Every version of every item is archived, along with each comment received in any review, including the reviews conducted by state clients. Reviewers have immediate access to all older versions, providing version control throughout development. All attributes of the item, including accommodation and accessibility tags, item statistics, and rubrics, are associated with the item and travel with it through each review step.

The metadata about each item or passage, which we call *attributes*, are fully configurable and can capture any data that we wish to maintain. For passages, ITS attributes generally include general passage information, such as grade level, genre, and readability measures. In addition, all information about sources, including copyright information and attached permission documents, are stored with the passage in ITS. Item attributes also include key blueprint-relevant metadata, such as grade level, content alignment, and depth of knowledge. In addition, we include item origin (for tracking any shared items), item formats, and item scoring attributes. These attributes allow ease in tracking and maintaining information about the item bank. At any time, if a client has questions about the item pool, we can run a summary table based on any attribute to quickly and accurately get a sense of the number of items in a particular category.

Item statistics are also available for any items that have been field tested. Item cards are part of the item in ITS and can be accessed at any time simply by clicking "Item Card" at the top of the screen. The item card renders beside the item so that users can see the item statistics and the item content simultaneously. Exhibit D1.2-37 provides a sample item card from a released item in the ICCR bank.

Exhibit D1.2-37: Sample Item Card from Released Item in ICCR Bank



The New Hampshire instructional support teams and management teams will have access to ITS and will be able to view items and generate these summary reports. The item bank can be accessed easily and securely through any location with Internet access. Our encryption software maintains bank security, and only approved users with login credentials can access the bank.

Topic 2.9.2 Use of Items from Other Sources

The security of the ICCR bank is extremely important to AIR. Our Item Tracking System (ITS) helps ensure the security of items by tracking their movement through the entire development process, including review by state clients and their stakeholder committees. Only users with approved credentials can log in to the bank and review the items. AIR tracks each state's use of items and passages through an attribute in ITS that identifies which states have selected them. This allows us to monitor the use of items/passages in each state and ensure that items remain in each state's secure operational pools.

Every year, we earmark items and passages from the ICCR bank for release. Our test developers first check the item pools in each affected state to verify that there would be no negative impact due to the release of the selected items and passages. Then, we share our selections with each state and request their explicit approval before releasing the items. The items are then moved into released status in ITS (both in the ICCR bank and the state bank) where they are no longer be eligible for use on secure tests.

AIR will offer ICCR items to the Department for use on multiple administrations. We will maintain the security of these items throughout their operational use just as we do for other states that use the ICCR bank. When items are selected for release, the Department will have the opportunity to review those items well in advance of the release date.

D1.2 Solution Technology

Topic 3 Technology Requirements

Introduction

AIR proposes to offer its full suite of services, providing to the Department AIR's test delivery platform, hosting site, test administration application, server, and application management services for the New Hampshire summative and interim assessments. AIR will also perform all required maintenance on the system components described below.

AIR is the leader in online testing for statewide assessment with approximately 30 million online tests administered in each of the 2014–2015 and 2015–2016 school years. Approximately 18.8 million of these test administrations were secure, summative assessments, and the vast majority of these (12.3 million) were adaptive. AIR has recently supported more than one-half dozen states, including New Hampshire as they moved almost instantly from an all-paper system to a virtually 100% online. Moreover, we assisted New Hampshire in moving students that require Braille and large-print accommodations to our online testing platform. We are proposing the same solution technology that New Hampshire already has in place to successfully deliver the Smarter Balanced assessments. Currently AIR's systems support New Hampshire students with nearly 100% of all Smarter Balanced Assessments administered online. In addition, we note that the College Board has recently selected AIR's test delivery system to administer the SATs. Thus, should the Department award AIR the contract for the 3–8 assessments, New Hampshire's grade level and high school assessments will be delivered on a common platform.

Software Requirements

AIR's test delivery system (TDS) has the smallest possible technology footprint in the industry and low end-user maintenance. We designed our system to work in the confines of real-life schools. These schools often have limited bandwidth, older hardware, and limited or nonexistent technical support. Our system is purely Internet-based and supports most platforms as long as their manufacturers support them, and often longer depending on the prevalence in schools. For example, we support Windows back to XP, which has been out of manufacturer support since April 2014, and Mac OSX 6, which has been out of manufacturer support since July 2011.

Installation of the system requires only the installation of our secure browser. The secure browser can either be installed from a network on many machines, or it can be individually installed. We can even offer installation that does not require administrative rights. Our browser works with existing equipment and bandwidth and requires little or no technical expertise, enabling entire states to go online quickly and smoothly.

AIR's solution is infinitely scalable and currently configured for nearly 1.5 million concurrent testers, with peak loads this year expected to be around 700,000–800,000 simultaneous users. Last year, peak loads exceeded 500,000 simultaneous users.

Our online testing system is accessible for all students. We have received Web Content Accessibility Guidelines (WCAG) 2.0 AA certification, a very high level of certification for accessible web applications. Our system boasts the most comprehensive set of tools, embedded supports, and accommodations currently in operational use in the nation. We are committed to accessibility for all students. The accessibility of our assessments is described in further detail in Topic 1.4 Accessibility and Fairness.

AIR's Test Delivery System

The key to our success in helping states to transition and maintain the administration of online assessments has been our respect for the infrastructure that exists in schools. We have designed our system with a small footprint. Our system works on the newest tablets as well as the oldest desktops, because we know that schools must work with the technology available. School technology should support teaching and learning, and assessments should work within that infrastructure. We consider our approach to test delivery to be device agnostic.

Our assessments require very little bandwidth, despite rich interactions and stimuli. They are engineered this way because we know that even schools with good Internet connections can sometimes have internal bottlenecks that throttle bandwidth. Most importantly, our assessments are designed to require virtually no technology expertise in schools. We know that schools and districts often lack access to skilled technicians. Our system requires only a single piece of software—the AIR secure browser—and no special hardware. The secure browser can be installed quickly and simply, either on a single machine or across a large network. Once installed, it takes care of itself. Exhibit D1.2-1 provides a list of supported operating systems and minimum system requirements.

Exhibit D1.2-1: Supported Operating Systems and Minimum System Requirements

Supported Technology	Minimum System Requirements
Windows Operating Systems	XP (Service Pack 3), Vista, 7, 8.0 Pro, 8.0 RT, 8.1, 10 Server 2003, 2008, 2012
Macintosh Operating Systems	10.6–10.11
Linux Operating Systems	Fedora 20–22, openSUSE 13.1, Red Hat Enterprise 6.5, Ubuntu (LTS) 10.04, 12.04, 14.04
Chrome Operating Systems	41–46, 48, 49, excluding v.47, which had manufacturer defects
Thin Clients	Win2003 and 2008 terminal servers, Ncomputing vSpace 6.6.2.3 with a host OS of Win 7
Disk Space	200 MB
Processor	Intel Pentium 4 or newer processor that supports SSE2
Memory	512 MB
Screen Size	10”
Screen Resolution	1024 x 768 or larger
Peripherals	Headsets/earbuds and mice are strongly suggested but not required.
Connection	Wired or wireless supported

Each of our private clouds supports hundreds of thousands of simultaneous users. Currently, each of our constellations is designed to support 400,000–500,000 students signing on within 20 minutes of each other and testing simultaneously. Across constellations, our capacity approaches 1.5 million simultaneous users. Each year, we project peak loads and can scale any constellation out as much as necessary. Last year, peak loads exceeded 500,000 simultaneous users. Peak loads in 2017 are expected to be around 700,000– 800,000 simultaneous users. Our system delivers a mean refresh time of less than one second. The peak number of students in New Hampshire testing at any given time in 2016 was 6,835.

On the other hand, technology is advancing rapidly. AIR is committed not only to keeping up with those advances but leveraging them to make test content more meaningful and accessible. Thus, AIR needs to ensure our system always has forward browser compatibility with the latest operating systems, including iOS, Android, and Chromebook devices, as well as assistive technology devices. Exhibit D1.2-2 provides the list of operating systems for new devices that AIR presently supports.

Exhibit D1.2-2: Operating Systems for New Devices Supported

Supported Technology	Minimum System Requirements	
Windows Tablets	8.1 (Pro)	Surface Pro, Asus Transformer, and the like
iOS Tablets	7.0, 7.1, 8.0–8.2, 9.2, 9.3	iPad 2, iPad 3, Fourth Generation (Retina Display), iPad Air, iPad Air 2
Android Tablets	4.3, 4.4, 5.0,5.1	Google Nexus 10, Motorola Xoom, Samsung Galaxy Note (2014 edition), Samsung Galaxy Tab (3 and 4), LearnPad Quarto
Screen Size	10"	
Screen Resolution	1024 x 768 or larger	
Peripherals	Headsets/earbuds and mice are strongly suggested but not required	
Connection	Wired or wireless supported	

AIR's test administration, delivery, and reporting systems reside on more than 800 servers in three data centers with multiple and redundant Internet connections and power supplies. AIR conducts extensive load testing each year on each of AIR's private clouds. Each of our private clouds supports hundreds of thousands of simultaneous users.

AIR balances testing loads across hundreds of servers in our private clouds. AIR's testing platform typically requires under 15 Kbps per student, making online testing an option for every school and ensuring comparable performance on wired and wireless networks across platforms and operating systems.

AIR's test delivery and reporting systems are protected against catastrophic events. Our systems reside at Rackspace, which implements redundant Internet lines, power supplies, cooling, and other necessary redundancies. Data is backed up nightly to a separate Rackspace site. All of our servers are redundant, and data are continuously being backed up during testing.

For over a decade, AIR has worked with many states to train and support districts, schools, and test coordinators in test administration requirements and procedures. While educators are often nervous when transitioning from a familiar testing system, our experience has been that training and support can ease fears and facilitate a smooth transition. Refer to Topic 13 Training Materials, and Topic 15 Software Implementation Training, for more information about training.

As we describe in this section, AIR makes available a diagnostic tool that schools are encouraged to use to perform a system check that evaluates whether the available bandwidth is sufficient to support the planned number of test-takers. In cases where bandwidth is not sufficient, the tool provides the number of test users that can be supported. AIR will also work with the Department to identify target schools and districts for additional support should the state desire.

Minimum and Preferred Technology Infrastructure Needed to Support Online Testing

As explained above, installation of the system requires only the installation of our secure browser. This browser can either be installed from a network on many machines, or it can be individually installed. We can even offer installation that does not require administrative rights. Our browser works with existing equipment and bandwidth and requires little or no technical expertise, enabling entire states to go online quickly and smoothly. Minimum system requirements are outlined in the explanation above; however, we note that we consider our approach to online testing to be device agnostic. Our system works on the newest tablets as well as the oldest desktops, because we know that schools must work with the technology available. School technology should support teaching and learning, and assessments should work within that infrastructure. We have designed our system with what we refer to as a small footprint.

AIR can help schools assess if they have sufficient infrastructure for online testing in two ways:

1. Using a diagnostic tool to be run on the machines
2. Assessing if the specifications of the student machines are supported

As mentioned, AIR's system has a minimal technology footprint. Very little infrastructure is required to enable a school to test online with our system. For example, installing proctor caching servers at the district or school is not required. The only requirements are

- sufficient bandwidth (10–15 kbps per student);
- sufficient machines;
- a secure browser; and
- minimal training.

Suggested Computer Lab Configurations.

We designed our system to work within the confines of real-life schools. Schools often have limited bandwidth, older hardware, and limited or nonexistent technical support. Our system is a purely Internet-based system and supports most platforms as long as their manufacturers support them, often longer depending on the prevalence in schools. For example, we support Windows back to XP, which has been out of manufacturer support since April 2014, and Mac OSX 6, which has been out of manufacturer support since July 2011. AIR's testing platform has a very small footprint in schools, working with existing equipment and bandwidth and requiring little or no technical expertise. Our system typically requires under 15 Kbps per student, making online testing an option for every school, and ensuring comparable performance on wired and wireless networks and across platforms and operating systems. Before testing, schools simply install our secure browser or stand-alone apps for iOS, Android, or Chrome OS. This can either be done from a network on many machines or individually installed. We can even offer installation that does not require administrative rights.

Diagnostic Tool

Our system has a web-based tool that performs a system check to assess the readiness of a particular computer for testing. One of the tools is a download and upload speed check that estimates the number of simultaneous test-takers who can test at the same location. This tool has proved useful to other K–12 programs.

This tool can be run on all platforms supported by AIR's test delivery system (TDS). Please see Exhibit D1.2-1 and Exhibit D1.2-2 for a list of the supported platforms.

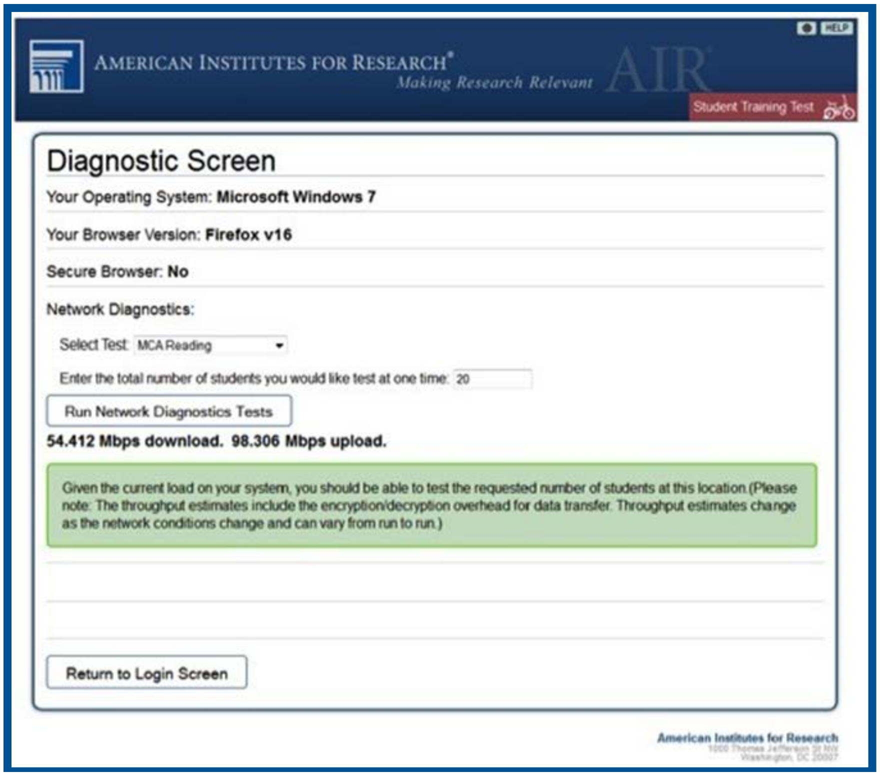
Typically, no network configuration is required at the school. Sometimes school districts or schools will only allow computers to access certain white-listed IP addresses. In these cases the schools must white-list our IP addresses. Unusual setups, such as certain proxy server arrangements, may need adjustment to allow traffic to pass appropriately to our servers or to prevent proxy servers from caching content that may be updated. We provide simple instructions to address these rare cases.

Maximum Concurrent Users

The diagnostic readiness tool uses a sophisticated statistical model of the testing process that models variation in instantaneous demand (e.g., the number of students who simultaneously press the Next button) and evaluates the likelihood that peaks will exceed network free capacity (not used by other processes) with a frequency likely to cause noticeable delays in testing. Exhibit D1.2-3 presents the user interface for our integrated diagnostic tool.

If the tool indicates that a school's intended number of users exceeds its current capacity, it will indicate the number of concurrent users that can be supported with existing resources.

Exhibit D1.2-3: Simple, Effortless, Accurate Diagnostic Tool



It is a good idea for schools and districts to coordinate a trial run to ensure that there are no unknown bottlenecks in the school or district. While such problems are extremely rare, it is better to catch them before operational testing. A trial run has the added benefit of making sure that test administrators (TA) and students have accessed the system before operational testing.

We recommend that schools (and districts, when Internet access is centralized through the district) coordinate a practice test day. On this day, TAs and students go to the computer lab as they would on an operational test day and access the practice test as though it were a real test.

Our servers capture information about network performance, so any schools or districts that experience trouble can contact us for rapid diagnosis of any issues.

Computer-Based Assistive Technologies

As described in Topic 1.4 Accessibility and Fairness, AIR's TDS currently supports a wide array of assistive technologies. AIR is working to expand the classes of assistive technologies supported. Our streamlined interface adheres to WCAG 2.0. We have achieved WCAG 2.0 AA certification, which means that the system is accessible and compatible with standards-compliant assistive technology. In addition, we know that our security mechanisms can interfere with some assistive technologies, and we have built in a permissive mode that relaxes these security restrictions for individual students who need to use such technologies.

AIR provides text-to-speech support with configurable text-to-speech tracking. The computer reads text and graphics aloud on directions, passages, and items. What is read and how it is read are configurable, and the Department can offer a variety of options. AIR's approach to text-to-speech uses well-specified guidelines to annotate the items for speech and generate the speech on the student machines. We use the voice packs on the machine (and can support any compliant voice pack supported). This does result in some variability in intonation and pronunciation from machine to machine; however, we have taken steps to mitigate the impact of these differences. In particular, the one popular operating system that offers a stock voice pack that is below our expectations is Windows XP. To mitigate this problem, we provide a high-quality Julie Voice Pack from NeoSpeech that can be downloaded and installed on Windows machines at no cost to the Department. Since Windows users who use this tool should already have this voice pack installed on their Windows devices from the current year test administration of Smarter Balanced in New Hampshire, this should impose no additional work on the behalf of users.

The system currently works with a wide range of refreshable Braille devices as well as screen readers, on-screen keyboards, and a wide array of input devices. While it is impossible for any organization to guarantee support for unknown hardware and software, AIR is committed to accessibility for all students and has demonstrated that commitment through our leadership in this area.
As stated in the requirements of this RFP, AIR can collect data on the use of assistive technologies for use by the Department.

Practice and Training Tests

A practice test is available for each grade and each subject. Each item on the practice test has a tutorial that can be accessed to help the student learn to interact with the item type. The availability of practice and training tests to allow students to become familiar with keyboard commands, navigation techniques, and tools used during interim and summative assessments is described in Topic 14 Practice Tests and Student Materials.

Uploading Student Demographic Data and Necessary Accessibility Tools and Supports

Student registration processes including the uploading of student demographic data are elaborated upon in Topic 1.3 Student Registration.

Processes for uploading accessibility tools and supports are detailed in Topic 1.4 Accessibility and Fairness.

Security of the Online Testing Environment

AIR understands the gravity of the responsibility that comes with handling student data. We have measures in place to ensure that data are accessed only by authorized individuals and that our systems housing sensitive data are well protected. As previously described, within the schools, the only software required is AIR’s secure browser. This browser locks down the computer, preventing the student from navigating away from the test or starting other software. It also disables keystrokes that can threaten the security of the test. For example, the secure browser disables screen captures and navigation and prevents test-takers from viewing the source and opening the taskbar. It continuously monitors other activity on the computer for possible threats and terminates testing if a threat is detected.

AIR fully recognizes the importance of protecting user and student information in all its various states, whether being transmitted, stored, or converted. AIR has built-in security controls in all its data stores and transmissions. Unique user identification is a requirement for all systems and interfaces. All of AIR’s systems conduct external data transfers on an secure file transfer protocol (SFTP) connection, and all communication with any of the online websites occurs over secure sockets layer (SSL). This ensures that passwords and secure student data are not passed over the network in clear text.

Security of test items and student information is protected at all times, with security procedures operating at three levels:

1. Physical security, preventing access to the machines on which data reside or are processed

2. Network security, including protection of our networks from infiltration and the secure transmission of data over our networks and others
3. Software security, ensuring that only authorized users access information on our systems and that their access is limited only to information they are authorized to view

Physical Security

State data will reside on servers at Rackspace, our hosting provider. Rackspace maintains 24- hour surveillance of both the interior and exterior of its facilities. All access is key card controlled, and sensitive areas require biometric scanning. Access credentials are assigned only for authorized data center personnel, and only they have access to the data centers. Visitors' identities are verified, and visitors are escorted at all times while in the facility. All data center employees undergo multiple background security checks before they are hired.

Secure data will be processed at AIR facilities and will be accessed from AIR machines. Similar security procedures are in place in our facilities. Access to our facilities is key card controlled. Visitors must sign in and be escorted while in our center. Our servers are in a secure, climate-controlled location with access codes required for entry. Access to our servers is limited to our network engineers, all of whom, like all AIR employees, have undergone rigorous background checks.

Staff at both AIR and Rackspace receive formal training in security procedures to ensure that they know and implement the procedures properly. AIR and Rackspace protect data from accidental loss through redundant storage, backup procedures, and secure off-site storage.

Network Security

Hardware firewalls protect our networks from intrusion. They are installed and configured to prevent access for services other than hypertext transfer protocol secure (HTTPS) for our secure sites. Our firewalls provide a first level of defense against intrusion, backed up by a capable second line: hardware and software intrusion detection and remediation. Our intrusion detection systems constantly monitor network traffic and raise alerts for suspicious or unusual network traffic.

Our systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts. Suspicious log entries are investigated and resolved.

All secure data transmitted across the public Internet are encrypted using secure shell (SSH) advanced encryption standard (AES), or an Internet protocol security (IPSec) virtual private network (VPN). Secure websites encrypt data using 128-bit SSL public key encryption.

The hosting environment is protected by an AlertLogic Threat Manager Intrusion Prevention System (IPS) appliance at the perimeter and by Symantec Antivirus Corporate Edition on each individual server. The AlertLogic IPS appliance combines intrusion protection and vulnerability management technology into a single integrated solution that offers both proactive and reactive protection from the latest threats. Symantec Antivirus offers real-time virus and malware protection for our servers along with centralized management and administration capabilities.

Software Security

All of our secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with New Hampshire's privacy laws, the Family Educational Rights and Privacy Act (FERPA), and other federal laws.

Our systems implement sophisticated, configurable privacy rules that can limit data access to only appropriately authorized personnel. Different states interpret the FERPA differently, and we support customized interpretations. Our system is designed to support these interpretations flexibly. AIR is committed to working with the Department to maintain data security according to its specifications.

Secure transmission and password-protected access are basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins.

Our systems use role-based security models that ensure that users access only the data to which they are entitled and that limit their ability to change those data according to their rights. User rights have two dimensions: the user's role and the user's data access rights. The role determines what actions a user can take, which types of reports he or she can view, and similar functional limitations. Data access rights tell, for example, which principal can view which teacher and student data. Data access rights are governed by relationships among entities in our Roster Tracking System (RTS), along with a configurable set of business rules. The business rules describe which access rights correspond to which relationships and enforce client-specific policies. We will work with the Department to maintain current business rules and adjust business rules as needed.

Training Protocols at the Local Level on Test Administration Procedures

Training for test administrators at the local level is discussed in Topic 13 Training Materials.

Regional Trainings on System Use and Test Administration Procedures Supplemented by an Online Webinar and other Online Training Materials

Regional trainings, webinars, and training modules for system use and test administration procedures are discussed in Topic 13 Training Materials.

Technical Support

Technical support is accessible via the dedicated Help Desk through phone, e-mail, or a live chat function for the duration of the contract. This is discussed in Topic 38 Help Desk Support.

Metrics for Monitoring and Documenting Systems Performance

We refer readers to Topic 10 Backup and Recovery for our approach to monitoring and logging systems performance.

Capacity of the System to Support Current and Future Range of Items Types

Our item development and test delivery system teams work tirelessly to produce innovative item types. The current item types are offered accessibly and perform flawlessly in our test delivery system. AIR strives to innovate to produce future item types that fully exploit the capabilities of systems technology and our test delivery system as they evolve together. For more information, we refer readers to Topic 2.6 Item Types and Number of Items.

Provide documentation regarding the application's capacity to import and export as applicable: items, student item response data, student registration, demographics, and data regarding eligible and utilized accommodations.

Items

We refer readers to Topic 2 Item Development for all information regarding the items we offer.

Student Item Response Data

We have a proprietary in-house system that allows authorized users to access student item response data in order to see how a student answered any item as it appears in our test delivery system. All authorized users understand the importance of confidentiality and comply with FERPA and other student privacy laws and take measures to protect student's personally identifiable information (PII). If requested, the AIR New Hampshire project team can provide the department with access to student responses through a secure viewing platform, the Student Item Response Viewing Engine (SIRVE). AIR would be please to provide more information about SIRVE to the Department upon award of the contract if its use interests the Department.

Student Registration, Demographics, and Data Regarding Eligible and Utilized Accommodations

As discussed in Topic 1.3 Student Registration, AIR supports the import and export of student registration and demographic data.

In the Test Information Distribution Engine (TIDE), it is possible to export a report of assigned tools, accommodations, and supports that differ from the default test settings for an individual student or in batches by school or district. AIR can produce a report of the accommodations that were used in the TDS by students in a school or district, if necessary.

Topic 4 Technology Requirements

AIR's Test Delivery System

AIR is pleased to propose a proven online testing solution to assess student proficiency for New Hampshire's Statewide Assessments in ELA, mathematics, and science. Educators will be familiar with the TDS because it is the same solution being used to deliver New Hampshire's current Smarter Balanced ELA and mathematics assessments. Moreover, the College Board has recently selected AIR's online testing system to administer the SATs, allowing the possibility of single platform for delivery of New Hampshire's grade level and high school assessments.

AIR offers a robust online TDS that meets and exceeds the Department's requirements. AIR continuously invests in research and development, helping all our clients improve measurement techniques, extend the range of skills that can be effectively measured, extend the accessibility of the system, and improve how features are deployed. Our active research and development program has enabled us to offer the industry's most accessible platform with the most item types and industry-leading quality assurance and monitoring tools to keep the system running smoothly. The investments that we are making today will extend these and other benefits to our clients tomorrow.

In this section we describe many of the features of our online TDS. AIR's system has the proven operational capabilities to deliver the full range of New Hampshire's statewide assessments. Some of the benefits of the AIR systems include the following:

- A robust, online TDS that has proven effective, requiring minimal equipment, bandwidth, expertise, or support, including the ability to deliver tests on iOS and Android tablets and Chromebooks
- A student experience that empowers students with the broadest range of access technology available in a statewide system while also offering a simple, accessible interface free of distraction
- The industry's richest repertoire of item types, using technology to probe deeper into students' understanding by asking them to construct responses graphically using language and simulations
- The industry's broadest set of available machine-scoring (sometimes called *artificial intelligence*) engines

Exhibit D1.2-1 and Exhibit D1.2-2 list the minimum requirements for all of AIR's systems.

System Summary

The AIR online testing system is made up of a set of six integrated modules. Shared access to all modules is managed by OpenAM, which enables these systems to appear to users as a single, integrated system. Once logged in, users can navigate the various components of the system securely and seamlessly. Exhibit D-1.2-4 summarizes each module and its role in the overall system. Below, we provide an overview of the key components of the system and their capabilities, followed by a detailed description of each component in subsequent sections.

Exhibit D1.2-4: Summary of Online Test Delivery Components

System	Description
Test Information Distribution Engine (TIDE)	TIDE is responsible for • student registration; • gathering of demographic data; and • materials ordering.
Test Delivery System (TDS)	The TDS's Test Administrator Interface (TA Interface) is the interface through which test administrators establish and monitor testing sessions and authenticate student users. The Student Interface is the testing system on which students take tests. The TDS delivers tests to students, records responses, and forwards data to downstream systems.
Quality Monitor (QM) System	The QM system receives the data, verifies the validity of the test administered and the scores assigned, gathers statistical data for ongoing quality reports, and forwards data to the database of record (DoR) and the Online Reporting System (ORS).
Database of Record (DoR)	The DoR maintains the authoritative records of tests administered and completed. Data in the DoR reflect appeals, verifications, and other post-administration adjustments to the data.
Online Reporting System (ORS)	The ORS provides a secure interface to assessment data and associated demographic information. It provides educators with a powerful tool to explore the data and turn them into actionable information.
Administrative Portal	The administrative portal provides a customized, centralized location for educators and other stakeholders and provides access to the other systems as well as practice sites, training sites, tutorials, and other resources.

System Description and Capabilities

To administer tests, the TDS needs information about students and test administrators (TAs), including authentication information. TIDE, our registration system, gathers data from districts, schools, or the state and stores those data in the RTS, a flexible database system shared by TIDE, the TDS, and the reporting systems. The RTS will maintain data about the educational networks in New Hampshire, such as which schools are in which districts, which teachers are in which schools, and which students are in which classes. TIDE maintains fully configurable data about the attributes of the various entities in the system, such as school addresses, student demographics, and virtually any other information that the state would like reflected in the final data delivered. TIDE also maintains authentication and role data, so it knows which personnel are authorized to administer tests and can allow them to log in and define sessions. Student and user enrollment processes are elaborated on in Topic 1.3 Student Registration.

After a test is administered to a student, the TDS passes the resulting data to our Quality Monitor (QM) system. The QM system rescores tests, checks that the tests meet blueprints, captures statistics on items, and runs a host of extensive quality checks. The QM system also runs a suite of analyses designed to detect cheating, which our psychometricians can access at any time. The entire quality checking process occurs in milliseconds, and the data are then passed to the electronic reporting system. In the rare event that the QM system identifies an anomalous test result, members of the project team are immediately notified via e-mail, and the results are held until verified.

From the QM system, data are passed to the database of record (DoR), which serves as our consistent repository of data. The extract system, which will interface with the state's systems to deliver test and item data, will draw data from the DoR. The DoR will send data to the Response Bank, our data

system that supports the ORS, as well as to our Score Reporting team. Shared access to the components of the system is accomplished by OpenAM, which enables a user to log in to any one of the systems—excluding the Item Tracking System (ITS)—and access any other system for which he or she is authorized.

OpenAM is an open-source access management solution and a federation server platform that is highly flexible and offers industry-standard security capabilities, including strict and rigorous password management, data encryption, and a scalable architecture. AIR will work with the Department to ensure that best security practices are enforced throughout the login and authentication processes.

Administrative Portal

Public-Facing Administrative Portal

AIR will maintain an administrative portal as a central point of access for New Hampshire stakeholders, including district test coordinators, school test coordinators, TAs, teachers, families, and students. The portal will house links to all online practice and training tests as well as the supporting materials required for test administration. The structure of the portal may be organized according to role (test coordinator, teacher, technology coordinator, student/parent) and separate, subject-specific tabs such as Secure Browsers, Training, General Resources/Manuals, FAQs, and Contact Us. The portal tabs function like a table of contents, helping users organize and locate information as needed. The portal homepage also offers an easy-to-access location for important announcements and updates. Aside from functioning as a central information repository, the portal also provides each user role with password-protected access to only those components of the online testing system that are relevant to each specific role. AIR staff will work with the Department to configure the portal in the most optimal way for administrators, coordinators, and educators.

We include a discussion of the portal in our draft agenda for the first day of the initial kickoff meeting as discussed in Topic 32 Management Meetings and as indicated in Appendix J. At that time, AIR will present alternative designs for the portal for Department review. A member of the AIR design team will be present at this meeting and will work with the Department to modify designs and settle on a final design.

To see the current design of the portal that AIR has created and maintained for New Hampshire's Smarter Balanced assessments, see http://nh.portal.airast.org/.

Secure Administrative Portal for Registered Users

TIDE is the secure administrative portal that also handles student and user enrollment, test registration, and scheduling test windows. TIDE is discussed in greater detail in Topic 1.3 Student Registration.

Test Registration and Scheduling

Student registration through AIR's TIDE is discussed in Topic 1.3 Student Registration.

Test scheduling may also be managed in TIDE. Annually, the Department can establish a statewide testing window in TIDE. Each district can customize testing windows in the state-level window for each school in their district. TIDE is configurable to allow district testing windows to be modified if changes are made to the district school-year calendar based on missed instructional days due to unforeseen circumstances (e.g., snow days). Testing windows may be created or edited manually or, if a district has many testing windows to create or edit, performing those transactions through a file upload using a pre-defined comma-separated value (CSV) or Microsoft Excel template may be less cumbersome.

Test Administration (Administrator Interface)

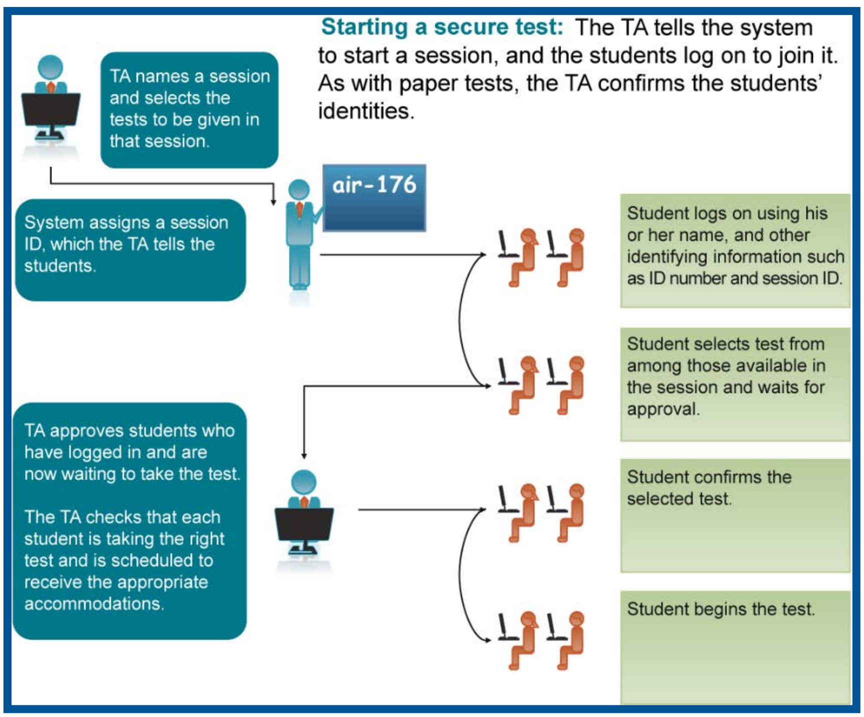
Once testing begins, the TA can monitor student progress and, if necessary, stop one or all students from testing via the TA Interface. The system is designed to automatically notify a TA of events that require an action on his or her part. For example, students with the appropriate accommodation may want to print a passage, and such requests alert the TA to print the passage.

The list of students currently participating in the session shows the TA who is actively testing, which test each student is taking, and how many items have been delivered to each student. The TA can also view approved individual student accommodations by clicking the binocular icon, pause and restart individual tests or the entire session at any time, and view the reason that a test is in paused status.

Test Administrator Interface

AIR's testing platform is easy to use for test administrators to create and manage test sessions for all online assessments in the TA Interface. This interface allows authorized TAs to administer test sessions, monitor activity, and respond to test-related issues in one convenient location. The secure interface helps ensure that the right student is taking the right test and lets the administrator focus on test administration. As with paper tests, the TA's primary role with online tests is to ensure the security of the test and make sure that students have a quiet environment, free of distraction, in which to take the test. The security of the test is managed through a simple interaction between the TA and the students, as summarized in Exhibit D1.2-5. Students can test only when the TA is logged on and has an active session. After a student pauses a test, the TA must readmit that student to the session.

Exhibit D1.2-5: Starting a Secure Test



Creating a Test Session

TAs create a session simply by selecting the tests that will be available in the session. The system assigns a session ID, and the student login is simply a request to enter a session. When students join the session, they can select from among the tests that they are offered in the session and those that they are eligible to take. In this manner, AIR's TDS offers TAs a seamless and integrated means of administering assessments in a single, easy-to manage space. Students can be admitted to any assessment based on the parameters established by the TA for a particular test session. The process, platform, and interactions between the TA and the student remain the same regardless of the assessment being delivered.

Students use the session ID to log in to a test session. Session IDs are unique to each test session. When a student enters the session ID, the system links the student requesting access to the test session to the TA, who then approves or denies the student's access request. When students join a session by logging in and entering the session ID, the TA verifies their identity and admits them to the session. Once admitted to the session, students are given a final chance to confirm that they have requested the correct test. Upon confirmation, the test begins.

Reviewing and Approving Test Settings

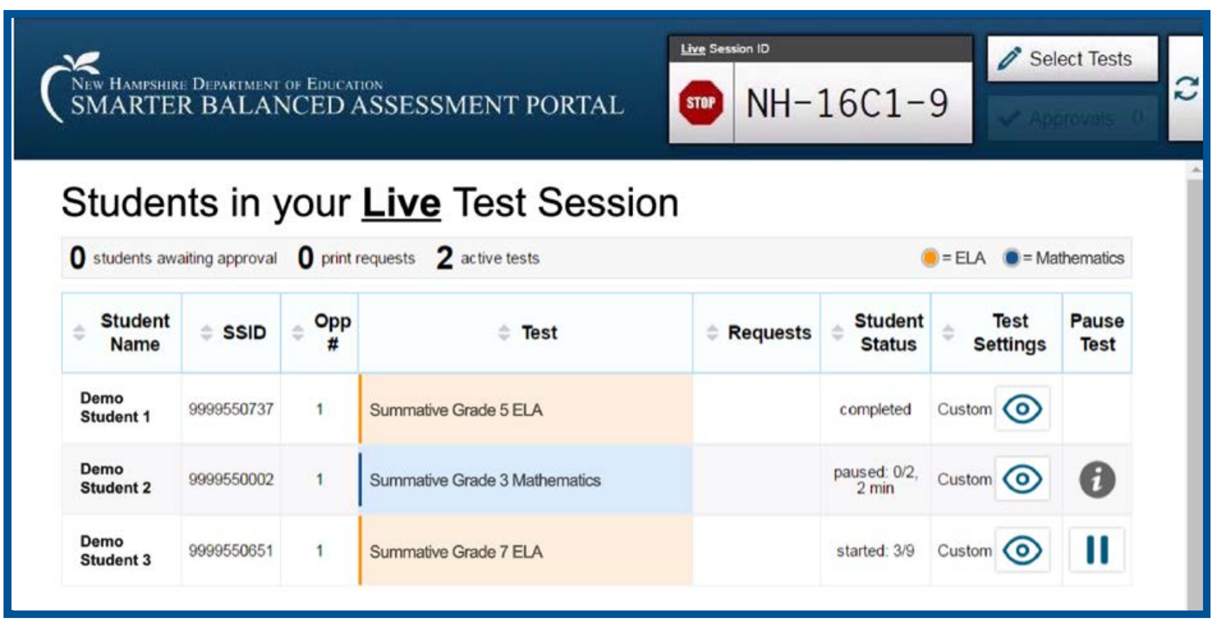
During the authentication process, TAs have an opportunity to review or adjust the test settings. These settings include any embedded supports or accommodations to which the student may require access. These can, and should, be set in advance in TIDE; however, we recognize that schools do not make these adjustments in advance. Hence, we make it possible for TAs to assign some or all accommodations (at the discretion of the Department).

The availability of access features, embedded supports, and accommodations is completely configurable. The state will be able to determine whether TAs can adjust settings at the beginning of the session or whether access to specific features requires higher-level authorization. For example, one state provides TAs with broad authority to adjust most features, but only authorized state users can authorize the read-aloud of reading passages.

Monitoring Students and Other TA Tools

Once testing begins, the TA can monitor student progress and, if necessary, stop one student or all students from testing. The system is designed to notify a TA automatically of events that require an action on his or her part. For example, students with the appropriate accommodation may want to print a passage, and such requests alert the TA to print the passage. Exhibit D1.2-6 shows the TA Interface during an active test session with three approved students.

Exhibit D1.2-6 Test Administrator Interface during an Active Session



During an active test session, the TA can add tests to offer in the session. This feature is useful, for example, if a student joins a testing session to make up or complete a test that had been offered earlier.

The list of students currently participating in the session shows the TA who is actively testing, which test each student is taking, and how many items have been delivered to each student.

The TA can also view approved individual student accommodations by clicking the binocular icon, and he or she can pause and restart individual tests or the entire session at any time. Student actions that require TA intervention appear in the "Requests" column. For example, our secure print-on-demand feature prompts the TA to authorize the request and to retrieve the paper from the authorized printer.

Test Delivery (Student Interface)

Students take secure tests through the Student Interface, which is essentially a secure website accessed through AIR's secure browser. This browser is the only software needed to take a secure test, and it is simply a secure build of the Mozilla (Firefox) browser or a secure testing app for tablets and other platforms.

AIR's secure browser operates in a full-screen mode, disables access to other applications, and prohibits navigation outside the test. The browser is designed to intercept all operating system hot-key combinations and print capabilities; it enables keyboard combinations specifically designed for test navigation. The system verifies that the test is being launched in AIR's secure browser and prevents the test taker from continuing if the test is launched in a normal browser window.

The Student Interface is designed to be intuitive and make it easy for students to track information on a single screen. The interface incorporates many tools and accessibility features, and various layouts are available for items that require reading passages, graphics, or other item stimuli. (We refer readers to Topic 1.4 Accessibility and Fairness, for a discussion on tools and accommodations.) While the layouts are completely configurable, we encourage states to select layouts that allow students to read left to right and top to bottom as they naturally would with a paper-pencil test. For example, most states, if they are using a passage along with a group of items, choose to put the passage on the left and the items on the right.

Our system offers complete flexibility regarding graphics layout and other media in text. Anything that can be formatted on a web page can be correctly displayed in our system. Our passage layout is configurable, allowing either scrolling or paging. An additional feature allows users to devote additional screen area to the passage while reading it. Similarly, the items associated with a passage may be laid out vertically and scrolled through or displayed on separate pages.

All tests will be taken using the same secure browser. In fact, the only logistical distinction occurs when the TA selects which assessment will be administered during a particular test session. The experience is seamless for students and TAs, using a single system to deliver a diverse array of content to all students.

Students are always presented with the same authentication sequence and interactions with the TA, regardless of the test being administered. To access a test, a student logs in using his or her name and other identifying information such as ID number and test session ID. This information is displayed on both the student's and TA's workstations.

Student Interface Layout

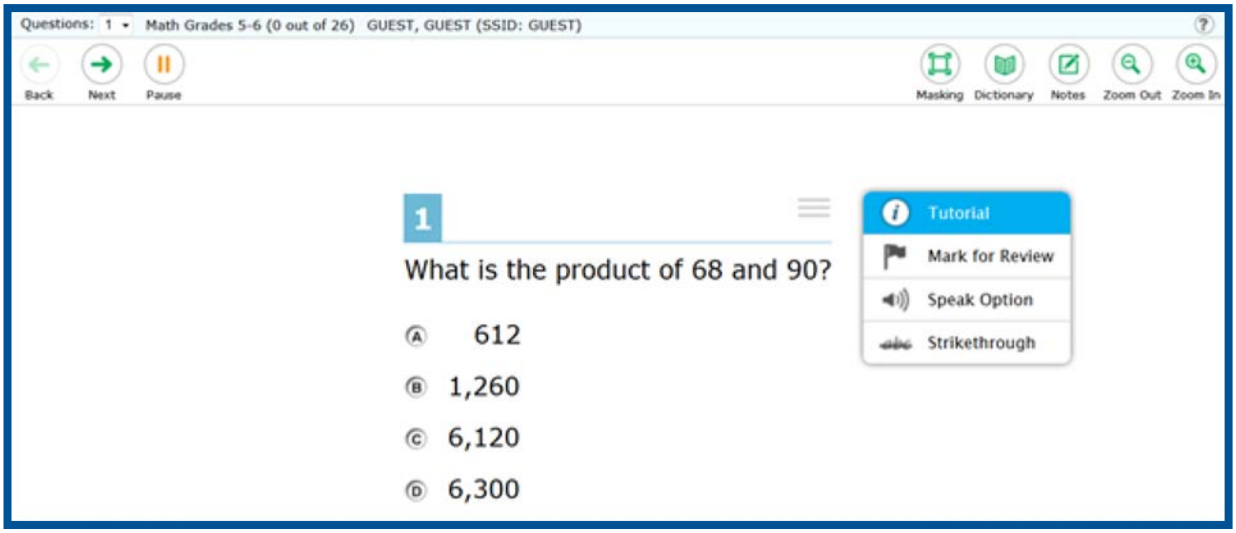
The Student Interface begins with a simple look and feel. This look and feel is customizable. Each test (grade, subject) can have its own defaults for font, layouts, or other components. The interface is not cluttered with tools or options. In general, the interface is simplified in two ways:

1. The selection of accessibility tools or accommodations is moved to precede the test. AIR usually recommends that the selections be made at registration time or by the TA at testing time. Some tools are configured to be available to all students taking a particular test.
2. Context menus are used to allow students to access the various tools and supports. This serves two purposes. First, it is part of AIR's strategy to make sure the entire system is

keyboard navigable using few keyboard commands. This strategy is necessary to support alternate input devices, such as switch arrays, that provide access for some students. Second, it supports all students by making all the relevant tools and options available in the context in which they are relevant, helping students to find what they need when they need it. Because students test only rarely, we cannot and do not expect them to become expert users of the software. Rather, we support students' navigation of the testing environment by presenting available tools and options when and where they are available.

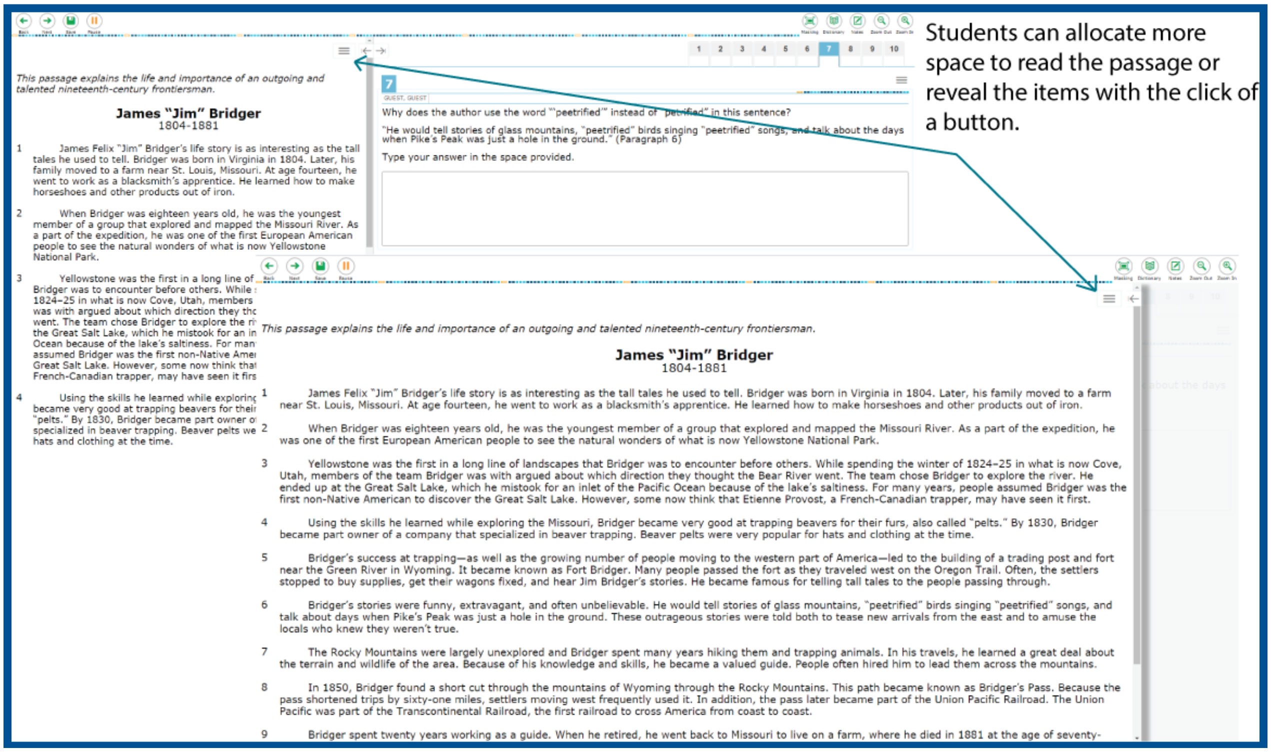
Exhibit D1.2-7 illustrates AIR's context menu strategy. One click (or keystroke) brings up the menu. The menu presents all the relevant options in the context. In this case, the menu includes viewing the tutorial, marking the item for review, using text-to-speech (TTS) for the question, and selecting a strikethrough option. When selecting text on the screen, students may also have the option of highlighting the text or hearing TTS for the selected portion of the text.

Exhibit D1.2-7 Context Menus Ensure Keyboard Accessibility and Find-as-Available Functionality



The system includes a host of small, intuitive features that help students navigate and use the test without getting distracted. For example, the expand-passage function allows the student to allocate more room to the passage as he or she reads, while still having immediate access to the test items. While the system can display one item in a set at a time, good testing practice presents students with the information they will need for reading the passage before (or along with) the passage itself. The ability to expand the passage reduces the load on working memory and reduces the need to scroll the page. Exhibit D1.2-8 shows an example of this capability.

Exhibit D1.2-8: Expand Passage Function



AIR's pure HTML5 Student Interface supports the industry's widest set of devices, ranging from desktops and laptops to tablets and Chromebooks. Using standard HTML5, we can deliver sophisticated animations, simulations, multimedia, and a wide range of input types. Exhibit D1.2-9 presents an online performance task that includes an HTML-5 simulation, which asks the test taker to perform an experiment and interpret the resulting data.

Exhibit D1.2-9: Integration with HTML-5 Animations and Simulations

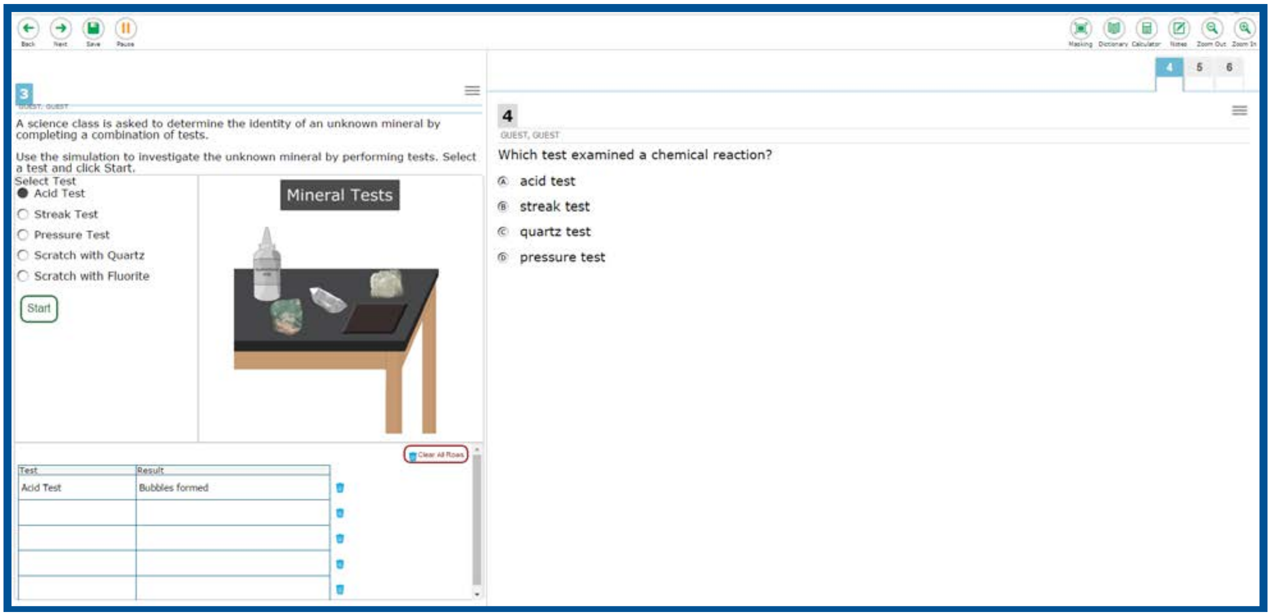
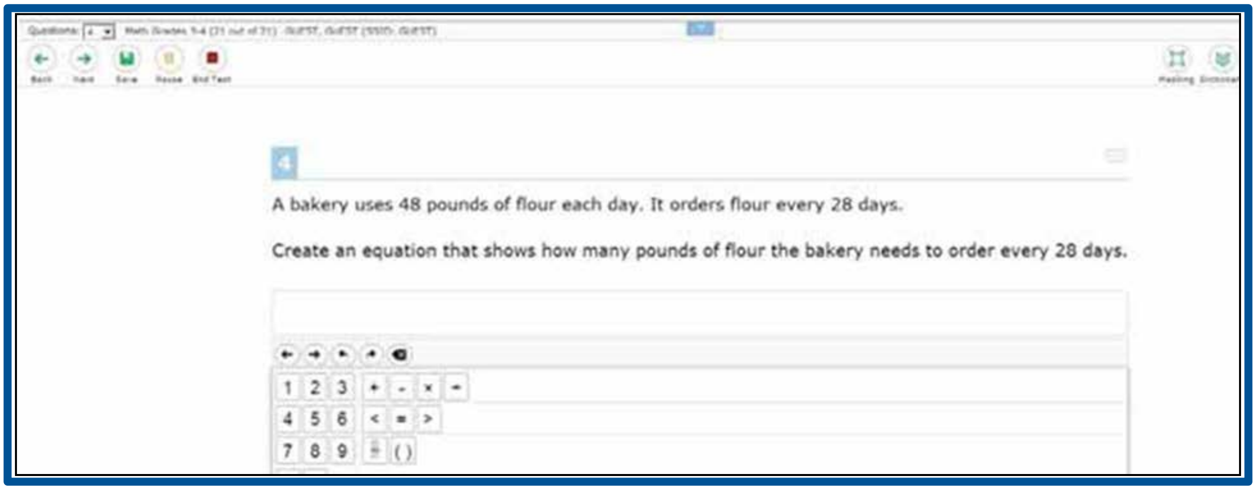


Exhibit D1.2-10 presents an item that asks the test-taker to write an equation using the HTML5 equation editor. The student response can be machine-scored using symbolic math logic to evaluate whether any expression the student writes is mathematically equivalent to the correct answer. The equation-scoring engine can also evaluate whether the response has the correct functional form, even though it might not be correct, and whether there are other properties for which partial credit might be awarded. The latter capability can be helpful in lower grades where the content may not demand mathematical precision.

Exhibit D1.2-10: Equation Editor Item



Test Client

Students access the assessments using our secure browser (test client), the only piece of required software. The secure browser is discussed in greater detail in Topic 4 Assessment Delivery Platform.

Key-Based and Rule-Based Machine Scoring

We refer readers to Topic 16 Machine-Scored Items, for a discussion of the key-based and rule-based machine-scoring processes.

Assessment Delivery Data Storage

We refer readers to Topic 10 Backup and Recovery, for a discussion of the data storage processes in place to protect assessment delivery data storage.

Student Toolset (e.g., Virtual Calculators, Protractor, Ruler, Notepad, Highlighter)

We refer readers to Topic 1.4 Accessibility and Fairness, for a discussion of available tools and accommodations.

Authentication/User Identity Management

TIDE, as well as all other AIR systems, enforces strict authentication policies. TIDE adds, edits, and deletes users and sets them with different authorizations based on roles. Since role creation is configurable, AIR will work with New Hampshire to define enough differentiated roles to support workflows at the schools and districts throughout the state. Currently, TIDE is configured for New Hampshire users to allow the roles of District Test Coordinator, School Test Coordinator, Teacher, and TA. AIR is prepared to keep the current designated user roles or configure new roles should the Department prefer.

Login information generated when adding users is communicated securely to the individual users. Passwords are not e-mailed to users. Instead, upon addition to TIDE, users are sent a secure link, one that expires after a configurable time, to a site where they set their password. Passwords in the TIDE databases are encrypted. A host of application-level security controls safeguard against unauthorized access to items or student data. Across all systems, our applications are password-protected.

AIR will work with New Hampshire to establish a user role hierarchy. Based on this hierarchy, TIDE can be configured such that users at higher levels can delegate tasks of adding users to other roles. A typical setup follows:

- State-level users can upload district-level users.
- District-level users can add school test coordinators.
- School test coordinators can add teachers and TAs.

Logging and Audit

Our TDS logs most events in the system, tracking when tests are started, paused, restarted, or stopped. It logs each change in an item response and many other system activities. TIDE keeps a complete record of changes to student and school attributes, so we can reconstruct that information for any point in time.

Hardware firewalls protect our networks from intrusion. They are installed and configured to prevent access for services other than Hypertext Transfer Protocol Secure (HTTPS) for our secure sites. Our firewalls provide a first level of defense against intrusion, backed up by a capable second line: hardware and software intrusion detection and remediation. Our intrusion detection systems constantly monitor network traffic and raise alerts for suspicious or unusual network traffic.

Our systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts. Suspicious log entries are investigated and resolved.

Logging and audit procedures are outlined in greater detail in Topic 10 Backup and Recovery.

System Monitoring and Alerting

Our systems monitor the network and all attached devices. We have monitoring at the network, server, and software levels. The system sends alerts to two network operations centers: Rackspace and AIR's Network Operation Center (NOC). When alerts hit specified levels, our NOC activates our emergency notification system, which assembles a team of engineers, managers, and leadership to respond. Usually, problems are resolved before they impact the field.

In addition to traditional monitoring, AIR has developed a predictive monitoring system that integrates information from all of our server sensors as well as data on software latencies. These data feed a sophisticated model that identifies servers that are behaving differently than they have in the recent past plus other servers performing the same roles. Servers that exhibit anomalous behavior are audited. Such audits have found disks with bad (but still working) clusters, poorly optimized but rarely called procedures, and other problems before they had any impact on system performance.

Our staff continuously monitor the functioning of the system. Our NOC monitors many server metrics and alerts our engineers and senior leadership if an anomaly arises. Part of the monitoring is our unique dashboard, which provides a visualization of the health and performance of every server in a constellation. This can identify anomalous behavior long before it becomes a problem in the field. We look forward to the opportunity to demonstrate the dashboard.

Usually, our monitoring systems enable us to prevent problems before they impact testing. In the rare event that this is not the case, our systems log errors to rapidly diagnose problems. If a problem arises, the project manager will immediately alert the Department by phone and, working with the Department, information can be disseminated to the field through

- announcements on the administrative portal and
- a messaging system that alerts TAs who are logged on or who attempt to log on to the TA Interface.

Common ID System

Typically, states maintain unique identifiers for each student. We accept this information in enrollment data, and encrypt and store it along with other sensitive personally identifiable information (PII). Each student is then assigned a unique AIR identifier for internal purposes and is identified in our systems using that anonymous key. We understand that New Hampshire assigns new student identification numbers to each student at the beginning of each test administration and AIR has been able to successfully support this approach to student identification in our TDS.

When student information is reported, all systems must access the single system hosting the PII to decrypt that information. In this way, we can always link test results to their real student identities, but we limit the risk to sensitive information.

The system ensures the proper identification of students and accurate matching of student identifiers to test results using the local student ID and a unique system-assigned ID to accurately identify student test records and demographic/accommodation information.

Detailed Description of the Interfaces and the System components Used for Processing.

A detailed description of various AIR online testing systems and interfaces is included at the beginning of Topic 4 Assessment Delivery Platform.

Topic 5 Data Exchange and Process

AIR works with each of our clients to specify automated data exchange protocols. AIR will provide data file deliveries to New Hampshire using the processes specified in this section at agreed-on intervals.

A typical protocol specifies file formats and business rules for importing data from state data systems as well as file formats and business rules for preparing files to be exported to state data systems. Our most common mechanism for transferring data is an automatically monitored SFTP site, with two-way data transfers occurring each evening.

We will work with the Department and your IT staff to specify file formats and transfer protocols that integrate seamlessly with your systems. The procedures for transferring data to and from the automatically monitored SFTP site are as follows. AIR currently transfers Smarter Balanced assessment data files from AIR to the Department for the 2016-2017 administration using our proposed SFTP file transfer method. During the summative testing window, this data file transfer occurs nightly. Following the close of the testing window, AIR delivers a final data file to the Department. AIR is prepared to continue to use this process for data exchange with the Department or we would be glad to work with the Department to implement new data exchange procedures.

All transmission of secure data with the Department will take place using SFTP. When the transferred data include student PII, the transfer software encrypts the files upon upload to the site. The files remain encrypted while they are at rest, and SSL technology encrypts the data in transit.

We implement a two-part procedure to prevent access to student PII, even in the unlikely event of a breach. Our strategy is to limit exposure of the identifying information and to encrypt those data where they absolutely must appear. Most of our systems carry only an AIR-assigned, arbitrary student identifier for each student. Only a single system links that to the student's statewide identifier and name and that system encrypts those fields. Systems needing this information request the encrypted data. The key needed to decrypt these data is stored on a server without any external access and is accessible only to the servers running our systems. Our systems never write the key anywhere. Internally, only very few engineers have access to this key.

In short, we stop bad actors from entering our system, and if they get past this defense, even if they steal data, the data will likely be devoid of PII. If they are lucky enough to obtain the data file with PII, they will find it encrypted, and they will not find the key anywhere.

Topic 6 Data Privacy and Security

Data Privacy

All of AIR's systems protect individual privacy and confidentiality in a manner consistent with the Department's privacy laws, FERPA, COPPA, FISMA, and other federal laws. All secure data transmitted across the public Internet are encrypted using SSH, AES, or an IPsec VPN. Secure websites encrypt data using 128-bit SSL public key encryption. When data are stored, they reside securely on database servers behind multiple firewalls and are secured through an encrypted connection.

Key PII such as student names and identifiers are stored only in our RTS database. These data are encrypted, with encryption keys being very closely held. Other systems store only an arbitrary AIR-assigned ID. When student data must be matched with identifiers (e.g., when presenting individual results in the ORS), the encrypted identifiers are merged onto the data and decrypted in real time.

Our systems implement sophisticated, configurable privacy rules that can limit data access to only appropriately authorized personnel. Different states interpret FERPA differently, and we support customized interpretations. AIR is committed to working with the Department to maintain data security according to its specifications.

For data encryption in transit, all data are encrypted using SSL. For data encryption at rest, PII is encrypted using a double-encryption scheme in which the passphrase itself is encrypted and stored on a server accessible only to our systems. We ensure that FERPA is implemented properly in all stages of the student data.

Ensuring Student Data Integrity

This section describes the many mechanisms at work to ensure the integrity of the student data. The system uses record-level locking when that is sufficient to prevent deadlock and race conditions, but at times broader segments of tables must be locked to ensure data integrity. Restores from backup are only done manually, because automated recovery risks data corruption. Our system has a sophisticated set of warnings to alert users to unusual conditions, and the system is backed up nightly, with full backups of key testing data.

We have disaster recovery plans in place, which can recover data from most failures. Catastrophic failure of an entire data center will not be recovered from immediately. It will be possible to recover the data to another data center, although this will require manual processes that take more time. Failure of any component will be recovered from automatically.

AIR warrants the functionality of our software. Flaws will be reported immediately and remedied in the shortest feasible time. We use up-to-date standards for all development and adhere to strict security guidelines.

Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered real-time backup processes prevent loss of student data, even in the unlikely event of system failure. We refer readers to Topic 10 Backup and Recovery for a full discussion of our disaster prevention and recovery plan.

Topic 6.1 Information Technology Standards

All AIR solution technology proposed in D1.2 Solution Technology is currently used for New Hampshire's Smarter Balanced assessments and complies with New Hampshire's information technology standards.

Open Source Software

We note that we consider our test delivery system (TA Interface and secure browser) to be proprietary software. However, the secure browser is simply a special build of the Mozilla (Firefox) browser that is modified to protect the security of the test and supports certain accommodations or a downloadable mobile app for mobile platforms.

AIR's single sign-on system, OpenAM, is an open-source access management solution and a federation server platform that is highly scalable and offers industry-standard security capabilities, including the following:

- Automated generation of secure passwords as soon as users are provisioned
- Forced password change after the user logs in for the first time with the system generated password
- Enforcement of passwords of certain length (e.g., a mix of numbers, letters, special characters)
- Secure way to reset passwords
- Configurable security questions for resetting passwords

OpenAM enables AIR's systems to appear to users as a single, integrated system. Once logged in, users can navigate the various components of the system securely and seamlessly.

Open Data Format

AIRCraft

AIRCraft is AIR's industry-leading item authoring tool, and the AIRCraft format is APIP compliant and validates against the QTI/APIP 2.2 standard. AIR's internal format for items in AIRCraft includes compliant extensions to support interaction types, accessibility features, layout, and other features not natively supported by the QTI/APIP standard.

Item Tracking System

ITS can import items from virtually any format, but we would write custom import routines for proprietary item formats. We export items in QTI/APIP 2.2.

Test Delivery System

AIR's test delivery system can natively deliver items represented in QTI/APIP format. However, APIP has several limitations that will prevent the Department from exploiting the full suite of accessibility capabilities offered by our test delivery platform. For example, AIRCraft allows words in the content to be tagged with definitions to allow the delivery engine to present multilingual thesaurus entries and dictionary definitions with embedded audio. Similarly, AIRCraft supports a variety of text-to-speech representations for tagged content (literal, descriptive, etc.), allowing the delivery engine to use the student's accommodations and test-level configurations to dynamically determine how to synthesize content to speech. Other accessibility features that we support, such as closed-captioning and transcripts, are not currently available in APIP.

We have a structured XML extension to APIP that supports representation of the advanced accommodations that go beyond those imagined in the APIP specifications. We would be pleased to share this with the Department upon award of the contract.

ARIA Tags

Students using assistive technology such as screen readers and refreshable Braille displays need to be able to take the online New Hampshire Statewide Assessments. The end user experience is heavily dependent on the accessibility markup embedded in the item content as ARIA tags. QTI does not require ARIA tags nor do most item authoring systems include them in their content.

Our item authoring tool, AIRCraft, ensures that ARIA tags are built into the content from inception and ensures that the items are accessible from the start. For example, when tables are used in content, AIRCraft requires the item writer to specify whether the table is used to present data or used to present regular content in a tabular layout. For data tables, AIRCraft requires item writers to provide table summaries that describe the table so that, for example, visually impaired students using screen readers can get an overview of the table before the table data are read aloud. For layout tables, AIRCraft marks up the table so that screen readers do not read them out to students as columns and cells. Our system includes a tutorial with each item. QTI/APIP does not provide a structure for this.

Compliance with Privacy and Confidentiality Standards

All of AIR's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA), COPPA, FISMA, and other federal laws. Our systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. Different states interpret the FERPA differently, and we support customized interpretations. Our system is designed to support these interpretations flexibly. AIR is committed to working with the Department to maintain data security according to its specifications.

All secure data transmitted across the public Internet are encrypted using SSH, AES, or an IPsec VPN. Secure websites encrypt data using 128-bit SSL public key encryption. Stored data reside securely on database servers behind multiple firewalls and are secured through an encrypted connection. Secure transmission and password-protected access are basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins.

Key personally identifiable information (PII), such as student names and identifiers, is stored only in our RTS database. These data are encrypted, with encryption keys very closely held. Other systems store only an arbitrary AIR-assigned ID. When student data must be matched with identifiers (e.g., when presenting individual results in the Online Reporting System [ORS], the encrypted identifiers are merged onto the data and decrypted in real time).

Our systems use role-based security models that ensure that users access only the data to which they are entitled and that limit their ability to change those data according to their rights. User rights have two dimensions: the user's role and the user's data access rights. The role determines what actions a user can take, which types of reports the user can view, and similar functional limitations. Data access rights tell, for example, which school test coordinator can view which teacher and student data. Data access rights are governed by relationships among entities in our Roster Tracking System (RTS), along with a configurable set of business rules. The business rules describe which access rights correspond to which relationships and enforce client-specific policies. We will work with the Department to maintain current business rules and adjust business rules as needed.

Security of test items and student information is protected at all times, with security procedures operating at three levels:

1. Physical security, preventing access to the machines on which data reside or are processed
2. Network security, including protection of our networks from infiltration and the secure transmission of data over our networks and others
3. Software security, ensuring that only authorized users access information on our systems and that their access is limited to information they are authorized to view

Student privacy and our compliance with FERPA, COPPA, FISMA, and other federal laws in addition to other system security features are discussed further in Topic 9 System Security.

Topic 7 Technical Compatibility

AIR's TDS has the smallest possible technology footprint in the industry and low end-user maintenance. We designed our system to work within the confines of real-life schools. These schools often have limited bandwidth, older hardware, and limited or nonexistent technical support. Our system is purely Internet-based and supports most platforms as long as their manufacturers support them, and often longer depending on the prevalence in schools. For example, we support Windows back to XP, which has been out of manufacturer support since April 2014, and Mac OSX 6, which has been out of manufacturer support since July 2011.

On the other hand, technology is advancing rapidly. AIR is committed not only to keeping up with those advances but to leveraging them to make test content more meaningful and accessible. Thus, AIR needs to ensure our system always has forward browser compatibility with the latest operating systems, including iOS, Android, and Chromebook devices, as well as assistive technology devices. Exhibit D1.2-2 and Exhibit D1.2-3 provide the list of operating systems for new devices that AIR presently supports.

As stated in Topic 3 Technology Requirements, installation of the assessment system requires only the installation of our secure browser. This browser can either be installed from a network on many machines or it can be individually installed. We can even offer installation that does not require administrative rights. Our browser works with existing equipment and bandwidth and requires little or no technical expertise, enabling entire states to go online quickly and smoothly. Minimum system requirements are outlined in the explanation above; however, we note that we consider our approach to online testing to be device agnostic. Our system works on the newest tablets as well as the oldest desktops, because we know that schools must work with the technology available. School technology should support teaching and learning, and assessments should work within that infrastructure. We have designed our system with what we refer to as a small footprint.

For an in-depth discussion of the technology requirements for AIR's online testing system, please see Topic 3 Technology Requirements.

Topic 7.1 Interoperability

AIR complies with industry interoperability standards. We support the standard Question and Test Interoperability (QTI) specifications, and we are currently developing the capability to import and export Accessible Portable Item Protocol (APIP)-compatible items. APIP builds on the QTI specifications. In addition, AIR's TIDE, which is responsible for student enrollment and registration, supports a variety of file import formats and interfaces. State data systems can send AIR flat files that are delimited (e.g., CSV), fixed-width files, Windows SIF files, IMS files, or files using custom XML formats.

In February 2015, we achieved WCAG 2.0 AA certification. In addition, our systems are compatible with modern technologies such as HTML5, MathML, and CSS. These technologies allow our user interfaces to scale and be compatible with the latest devices available on the market.

AIR will deliver the items, complete with all necessary accessibility markup, in our IAT format, which is compatible with QTI/APIP 2.2 and validates against that schema. In any case that QTI is deficient, we have extended it using the extension points available in QTI, such as custom interactions and custom operators. Renderers for such items would need to render those extensions, and AIR will provide the necessary documentation and schemas to support this. Presentation of the items, including ensuring that the presentation meets web accessibility standards, depends on the renderer rather than the representation of the items. At the Department's request, AIR will offer a no-cost license to the Department for our renderer, which can render these accessible items for the duration of the contract. We will provide technical assistance or, at the Department's discretion, host the renderer for released items for the duration of the contract.

D1.3 Security and Protection of Data

Topic 8 Security and Forensics

AIR recognizes the gravity of our responsibility as stewards of student data and secure test items, and we take every available precaution to protect these data. These data are protected from intrusion through our applications or intrusion into the host systems themselves.

A host of application-level security controls safeguard against unauthorized access to items or student data. Across all systems, our applications are password protected, and password complexity requirements are enforced, periodically audited, and periodically revised. Role-based permissions restrict authorized users' access to functionality and data. Our systems also restrict district and school user access by jurisdiction, allowing access only to those students over whom the user has legitimate educational jurisdiction.

All of our networks are protected by firewalls, automated intrusion detection systems, and extensive network monitoring. These safeguards protect against bad actors who attempt to bypass our applications, hack into our systems, and steal data. All of our systems, including servers, firewalls, and load balancers, are updated regularly with security patches. These systems have proven effective in protecting our clients' secure data.

We implement a two-part procedure to prevent access to students' personally identifiable information (PII) even in the unlikely event of a breach. Our strategy is to (1) limit exposure of the identifying information and (2) encrypt that data where it absolutely must appear. Most of our systems carry only an AIR-assigned, arbitrary student identifier for each student. Only a single system links that data to the student's statewide ID and name, and that system encrypts those fields. Systems needing this information must request the encrypted data. The key needed to decrypt these data is stored on a server without any external access and is only accessible to the servers running our systems. Our systems never write the key anywhere. Internally, only a very few engineers have access to this key.

In short, we stop bad actors from entering our system; if they get past this defense, even if they steal data, the data is likely devoid of PII. If they get lucky enough to get the data file with PII, they will find it encrypted, and they will not find the key anywhere.

One circumstance where PII is necessarily included and exposed is in the printing of student reports, which is managed securely.

All transmission of secure data takes place using secure file transfer protocol (SFTP). When the transferred data includes student PII, the transfer software encrypts the files upon upload to the site. The files remain encrypted while they are at rest, and Secure Sockets Layer (SSL) encrypts the data in transit. We recommend that our clients use similar SFTP software when the data arrive at their site to re-encrypt data when at rest on their servers.

All of our personnel undergo background checks and participate in our extensive security training, which is updated and revisited periodically.

Topic 8.1 Test Security

AIR shares the Department's concerns about ensuring security through all phases of developing, administering, and reporting these assessments. To be effective, security measures must be enforced and evaluated throughout the life cycle of the contract. Safeguarding restricted materials also helps to protect the integrity and quality of final products.

Security of Test Items

In addition to the online protection of test items, we have policies and procedures in place to prevent accidental disclosure of items. We have a policy against copying secure materials onto portable storage such as flash drives. Every staff member's laptop has an encrypted hard drive, so files stored in temporary folders or work in progress is not at risk if a portable computer is lost.

All item writing, reviewing, and test creation at AIR occurs in the Item Tracking System (ITS). This is a secure system, and only AIR staff working on the state project may access the state item bank. In addition, only grade-level leads and content leads will have access to all of the item development levels in the item bank, and the access for item writers, editors, and item reviewers is limited to the level at which they work. Once an item leaves the level that the reviewer/editor is certified to work in, that reviewer/editor can no longer interact with or view the item. In some limited instances, such as when reviewing item sets for cueing and overlap of distractors among items, the reviewers may print out the items to check more closely among the items at one time. AIR printers are located throughout the secure AIR facility, and AIR staff are instructed to be present at the printer when printing the items. AIR procedures dictate that all printed copies of items must be kept locked in a secure storage area in the staff member's office. Once a staff member has completed the review of the item set, the printed items are placed in a secure, locked shredding bin to be destroyed.

Security during Committee Meetings

AIR holds both online and paper-pencil committee meetings for states. For online committee meetings, items are accessed directly from AIR's secure ITS and therefore no printed copies need to be transported to meetings. For paper-pencil committees, item review booklets are distributed to participants only after they have signed the AIR and state non-disclosure forms. Prior to handing out the booklets, AIR staff review the security procedures with the committee members. Items in review booklets are printed on light green paper so that they are easily identified as being secure. Each page is also marked as secure in the header and/or footer. The grade-level lead at the committee meeting keeps a committee booklet log that indicates the names of the committee members and approved state staff members who were assigned a book. AIR then ships the review booklets in secure boxes with a tracking label to an agreed-upon location, and AIR staff keep them securely locked. AIR staff and the state's content staff review the committee members' comments and decide how to address each comment. The committee members' booklets are then shipped back to AIR where they can be securely destroyed or kept in secure AIR facilities, as determined by the Department.

Security of Items and Documents Related to Item Development

Once non-AIR staff item writers submit an item to AIR in ITS, they no longer have access to the item and cannot follow its review in our secure system. All of the editors reviewing the items are permanent AIR staff or temporary editors hired by AIR. These temporary editors are housed in AIR offices and work directly in ITS on computers owned by AIR. Non-AIR staff item writers must sign a non-disclosure form prior to beginning work on any AIR state project. In addition, AIR explicitly requests that all item writers submit a signed affidavit as proof that they have destroyed any hard copies of items or secure documents and deleted any copies of secure materials saved on the hard drives of their computers. AIR also requests that item writers send directly to AIR all secure materials they received from AIR while working on items. These affidavits and hard copy shipments are tracked internally at AIR offices and their receipt is documented.

Security of Items during Computer-Based Testing

AIR maintains the security of the test within the school through mechanisms that simultaneously exceed industry standards and minimize the workload on teachers. Using AIR's system, schools do not have to laboriously construct rosters of students who will test together. As long as a student is enrolled in the system, the testing rosters are constructed as students request entry into testing session. Students are prevented from leaving the secure test environment by the use of our secure browser, which also blocks prohibited software during testing. Secure material is never written to disk in any form during testing. All items are transmitted over HTTPS, which encrypts them in transit.

Protection of Personally Identifiable Information

Information security is a key requirement for any system dealing with high stakes assessments. AIR testing systems employ industry-standard processes and tools for data confidentiality, integrity, and availability. All sensitive data in our systems are encrypted using strong ciphers that meet FIPS-140 guidelines. All student personally identifiable information (PII) is encrypted both at rest and in transit. We employ a single sign-on for all of our systems using industry standard practices for user authentication (high-strength passwords, password change policies, etc.). All of our systems utilize role-based access controls to limit which users have access to what features and data in a system. We maintain audit logs to track all operations on a test opportunity or session. For data exchanges between systems, we use secure data transfer systems and/or digitally sign data using FIPS-140 compliant hashes for non-repudiation. All of our back end databases are backed up regularly and stored offsite. Access to our back end systems is limited only to users that need access such as AIR NOC personnel and Rackspace network management staff. These individuals are required to take mandatory training programs on AIR’s information security policies. We employ a third party penetration testing firm to independently audit our testing platform to identify any security vulnerabilities or risks.

All student PII is stored in 3DES encrypted form in our demographics databases. For each student imported through the Test Information Distribution Engine (TIDE), we generate an internal unique key for the student. This surrogate key is used within AIR systems to track students’ test progress and testing history. This ensures that student demographic data is never duplicated into any of our other system databases and that it remains centralized in our demographics servers. The keys used to encrypt/decrypt the demographic data are themselves secured and stored on a different set of dedicated servers and can be accessed only through a secure interface. Whenever any PII is needed by an application (for example, to present test scores with student names on the reporting interface to authorized users), it must fetch a passcode from our key servers through a secure protocol and then access the demographic servers to decrypt as needed the required pieces of data.

Physical Security of AIR Facilities

Every AIR location has card-controlled access. In our main office, card entry is not required for the building during business hours but is required for entry onto any of the floors where work is done. The entry is monitored by a guard at all times when the exterior doors are not locked. In each facility, visitors must sign in upon entry and must be escorted by an AIR staff member.

When in printed form, test items are protected by at least two locks when an AIR staff member is not present: locked doors require a card reader to enter the office blocks where work is done, and staff offices and workrooms are locked when secure material is present and the staff member is not.

Similar locks and controls prevent unauthorized access to servers on our network. Most of our data and test items are stored on AIR-dedicated servers at Rackspace, which has state-of-the- art security, including biometric access control to sensitive areas.

Physical Security of Rackspace Facilities

Below is a summary of physical security at Rackspace:

- Data center access limited to authorized Rackspace data center personnel
- Keycard protocols
- Thorough background security checks for every data center employee
- Biometric scanning for controlled data center access

- Security camera monitoring at all data center locations; around-the-clock interior and exterior surveillance
- 24/7 on-site staff provide additional protection against unauthorized entry
- Unmarked facilities help maintain low profile
- Physical security includes locking down and logging all physical access to servers at our data center
- Physical security audited by an independent firm

Topic 8.2 Data Forensics

The validity of test score interpretation depends critically on the integrity of the test administrations on which those scores are based. Any irregularities in the administration of assessments can therefore cast doubt on the validity of the inferences based on those test scores. Ensuring that tests are administered properly requires multiple facets, including clear test administration policies, effective test administrator training, and proven tools to identify possible irregularities in test administrations.

Monitoring for Testing Irregularities

Help Desk staff are currently trained to refer all calls about testing irregularities to the Department. AIR will work with New Hampshire to establish Help Desk procedures that best serve the Department. Procedures may include having any report of irregularities in test administration procedures or suspicion of cheating be immediately elevated to Tier 3 for project management notification, which requires project management to immediately notify the Department and provide details about the report.

AIR psychometric staff are also able to actively monitor test administrations for testing irregularities. Throughout the testing window, psychometricians are able to generate a series of quality assurance (QA) reports that allow them to monitor the quality of test administrations. The forensic analysis component of the QA reports, which we describe here, is geared toward detection of possible cheating, aggregating unusual responses at the student level to detect possible group-level testing anomalies.

Online test administration allows the Department to track information that is not possible to track in the context of the paper-pencil tests. This information includes not only item responses but also item response changes, latencies between item responses and changes, number of revisits to an item or items, test start and end times, scores in each opportunity in the current year, scores in the previous year, and other selected information in the system (e.g., accommodations). Our test delivery system (TDS) captures all of this information.

Unlike with paper-pencil assessments where data analysis must await the close of the testing window and processing of answer documents, AIR's TDS allows AIR psychometricians and state assessment staff to monitor testing anomalies throughout the testing window. Evidence evaluated includes changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. The analyses used to detect the testing anomalies can be run any time within the testing window. Analyses are performed at student-level and summarized for each aggregate unit, including testing session, test administrator, and school.

Changes in Student Performance

Score changes are examined between summative test administrations using a regression model. The scores between past and current years are evaluated, with the current-year score regressed on the test score from the previous year. Between-year comparisons can be performed starting with the second year of the test administration.

A large score gain or loss between grades is detected by examining the residuals for outliers. The residuals are computed as observed value minus predicted value. To detect unusual residuals, we compute the studentized *t* residuals. An unusual increase or decrease in student scores between opportunities is flagged when studentized *t* residuals are greater than |3|.

The number of students with a large score gain or loss is aggregated for a testing session, test administrator, and school. Unusual changes in an aggregate performance between opportunities and between years are flagged based on the average studentized *t* residuals in an aggregate unit (e.g., a testing session or a test administrator). For each aggregate unit, a critical *t* value is computed and flagged when *t* is greater than |3|,

$$t = \frac{Average\ residuals}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^{n} var(e_i)}{n^2}}},$$

where s = standard deviation of residuals in an aggregate unit; n = number of students in an aggregate unit (e.g., testing session or test administrator); and $var(e_i) = \sigma^2(1 - h_{ii})$. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

If the aggregate unit size is 1–5 students, the aggregate unit is flagged if the percentage of flagged students is greater than 50%. The aggregate unit size for the score change is based on the number of students included in the between-year regression analyses in the aggregate unit.

Item Response Latency

Item response latency is captured as the item page time (the time each item page is presented) in milliseconds. Discrete items appear on the screen one at a time. However, for stimulus-based items selected as part of an item group, all items associated with the stimulus are selected and loaded as a group. Total test time is the sum of the item page times.

Large deviations from average response times may indicate aberrant testing conditions. We might expect, for example, that response times will be much faster for students with prior knowledge of test items. Conversely, if a test administrator kept sessions active to review and change student responses, the testing times for such sessions would be much longer than expected.

The average and the standard deviation of test-taking time are computed across all students. Students and aggregate units are flagged if the test-taking time is greater than |3| standard deviations of the state average. The state average and standard deviation is computed based on all students at the time the analysis is performed. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

Inconsistent Item Response Pattern (Person Fit)

In Item Response Theory (IRT) models, person-fit measurement is used to identify test-takers whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test-taker copies responses, is provided answers by the test administrator, or has some responses changed prior to closing the test administration, the student will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student.

The person-fit index is computed using a standardized log-likelihood statistic, l_z, for systematic flagging of aberrant response patterns (Sotaridona, Pornell, and Vallejo, 2003). Students with l_z values greater than |3| are flagged. Aggregate units are flagged with *t* greater than |3|.

$$t = \frac{Average\ l_z \text{ values}}{\sqrt{(s^2)/n}},$$

where s = standard deviation of l_z values in an aggregate unit and n = number of students in an aggregate unit. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

Response Change and Response Similarity

Although AIR provides response change analysis and response similarity reports for fixed-form assessment systems, we have not as yet implemented these reports into the online QA system for several reasons. First, adaptive testing by its nature serves as a deterrent to student copying. Because students are being administered thousands of distinct test forms, the ability of students to copy answers from one another is reduced significantly. Moreover, the administration of unique test forms in the adaptive environment precludes even the possibility of investigating patterns of response similarity.

Should the Department desire, AIR will additionally provide a response change report that is modeled after the reports produced for fixed-form test administrations. Although the TDS records all responses the student makes for each item, passing all that information through the Quality Monitor (QM) for processing and subsequent storage in the Database of Record (DoR) would require developing infrastructure to store additional data that are not used for test scoring or reporting. To conduct the response change analysis, we therefore retrieve from the TDS the two most recent responses for each item, including the final scored response and the immediately preceding response, if any.

The response change report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit. Our experience has shown that values used to flag paper-pencil and online test administrations will be quite different since the frequency of response changes in online test administrations is substantially higher than for paper.

To provide a more in-depth examination of flagged test sessions, complete response records for test sessions with suspected testing irregularities can be pulled for further review.

Topic 8.3 Test Monitoring

Preemptive Practices to Ensure Fidelity of Test Administration and Security Procedures

While test administration and security procedures should be defined by the Department, AIR recognizes that implementing processes to maintain the security of assessment instruments and the confidentiality of student personally identifiable information (PII) are essential to producing valid, reliable, and fair student assessment results.

In developing the Test Coordinator Manual (TCM) and Test Administration Manual (TAM) with the Department, AIR commits to ensuring that all test security procedures are available to everyone involved in test administration. Additionally, each manual will include protocols for reporting any deviations in test administration procedures or the distribution and collection of secure materials as a test security incident. We will work with the Department to define breaches of security and established protocols for test administrators and school test coordinators to report these incidents to the District Test Coordinator and similar instructions for the District Test Coordinator to report suspected breaches of test security to the Department. Additionally, test security will be an agenda topic to be addressed by the Department during each of the four half-day training sessions.

We include a sample explanation of test security requirements included in the TCM that AIR produced with the Department for the 2016–2017 administration of the Smarter Balanced assessments in D1.3-1. AIR will also work with the Department to produce a Test Security/Confidentiality Agreement to be signed by each test administrator prior to administering the New Hampshire Statewide Assessments to acknowledge understanding of test security procedures if it is the preference of the Department.

Exhibit D1.3-1: Security Requirements of the Test Environment

REQUIREMENTS BEFORE TESTING	
Instructional materials removed or covered	Instructional materials **must be removed or covered, including but not limited to** information that might assist students in answering questions. This includes materials that may be displayed on bulletin boards, chalkboards or dry-erase boards, or on charts (e.g., wall charts that contain literary definitions, maps, mathematics formulas, etc.).
Student seating	Students must be seated so that there is enough space between them to minimize opportunities to look at each other's work, or they should be provided with table-top partitions.
Signage	If helpful, place a "TESTING—DO NOT DISTURB" sign on the door or post signs in halls and entrances rerouting hallway traffic in order to promote optimum testing conditions.
REQUIREMENTS DURING TESTING	
Quiet environment	Provide a quiet environment void of talking or other distractions that might interfere with a student's ability to concentrate or might compromise the testing situation.
Student supervision of electronic devices	Students are actively monitored and are **prohibited** from access to unauthorized electronic devices that allow availability to outside information, communication among students, or photographing or copying test content. This includes any device with cellular, messaging, or wireless capabilities, including but not limited to cell phones, iPods, cameras, and electronic translation devices.
Student access to allowable resources only	Students must only have access to and use of those allowable resources identified by Smarter Balanced (see Section IX.I Establishing Appropriate Testing Conditions for examples) that are permitted for each specific test (or portion of a test).
Access to assessments	Only students who are testing can view test items. Students who are not being tested or unauthorized staff or other adults must not be in the room where a test is being administered. Trained TEs/TAs may have limited exposure to test items in the course of properly administering the assessments; however, DCs, SCs, TEs, TAs,* and other trained staff may not actively review or analyze any test items.
No answer key development	No form or type of answer key may be developed for test items.
Testing through secure browser	Administration of the Smarter Balanced assessments is permitted only through the Student Interface via the secure browser.
No unauthorized logging in to the test delivery system	Only students may log in to the Student Interface. Students may not use another student's login credentials. Only adults may log in to the Test Administrator (TA) Interface with an authorized account. An adult may not use another person's username and password.
REQUIREMENTS DURING AND AFTER TESTING	
No access to responses	DCs, SCs, TEs, TAs,* and other staff are not permitted to review student responses in the testing interface or students' notes on scratch paper.
No copies of test materials	Unless needed as a print-on-demand or Braille accommodation, no copies of the test items, stimuli, reading passages, performance task materials, or writing prompts may be made or otherwise retained.
No access to digital, electronic, or manual recording devices	No digital, electronic, or manual device may be used to record or retain test items, reading passages, or writing prompts. Similarly, these materials must not be discussed with or released to anyone via any media, including fax, e-mail, social media websites, etc.

No retaining, discussing, or releasing test materials	Descriptions of test items, stimuli, printed reading passages, or writing prompts must not be retained, discussed, or released to anyone.
No reviewing, discussing, or analyzing test materials	DCs, SCs, TEs, TAs,* and other staff may not review, discuss, or analyze test items, stimuli, reading passages, or writing prompts at any time, including before, during, or after testing.
All test materials must remain secure at all times	Printed materials from the print-on-demand accommodation, scratch paper, and documents with student information must be kept in a securely locked room or locked cabinet that can be opened only by staff responsible for test administration.
No test materials used for instructional purpose	Test items, stimuli, reading passages, or writing prompts must not be used for instructional purposes.
Destroy test materials securely	Printed test items/passages, including embossed Braille print-outs, and scratch paper must be collected and inventoried at the end of each test session and then immediately destroyed. See Section I.IV Secure Handling of Printed Materials for details.

Monitoring the Fidelity of Test Administration and Security Procedures

Should a District Test Coordinator contact AIR via our Help Desk to report any issue relating to test administration and security procedures, AIR understands and agrees to redirect the District Test Coordinator to the Department as described in Topic 38 Help Desk Support and Topic 39 Support Center. Help Desk agents will be provided with an appropriate, approved response to the customer and notify AIR New Hampshire project team staff. The AIR project team will also contact the New Hampshire project team to ensure that the Department is aware of the situation. Similar procedures will apply should any test administrator or school test coordinator contact the AIR Help Desk.

References

Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Philippine Statistician, 52*(1–4), 81–92.

Topic 9 System Security

FERPA Regulation Compliance

All of AIR’s secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA), the Children’s Online Privacy Protection Act (COPPA), the Federal Information Security Management Act (FISMA), and other federal laws. Our systems implement sophisticated, configurable privacy rules that can limit data access to only appropriately authorized personnel. Different states interpret the FERPA differently, and we support customized interpretations. Our system is designed to support these interpretations flexibly. AIR is committed to working with the Department to maintain data security according to its specifications.

All secure data transmitted across the public Internet are encrypted using secure shell (SSH), advanced encryption standard (AES), or an Internet protocol security (IPsec) virtual private network (VPN). Secure websites encrypt data using 128-bit Secure Sockets Layer (SSL) public key encryption. When data is stored, it resides securely on database servers behind multiple firewalls and is secured through an encrypted connection. Secure transmission and password-protected access are basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins.

Key personally identifiable information (PII), such as student names and identifiers, is stored only in our Roster Tracking System (RTS) database. These data are encrypted, with encryption keys very closely held. Other systems store only an arbitrary AIR-assigned ID. When student data must be matched with identifiers (e.g., when presenting individual results in the Online Reporting System (ORS), the encrypted identifiers are merged onto the data and decrypted in real time.

Our systems use role-based security models that ensure that users access only the data to which they are entitled and that limit their ability to change those data according to their rights. User rights have two dimensions: the user's role and the user's data access rights. The role determines what actions a user can take, which types of reports the user can view, and similar functional limitations. Data access rights tell, for example, which school test coordinator can view which teacher and student data. Data access rights are governed by relationships among entities in our RTS, along with a configurable set of business rules. The business rules describe which access rights correspond to which relationships and enforce client-specific policies. We will work with the Department to maintain current business rules and adjust business rules as needed.

Security of test items and student information is protected at all times, with security procedures operating at three levels:

1. Physical security, preventing access to the machines on which data reside or are processed
2. Network security, including protection of our networks from infiltration and the secure transmission of data over our networks and others
3. Software security, ensuring that only authorized users access information on our systems and that their access is limited to information they are authorized to view.

System Assurance Provisions

Processes to Ensure Data Integrity

Topic 10 Backup and Recovery outlines the many mechanisms at work to ensure the integrity of the student data. The system uses record-level locking when that is sufficient to prevent deadlock and race conditions, but at times broader segments of tables must be locked to ensure data integrity. Restores from backup are only done manually, because automated recovery risks data corruption. Our system has a sophisticated set of warnings to alert users to unusual conditions, and the system is backed up nightly, with full backups of key testing data.

We have disaster recovery plans in place, which can recover data from most failures. Recovery from a catastrophic failure of an entire data center will not occur immediately. It will be possible to recover the data to another data center, although this will require manual processes that take more time. Recovery from a failure of any component will occur automatically.

AIR warrants the functionality of our software. Flaws will be reported immediately and remedied in the shortest feasible time frame. We use up-to-date standards for all development and adhere to strict security guidelines.

Reliance on System Assurance Capabilities of the Relational Database Management System (RDMS)

AIR's system is based on a private cloud design that allows every function to be distributed across many database and application servers, while maintaining responsiveness. Each of AIR's private clouds supports hundreds of thousands of simultaneous users. In our secure browser, Every time a student answers a question, the response is saved to our servers. A typical student waits less than half a second after finishing one item before the next item is fully displayed on his or her screen. For longer responses, such as essays, the system can be configured to save periodically (say, every minute) and when the student presses the Save button. If the system is unable to reach the server, the student is stopped from testing, so work is not lost. We prevent data loss by saving responses in real time to AIR servers.

The safe, asynchronous system reports the successful save back to the browser. If the browser does not receive the response after a configurable amount of time (usually 30–90 seconds), the system stops the student from testing until connectivity is restored.

The system is designed to make multiple attempts to reach the server. Therefore, even if connectivity is temporarily lost, no work is lost. Often, the connection will be reestablished without the student being aware that it was ever lost.

Data integrity is paramount. Our system is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. Our architecture, described in Topic 10 Backup and Recovery, is designed to recover from failure of any component with little interruption.

Synchronization among Databases

Active/passive clusters of database servers are configured so that the passive node takes over in the event of failure of the active node. For more information about our system architecture and backup and recovery procedures, we refer readers to Topic 10 Backup and Recovery and Appendix B, which includes our Business Continuity Plan.

Security Testing

User Identification and Authentication

Students

AIR recognizes that the login process should be simple for students. We implement a novel and effective student authentication system. Students do not have to use any special login to access their tests. They can use something easy to remember like their first name and unique student identifier (e.g., state student identifier [SSID]). With this information, students can request access to the session that the test administrator (TA) has created in the TA Interface. Students have a two-part authentication where once they log in, a TA has to approve their entry into a test session from his or her workstation or mobile device after verifying the student's identity. This provides more security than a simple password system, and simplifies the administrative tasks around administration while still keeping the login process simple for students.

Users

Our systems use role-based security models that ensure that users access only the data to which they are entitled and that limit their ability to change those data according to their rights. User rights have two dimensions: the user's role and the user's data access rights. The role determines what actions a user can take, which types of reports the user can view, and similar functional limitations. Data access rights tell, for example, which school test coordinator can view which teacher and student data. Data access rights are governed by relationships among entities in our RTS, along with a configurable set of business rules. The business rules describe which access rights correspond to which relationships and enforce client-specific policies. We will work with the Department to maintain current business rules and adjust business rules as needed.

Authorized Access to Data and Services

User Access

All of our secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the FERPA, COPPA, FISMA, and other federal laws.

Secure transmission and password-protected access are basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins.

Our systems use role-based security models that ensure that users access only the data to which they are entitled and that limit their ability to change those data according to their rights. User rights have two dimensions: the user's role and the user's data access rights. The role determines what actions a user can take, which types of reports the user can view, and similar functional limitations. Data access rights tell, for example, which school test coordinator can view which teacher and student data. Data access rights are governed by relationships among entities in our RTS, along with a configurable set of business rules. The business rules describe which access rights correspond to which relationships and enforce client-specific policies. We will work with the Department to maintain current business rules and adjust business rules as needed.

Student Access

Students take secure tests through the Student Interface, which is essentially a secure website accessed through AIR's secure browser. Test-takers are only permitted to access those tests that TAs have granted them permission to access. TAs create a session simply by selecting the tests that will be available in the session using the TA Interface, which TAs use to assign tests and proctor students.

To access a test, a student logs in using his or her name and other identifying information, such as ID number and test session ID. This information is displayed on both the student's and TA's workstations. A student's test opportunity is associated with the internal AIR student ID and test ID when it is started.

AIR recognizes that test content must be kept in a secure test environment. Thus, AIR has included the following security features on the Student Interface:

Item visibility. After the items have been administered to the student, item visibility is the ability for these items to remain visible to a student only for a configurable period of time. In general, the recommended rule is 20 minutes.

Idle timeout. This feature allows the Student Interface to log the student out if the student is idle for a configurable period of time.

Automatic pausing. The test is automatically paused once the TA closes the session in the TA Interface.

AIR's system does not write any of the content (whether visual or audio, such as text-to-speech) to the file system. All content is stored in memory. Thus, closing the secure browser automatically removes all content.

Immunity from Unauthorized Malicious Programs

Secure Browser

AIR's secure browser is device-specific, which ensures that features unique to that device are kept secure in the testing environment. The secure browser operates in a full-screen mode, disables access to other applications, and prohibits navigation outside the test. The browser is designed to intercept all operating system hot-key combinations and print capabilities; it enables keyboard combinations specifically designed for test navigation.

Forbidden applications. This feature automatically detects certain applications that are prohibited from running on a computer while the secure browser is open. The secure browser checks the applications currently running on a computer when it is launched. If a forbidden application is detected, the student is denied entry and receives a message indicating the open application. Similarly, if a forbidden application launches while the student is already in a test (e.g., scheduled tasks), the student is logged out and a message is displayed.

TA Interface

There are many security features built into the TA Interface:

Navigating away from the TA Interface. The TA needs to focus only on managing the session; if he or she navigates away from the interface, this may pose a security issue, and doing so logs out and closes the session, which will log the student out of the test.

User interface timeout. The timeout indicates how long (in minutes) a TA and students can be idle before the TA is automatically logged out of the test session.

Check-in time. This feature indicates how long (in minutes) a TA's test session can be idle when a test is begun before the system automatically closes the test session.

Pausing/closing sessions closes all tests. When a session is paused (either explicitly by the TA or implicitly by logging out of the TA Interface), all tests belonging to that session are automatically paused.

Network Security

Hardware firewalls protect our networks from intrusion. They are installed and configured to prevent access for services other than hypertext transfer protocol secure (HTTPS) services for our secure sites. Our firewalls provide a first level of defense against intrusion and are backed up by a capable second line: hardware and software intrusion detection and remediation. Our intrusion detection systems constantly monitor network traffic and raise alerts for suspicious or unusual network traffic.

Our systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts. Suspicious log entries are investigated and resolved.

All secure data transmitted across the public Internet are encrypted using SSH, AES, or an IPsec VPN. Secure websites encrypt data using 128-bit SSL public key encryption.

The hosting environment is protected by an Alert Logic Threat Manager Intrusion Prevention System (IPS) appliance at the perimeter and by Symantec AntiVirus Corporate Edition on each individual server. The Alert Logic IPS appliance combines intrusion protection and vulnerability management technology into a single integrated solution that offers both proactive and reactive protection from the latest threats. Symantec AntiVirus offers real-time virus and malware protection for our servers, along with centralized management and administration capabilities.

The methods used to ensure that communications and data integrity are not intentionally corrupted via unauthorized creation, modification or deletion.

AIR fully recognizes the importance of protecting user and student information in all its various states, whether it is being transmitted, stored, or converted. AIR has built-in security controls in all of its data stores and transmissions. Unique user identification is a requirement for all systems and interfaces. All of AIR's systems conduct external data transfers on a secure file transfer protocol (SFTP) connection, and all communication with any of the online websites occurs over SSL. This ensures that passwords and secure student data are not passed over the network in clear text.

The methods used to ensure that the parties to interactions with the application cannot later repudiate or rebut those interactions.

AIR's system does not write any of the content (whether visual or audio, such as text-to-speech) to the file system. All content is stored in memory. Thus, closing the secure browser automatically removes all content.

The intrusion detection methods used to ensure the detection, recording and review of attempted access or modification by unauthorized individuals.

Hardware firewalls protect our networks from intrusion. They are installed and configured to prevent access for services other than hypertext transfer protocol secure (HTTPS) services for our secure sites. Our firewalls provide a first level of defense against intrusion and are backed up by a capable second line: hardware and software intrusion detection and remediation. Our intrusion detection systems constantly monitor network traffic and raise alerts for suspicious or unusual network traffic.

Our systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts. Suspicious log entries are investigated and resolved.

All secure data transmitted across the public Internet are encrypted using SSH, AES, or an IPsec VPN. Secure websites encrypt data using 128-bit SSL public key encryption.

The hosting environment is protected by an Alert Logic Threat Manager Intrusion Prevention System (IPS) appliance at the perimeter and by Symantec AntiVirus Corporate Edition on each individual server. The Alert Logic IPS appliance combines intrusion protection and vulnerability management technology into a single integrated solution that offers both proactive and reactive protection from the latest threats. Symantec AntiVirus offers real-time virus and malware protection for our servers, along with centralized management and administration capabilities.

The privacy methods used to ensure that confidential data and sensitive communications are kept private.

As mentioned above, AIR considers the privacy of student data of utmost importance and we take every available precaution to protect these data. We are committed to working with the Department to maintain data security according to its specifications.

Student data are protected from exposure in our applications and from intrusion into the host systems themselves.

All of our secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality. Our systems implement sophisticated, configurable privacy rules that can limit data access to only appropriately authorized personnel. Secure transmission and password-protected access are basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins.

AIR's systems also use role-based security models, which ensure that users access only the data to which they are entitled and that also limit their ability to change those data according to their role.

To prevent access to students' PII, even in the unlikely event of a breach, we implement a two-part procedure to prevent access. Our strategy is to limit exposure of the identifying information and to encrypt that data where it absolutely must appear. Most of our systems carry only an AIR-assigned, arbitrary student identifier for each student. Only a single system links that identifier to the student's statewide identifier and name. That system encrypts those fields. Systems needing this information request the encrypted data. The key needed to decrypt these data is stored on a server without any external access and is accessible only to the servers running our systems, and our systems never write the key anywhere. Internally, only a very few engineers have access to this key.

AIR fully recognizes the importance of protecting user and student information in all its various states, whether it is being transmitted, stored, or converted. AIR has built-in security controls in all of its data stores and transmissions. Unique user identification is a requirement for all systems and interfaces. All of AIR's systems conduct external data transfers on a SFTP connection, and all communication with any of the online websites occurs over SSL. This ensures that passwords and secure student data are not passed over the network in clear text.

The system maintenance methods used to ensure that unauthorized system maintenance does not unintentionally disrupt the security mechanisms of the application or supporting hardware.

Planned maintenance downtime to AIR systems will be coordinated with the Department, and mutually acceptable windows for system updates and improvements will be negotiated. All regular, daily maintenance will be performed at night and over weekends; other downtime may be necessary and will be negotiated.

Our systems will be available during mutually agreed-upon windows. Some operations require downtime that can be scheduled. Only routine daily maintenance will be conducted nightly.

AIR is proud to note that our system operates reliably, without interruption. Our systems incorporate redundancies and fail-safe measures at every potential point of failure, and we maintain redundant data stores in the case of a catastrophic event. In the unlikely event of interruption, data remain secure, and the system is designed to retain test validity even as students stop testing and return. As our recommendations from other states attest, we can deliver tests with virtually no systemic interruptions.

Scheduled unavailability, or maintenance with associated downtime, is scheduled and communicated proactively. Emergency downtime may be necessary in the event of component failure.

The testing methods conducted to load and stress test your system to determine its ability to withstand Denial of Service (DoS) attacks.

All of AIR's production systems are hosted at Rackspace and are protected against Distributed Denial of Service (DDoS) attacks through the automated deployment of Arbor DDoS mitigation software. Arbor provides industry leading DDoS protection and sits in front of all AIR assessment application systems as an automated *always on* protection service.

DDoS attacks generally fall into one of three categories:

1. Volumetric DDoS Attacks

   Attempt to consume the bandwidth either within the target network/service, or between the target network/service and the rest of the Internet.

2. TCP State-Exhaustion DDoS Attacks

   Attempts to consume the connection state tables which are present in many infrastructure components such as load-balancers, firewalls and the application servers themselves.

3. Application Layer DDoS Attacks

   This is the most deadly kind of DDoS attack. It can be very effective with as few as one attacking machine generating a low traffic rate (this makes these attacks very difficult to proactively detect and mitigate).

During the spring of 2015 we turned the Arbor software from manual to automated and we have successfully mitigated numerous DDoS attack vectors against our systems. The actual attacks have come in the following vectors signatures: TCP RST, ICMP, IP Fragmentation, TCP SYN, Total Traffic, UDP, NTP Amplification, DNS Amplification, and SSDP Amplification. Several of these attacks were large in scale yet, Arbor successfully mitigated the attacks without any impact to users in the field, including student testing.

These actual attack incidents give us confidence that our DDoS mitigation is capable to withstand DDoS attacks. We also realize that bad actors will continue to attempt attacks hence, we continue our diligence to work with Rackspace to identify processes and software options to deflect DDoS traffic.

Your software patch schedule employed to protect the software from new security vulnerabilities as they arise.

On a monthly basis, we have a rigorous patching policy that we engage in on all our servers to ensure that we have the latest security patches. See Appendix R for our Patching Policy.

The ability of your system's software to be installed in a "locked-down" fashion so as to turn off unnecessary features (user accounts, operating system services, etc.) thereby reducing the software's security vulnerabilities and attack surfaces available to system hackers and attackers.

AIR's secure browser is device-specific, which ensures that features unique to that device are kept secure in the testing environment. The secure browser operates in a full-screen mode, disables access to other applications, and prohibits navigation outside the test. The browser is designed to intercept all operating system hot-key combinations and print capabilities; it enables keyboard combinations specifically designed for test navigation.

Forbidden applications. This feature automatically detects certain applications that are prohibited from running on a computer while the secure browser is open. The secure browser checks the applications currently running on a computer when it is launched. If a forbidden application is detected, the student is denied entry and receives a message indicating the open application. Similarly, if a forbidden application launches while the student is already in a test (e.g., scheduled tasks), the student is logged out and a message is displayed.

CONFIDENTIAL

Topic 10 Backup and Recovery

Prevention and Recovery of Disruptions in Test Delivery System

AIR is continuously improving our ability to protect our systems from interruptions. We intend to exceed 99% uptime.

Data integrity is mission-critical. AIR's test delivery system (TDS) is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. Our architecture, described below, is designed to recover from failure of any component with little interruption. Each system is redundant, and critical student response data is transferred to a different data center each night.

AIR has developed a unique monitoring system that is very sensitive to changes in server performance. As with most monitoring systems, our system provides warnings when something is going wrong, but it also provides warnings when any given server is performing differently from its performance over the prior few hours, or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them *before* a failure. For example, earlier this year, the system alerted us to a drive that had some sectors going bad. On multiple occasions, this has enabled us to make adjustments and replace equipment before any problems occurred.

AIR has also implemented an escalation procedure that enables us to alert clients within minutes of any disruption. Our emergency alert system notifies our executive and technical staff by text message, who immediately join a call to understand the problem. Our vice president of assessment programs and client services, Ms. Hayes, is among the participants, and she is charged with ensuring that clients are promptly notified if there is (or is likely to be) impact on the field.

Below, we present our system architecture, and describe how it recovers from device failures, Internet interruptions, and other problems.

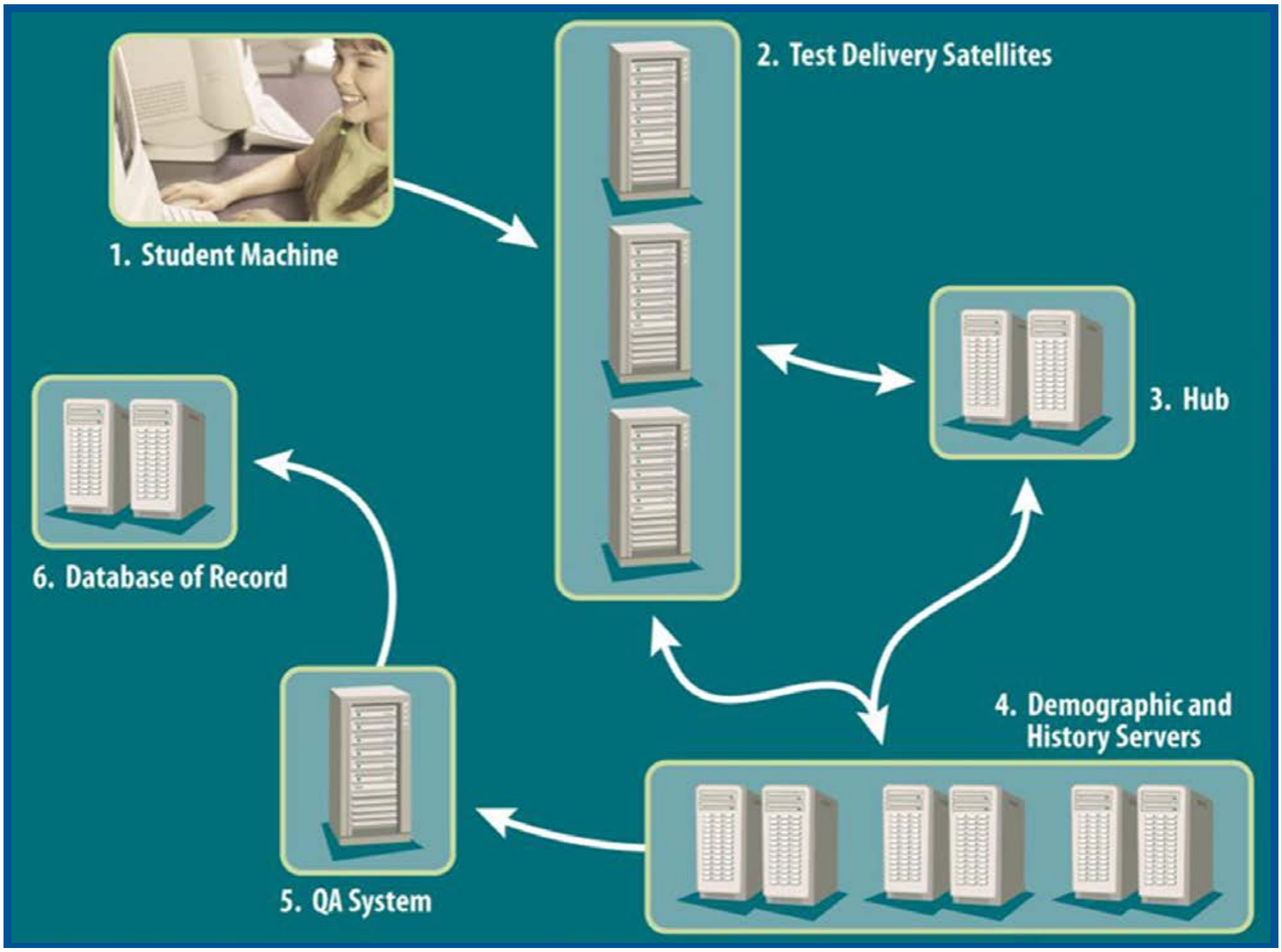
High-Level System Architecture

Our architecture provides redundancy, robustness, and reliability required by a large-scale, high-stakes testing program. Our general approach, which has been adopted by Smarter Balanced as standard policy, is pragmatic and well-supported by our architecture.

Any system built around an expectation of flawless performance of computers or networks within schools and districts is bound to fail. We know that school or district computers and networks will occasionally experience technology problems. Our system is designed to ensure that the testing results and experience are able to respond robustly to such inevitable failures. Thus, AIR's TDS is designed to protect data integrity and prevent student data loss at every point in the process.

wwork at each point in the system. Fault tolerance and automated recovery are built into every component of the system, as described below.

Exhibit D1.3-2: Key Elements of the Testing System



1. The Student Machine

Student responses are conveyed to our servers in real time as students respond. Long responses, such as essays, are saved automatically at configurable intervals (usually set to one minute), so that student work is not at risk during testing.

Responses are saved asynchronously, with a background process on the student machine waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated time (usually set to 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning at a later time. For example:

- If connectivity is lost and restored within the designated time period, the student may be unaware of the momentary interruption.
- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying the save.
- If the system fails completely, upon logging back in the system, the student returns to the item at which the failure occurred.

In short, data integrity is preserved by confirmed saves to our servers and prevention of further testing if confirmation is not received.

2. The Test Delivery Satellites

The test delivery satellites communicate with the student machines to deliver items and receive responses. Each satellite is a collection of web and database servers. Each satellite is equipped with redundant array of independent disks (RAID) systems to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

One server for every four satellites serves as a backup hub. This server continually monitors and stores all changed student response data from the satellites, creating an additional copy of the real-time data. In the unlikely event of failure, data are completely protected. Satellites are automatically monitored, and upon failure, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described below), with backup copies remaining on the drive arrays of the disabled satellite.

If a satellite fails, students will exit the system. The automatic recovery system enables them to log in again within seconds or minutes of the failure, without data loss. This process is managed by the hub. Data will remain on the satellites until the satellite receives notice from the demographic and history servers that the data are safely stored on those disks.

3. The Hub

Hub servers are redundant clusters of database servers with RAID drive systems. Hub servers continuously gather data from the test delivery satellites and their mini-hubs and store that data as described above. This real-time backup copy remains on the hub until the hub receives notification from the demographic and history servers that the data have reached the designated storage location.

4. Demographic and History Servers

The demographic and history servers store student data for the duration of the testing window. They are clustered database servers, also with RAID subsystems, providing redundant capability to prevent data loss in the event of server or disk failure. At the normal conclusion of a test, these servers receive completed tests from the test delivery satellites. Upon successful completion of the storage of the information, these servers notify the hub and satellites that it is safe to delete student data.

5. Quality Assurance System

The quality assurance (QA) system gathers data used to detect cheating, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QA system, and any anomalies (such as unscored or missing items, unexpected test lengths, or other unlikely issues) are flagged, with an immediate notification to our psychometricians and project team.

6. Database of Record

The Database of Record (DoR) is the final storage location for the student data. These clustered database servers with RAID systems hold the completed student data.

Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered real-time backup processes prevent loss of student data, even in the unlikely event of system failure.

CONFIDENTIAL

Topic 11 Assurance of Business Continuity

AIR’s business continuity plan addresses our approach to disaster recovery. We propose a plan that ensures that the system is protected against normal contingencies.

AIR’s test delivery system and related systems are designed to be extremely fault-tolerant. The system can withstand failure of any component with little or no interruption of service. One way that we achieve this robustness is through redundancy. Key redundant systems are as follows:

- Rackspace, our hosting provider, has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, they can operate indefinitely. They maintain an n+1 configuration of 16 diesel generators that, at maximum capacity, can supply up to 2.0 megawatts each.
- Our hosting provider has multiple redundancies in the flow of information to and from our data centers by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- Every installation is served by multiple web servers, any one of which can take over for an individual test upon failure of another.
- Active/passive clusters of database servers are configured so that the passive node takes over in the event of failure of the active node. Each database server in a cluster has dual connections to the disk arrays containing the system data.
- Each disk array is internally redundant, with multiple disks containing each data element. Failure of any individual disk recovers immediately by accessing the redundant data on another disk.

Data are further protected by nightly backups. We complete a full weekly backup and incremental backups nightly. AIR has two sets of backups:

- CommVault: We run a full backup each weekend with daily differential backups the rest of the week.
- SQL Maintenance Plans: We run SQL maintenance plans nightly, run full backups on each of the Shard servers, and then copy each backup to an archive server at the other data center.

Our full Business Continuity Plan is provided in Appendix B.

D1.4 Training and Support

AIR assumes the responsibility for providing training and support required to ensure successful administration of the New Hampshire Statewide Assessments. We propose to provide a help desk, detailed in Topic 38 Help Desk Support, for annual quality support to districts and schools throughout the registration, testing, and reporting cycles.

AIR also assumes responsibility for producing the Test Coordinator Manual (TCM) and Test Administration Manual (TAM) in order to ensure the successful administration of the New Hampshire Statewide Assessments. The production processes for these manuals are described in Topic 12.1 and Topic 12.2, respectively. AIR will provide all manuals in Adobe PDF format and allow access via the Internet for users to view and download. These materials will adhere to an ADA-compliant format. AIR proposes to provide these ADA-compliant manuals for access via the administrative portal, which is discussed in further detail in Topic 4 Assessment Delivery Platform.

Topic 12 User Manuals and Guides

AIR will work with the Department to develop and provide manuals, documents, and tools to assist students in accessing test items and to support test coordinators and test administrators in delivering the statewide assessments. Online versions are integral to the testing platform and, the goal of each user guide is to motivate users to make the best possible use of the system and the data it provides. All user guides provide "how-to" information to enable users to access and navigate the systems they are authorized to use. Each user guide begins with an overview of the system and includes screenshots and descriptions on how to accomplish tasks or move through the system.

For example, the *Online Reporting System (ORS) User Guide* walks users through the ORS. The guide begins by describing how the data in the system can support users as they make instructional decisions. Then, it walks users through scenarios demonstrating appropriate uses of the data. In introducing each set of reporting capabilities, the system will describe why those reports are available, how they can help at the school or in the classroom, and what the limitations of the data are. The guide also shows users how to create classes and other rosters of students and how to use these rosters to create aggregate statistics or track groups of students over time. A similar level of detail is provided in the *Test Information Distribution Engine (TIDE) User Guide*. AIR will work closely with the Department to update and revise the guides for each school year and ensure all systems documentation reflects updates to each system.

In addition to the TCM and TAM, such ancillary materials may include but are not limited to the following:

- **Systems User Guides.** User guides that provide instructions and support for users accessing any AIR online system. These user guides may include but are not limited to the *Test Information Distribution Engine (TIDE) User Guide*, *Online Reporting System (ORS) User Guide*, and *AIRWays User Guide*.
- **Practice and Training Test User Guide**. Provides an overview of the Practice and Training Test Site and information about accessing the Practice and Training Tests.
- **Braille Requirements and Testing Manual**. This manual will include information about supported operating systems and required hardware and software for Braille testing. It provides information on how to configure JAWS, navigating an online test with JAWS, and how to administer a test to a student requiring Braille.
- **Technical Specifications for Online Testing Manual.** This manual provides technology staff with the technical specifications for online testing, including information on Internet and network requirements, general hardware and software requirements, and the text-to-speech function.

- **System Requirements for Online Testing,** This document outlines the basic technology requirements for administering an online assessment, including operating system requirements and supported web browsers.
- **Secure Browser Installation Manual.** This manual provides instructions for downloading and installing the secure browser on supported operating systems used for online assessments.

AIR will ensure that all ancillary materials are made of high-quality materials and meet the Department's specifications. Samples of such ancillary materials will be provided to the Department for review and approval before final order/production of the materials.

AIR will develop ancillary materials that are

- fully customizable to include policies and procedures unique to the New Hampshire Statewide Assessments;
- compliant with the Department's test administration material workflow process;
- scheduled to ensure that there is minimal overlap with other test administration material under development;
- designed in consultation with layout and graphic design experts to create a look and feel reflective of the New Hampshire Statewide Assessment program; and
- reviewed and produced by accessibility experts to ensure that they are fully ADA compliant.

While developers at AIR have considerable experience creating administration scripts and instructions in support of AIR's test delivery system (TDS) for various states, we recognize that each state is different. Each state has its own test administration requirements that must be communicated to districts and schools. AIR will work closely with the Department to ensure inclusion of policies and procedures familiar to districts, schools, teachers, and especially students.

AIR understands that the Department has limited staff available to review materials and suggests the following document creation and review process for the first written draft and all subsequent versions of all materials and products developed for this project:

- AIR will discuss all upcoming review dates with the Department during the regular weekly meeting. We will send reminders to reviewers in advance of sending documents for review. The schedule for sending reminders will be determined in consultation with the Department.
- Each review period will last a minimum of five work days and will begin with the posting of an electronic version of the document to a secure website by 9:00 a.m. (Eastern Standard Time) on the first day of review.
- Development and review of administration materials will be staggered to minimize the number of materials under review by the Department at any one time.
- AIR will propose for the Department's approval a detailed materials development schedule for managing the workflow, volume, and format of documents to be reviewed in each five-day period.
- Before delivery, AIR will proofread all documents following well-defined procedures to ensure that all drafts are proofread for accuracy by qualified proofreaders.
- AIR will provide all user manuals and guides in both PDF and Word format for posting to the Department and vendor websites and for the use of the Department in creating training and informational materials.

Topic 12.1 Test Coordinator Manual

AIR, in collaboration with New Hampshire, will develop and electronically distribute TCMs that clearly explain all procedures relative to the organization of testing at the district and building levels. The Department will have authority to approve all language, content, and format of these TCMs. The TCMs will include

- appropriate processes for ordering and returning any paper testing materials;
- training requirements for test administrators;
- appropriate processes for handling accommodations;
- appropriate processes for protecting test security at the district level;
- suggested or required times for test sections and suggestions for district test scheduling;
- appropriate processes for including special populations of students in testing;
- important dates leading up to, during, and after the testing window(s); and
- how to handle student absences and other unique testing situations (e.g., testing of homebound students, and testing of students moving into and/or out of the school district during the testing window).

We divide the development process for the TCM, as well as the TAM, into two phases:

1. A drafting phase, during which we obtain agreement on the content and design of the document.
   a. **State Manuscript Round/Test Coordinator Manual (TCM) or Test Administration Manual (TAM) Information Sheet.** At this stage, AIR and the Department identify portions of the prior administrations' TCMs/TAMs to be used as a starting point for the current manual; identify high-level changes, additions, and omissions to the scripts and other sections to be included; and introduce known or potential policy updates.
   b. **AIR Round 1.** AIR uses the markup of the manuals provided by the Department along with the TCM/TAM information sheet to draft the manual. AIR updates the TCM/TAM with known changes, including new screen captures, and embeds queries to the Department for its consideration and input. This document is considered a draft given the outstanding questions and placeholder text/images prevalent in this round.
2. A production and review phase, during which we produce the document to the specifications developed in the first phase, accommodate any final adjustments, coordinate client and internal reviews, and ensure that the camera-ready proofs are error-free.

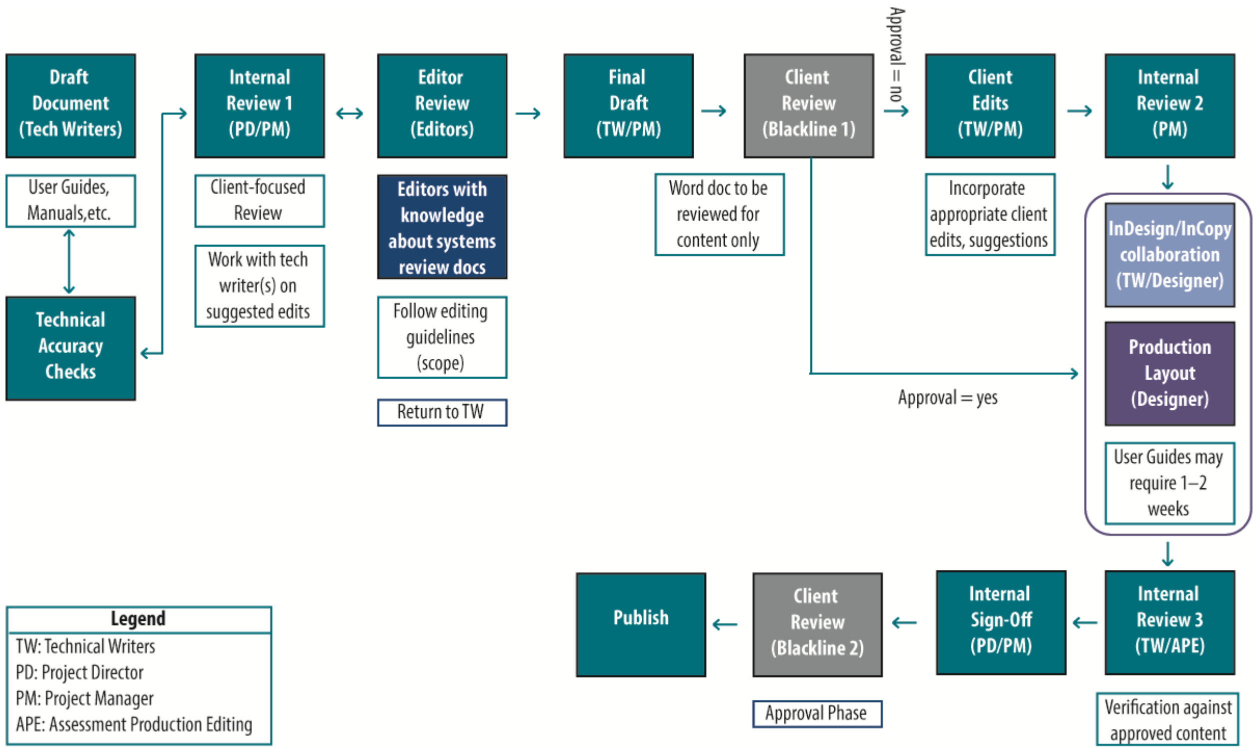
The drafting phase will begin with conversations with the Department, during which we will discuss issues of both design and content. Following these discussions and an exchange of materials, such as the Department manuscript, the TCM/TAM information sheet, and a style guide, AIR will begin to develop the draft. AIR will develop drafts in Word format and will produce the final draft of each manual in both Word and PDF format for posting to the Department and vendor websites and for the use of the Department in creating training and informational materials.

We view the drafting process as iterative. Especially with new or significantly altered materials, we understand that it is not possible to make all final decisions before a first draft. Manuals and other ancillary documents are produced over several weeks, and annual policy changes or clarifications may not be finalized before the beginning of the production process. Additionally, new reviewers enter the process as documents approach completion. Interactions among design or content elements can sometimes be unexpected, and sometimes a seemingly good idea at an early stage simply does not work. Therefore, we build in time for multiple drafts of each product before it enters the production process. During the drafting process, the Department will receive formatted documents in either Word or Adobe PDF format for review depending on the preference of the Department. These drafts, however, will not reflect the final formatting. The first blackline version of the document (Round 1), which results from the first iteration through our production process, is designed to reflect the final layout and formatting. This, too, is an iterative process, and we will work collaboratively with New Hampshire to ensure an error-free product.

Each document's content is fully drafted before it enters the production system. The Department must be assured of error-free production of materials. This is an area where AIR has had considerable success, though we strive for continuous improvement. Our production department is headed by Kara Prezocki, who brings more than a decade of experience in publications management. The staff of the production department consists of senior editors, editors, proofreaders, graphic designers, and layout specialists. Every public assessment document leaving AIR travels through this department for thorough review.

Our production processes include multiple editorial reviews, and each document must receive senior staff sign-off. When the senior staff member overseeing the content is satisfied with the content, the document enters the production system. Its first stop is an editorial review. Exhibit D1.4-1 provides an overview of our editorial review process.

Exhibit D1.4-1: Overview of AIR's Editorial Review and Revision Process



After the initial drafting phase, AIR will provide a copy of each TCM in both Word and Adobe PDF format no less than 60 days before the beginning of each test administration window for the Department to review. No less than 30 days before each administration window, the final TCM will be available electronically in Adobe PDF format so training can be provided to school and district test coordinators. The final TAM will be posted on New Hampshire's administrative portal, described in further detail in Topic 4 Assessment Delivery Platform.

Topic 12.2 Test Administration Manual

AIR, in collaboration with the Department, will develop and electronically distribute TAMs that clearly explain all procedures relative to test administration for the New Hampshire Statewide Assessments. The

TAMs will include information on administering both accommodated and non-accommodated tests. The Department will have authority to approve all language, content, and formatting of these TAMs.

Each TAM will include

- specific instructions for the administration of the applicable assessment(s);
- timing information (as appropriate);
- scripts or teacher test directions for the administration of each assessment to ensure that consistent and appropriate instructions are given to students; and
- any other information test administrators need to be able to give the assessment.

The TAMs will follow the same drafting procedure as described for the TCM in Topic 12.1 Test Coordinator Manual. After the initial drafting phase, AIR will provide a copy of each TAM in both Word and Adobe PDF format no less than 60 days before the beginning of each test administration window for the Department to review. No less than 30 days before each administration window, the final TAM will be available electronically in Adobe PDF format so training can be provided to test administrators on delivering the test and interpreting results of the assessments. The final TAM will be posted on New Hampshire's administrative portal, described in further detail in Topic 4 Assessment Delivery Platform.

Topic 13 Training Materials

Proposed Training Plan

Below we propose a training plan for district and school test assessment and technology coordinators as well as test administrators on all aspects of New Hampshire's Statewide Assessment program. The Department, together with AIR, will determine all training topics, the audience, frequency, and mode of delivery for the training sessions. Existing AIR systems training modules will provide the foundation for training along with other training materials as required by the Department. Our proposed plan includes four half-day regional trainings plus a series of webinars, presentations, and self-paced tutorials on each of our online systems to support the regional trainings and annual assessment administrations. The web-based trainings are available to users at any time on the administrative portal (discussed in Topic 4 Assessment Delivery Platform) and will serve as stand-alone trainings for each year of the contract. AIR's training materials are designed to teach both sophisticated technology users and users new to the systems their roles and responsibilities in context of the new online systems. While sophisticated users may not need instruction on all steps of a specific function and can progress to the point in the training they find useful, new system users typically find it useful to have instruction at a very detailed level. A detailed description of each training presentation listed in Exhibit D1.4-2 is also discussed on the following pages.

Exhibit D1.4-2: Proposed Presentations and Formats for Annual Trainings

Training Phase I: Preparing for Online Testing		
Presentation Title and Primary Audience	**Proposed Training Topics**	**Proposed Training Formats**
User Roles — Teachers, Test Administrators, School Test Coordinators, District Test Coordinators, District/ School Technology Coordinators	User roles: Who does what and in which system?	1. User role chart 2. Included as a component of each webinar and onsite trainings (also summarized in each user guide and manual produced with the Department)
Technology Requirements for Online Testing — District/School	Steps for secure browser installation and minimum hardware requirements.	1. Webinar presentation 2. Instructions on administrative portal

Training Phase I: Preparing for Online Testing		
Presentation Title and Primary Audience	**Proposed Training Topics**	**Proposed Training Formats**
Technology Coordinators		

Exhibit D1.4-2: Proposed Presentations and Formats for Annual Trainings (continued)

Training Phase I: Preparing for Online Testing		
Presentation Title and Primary Audience	**Proposed Training Topics**	**Proposed Training Formats**
TIDE: Registering Users and Modifying Student Settings — Teachers, Test Administrators, School Test Coordinators, District Test Coordinators	Learn how to register users in the Test Information Distribution Engine (TIDE), update student test settings and restrictions.	1. Webinar presentation 2. On-site training 3. Online training module
Training Phase II: Administering Online Tests		
Online Test Administrator Certification Course: — Test Administrators	Learn how to use the new online testing system, set up a test session, manage and monitor testing, assist students with online tools, and understand accessibility and accommodations. Certification earned only through passing mandatory quiz.	1. Online Test Administration Certification course
Online Practice Tests — Test Administrators, Teachers	Practice setting up, pausing, resuming, and ending live test sessions and setting accommodations using the same functionality as the operational test administrator site.	1. Online practice tests
Online Practice Tests — Students, Parents, Teachers	Practice signing in to a live test session using the same functionality as the operational student testing site, including using text-to-speech, zoom, highlighter, and strike-through, and answering all item types.	1. Online practice tests
How to Start and Monitor Online Testing and Modify Test Settings — Test Administrators, Teachers	Learn how to set up a test session, modify test settings, and monitor participation in multi-opportunity testing. Additionally, learn how to reset and invalidate tests.	1. Webinar presentation (recorded) 2. On-site training 3. Online training module
Training Phase III: Post-Testing		
How to Use the AIRWays Reporting System — Teachers, School Test Coordinators, District Test Coordinators How to Use the Online Reporting System (ORS) — Teachers, School Test Coordinators, District Test Coordinators	Learn how to access student scores, create class rosters, and drill down from district and school results.	1. Webinar presentation (recorded) 2. On-site training 3. Online training module

A detailed description of proposed training presentations and materials in Exhibit D1.4-2 is also provided below:

1. **User Role Chart.** This reference document indicates, at a glance, the access each user role is permitted in each online system. System access for each user role will be reviewed in each webinar, as appropriate. The user role chart will be posted on the administrative portal and will be available in the manuals and form part of the online, self-guided tutorials as determined in collaboration with the Department.
2. **Technology Requirements for Online Testing.** This training includes information for school and district network administrators, provides guidelines for varied technical setups, and technical support available at the schools. Technical setups include multiple secure browser installation methods such as manual installation on individual machines, installation on machines through a network, access from a shared network drive, and thin-client setup. Minimum hardware requirements will also be reviewed in detail. The recorded webinar will be posted to the administrative portal for ease of access.
3. **TIDE: Registering Users and Modifying Student Settings.** This training includes instructions on how to add users at the district and school level so they can access appropriate online systems, as well as how to upload students for test eligibility. The training will also cover other TIDE functionality including updating student test settings and using the test management reports. This content will be delivered in a webinar and module format as well as during the on-site trainings. The recorded webinar and training module will be posted to the administrative portal for ease of access.
4. **Online Test Administrator Certification Course.** This self-paced course includes detailed instruction on creating and managing test sessions, monitoring student activity, setting accommodations, and implementing test security measures. The course is designed to familiarize test administrators with the Student Interface, the approval process, and the tools students will use during testing. To ensure course-takers have learned the skills necessary to proctor tests using the new online system, each user must complete the course, including a quiz, and sign a test security agreement form before being permitted to sign in to the online testing system. Users who need to refresh their memory of course content can retake the course as many times as needed. For on-site trainings and certification, the test security agreement (compliance form) will be made available on paper for test administrators to sign. The online Test Administrator Certification course will be made available on the administrative portal.
5. **Online Practice Tests.** Online practice tests for all subject- and grade-level assessments will be available prior to the opening of the test window. Each practice test will be composed of approximately 20 items that are a subset of the blueprint for the operational test, thus providing users with exposure to all item types. The online practice test site uses the same applications as the operational test site, such as the Test Administrator Interface, the Student Interface, and the test management reports. This design ensures that students, educators, and teachers become familiar with the online testing system before operational testing begins. In addition, user guides and systems documentation will be made available when the online practice tests open, to help educators prepare their materials prior to testing. The online practice tests will be available throughout the test window. These online practice tests will be available on the administrative portal and are discussed in further detail in Topic 14 Practice Test and Student Materials.
6. **How to Start and Monitor Testing and Modify Test Settings.** This webinar training for test administrators includes the same content as the online Test Administrator Certification course but will also be delivered in a webinar and during the on-site trainings.
7. **How to Use the Online Reporting System and the AIRWays Reporting System.** This webinar training teaches authorized users how to access and correctly interpret score reports in both the Online Reporting System (ORS; the reporting system for summative assessments) and AIRWays (the reporting system for interim assessments). Users learn how to view district, institution, personnel, roster, and individual student reports in table

and graph format; access and use longitudinal reports for their diagnostic need; aggregate or disaggregate scores, filter score reports by student subgroup (e.g., gender, ethnicity, English language learner); and manage online rosters (groups). Users will also be trained on how to accurately interpret the information from the summative assessments provided in the ORS and from the interim assessments in AIRWays and how to draw correct inferences from the data so that educators can use the information provided in these systems to inform classroom instruction. Educators will also be trained to explain the reports and scores to parents and guardians. AIR will conduct webinar trainings on the ORS and will include this material in the on-site trainings.

Each year, AIR will collaborate closely with the Department to update the set of webinar presentations and online tutorials as needed. In the first year, we propose to focus all presentations on the basics of the program and the features of the online systems. In subsequent years, AIR proposes to focus training on the updates to each system and providing basic refresher training on each system. On a mutually agreed-upon date with the Department, a training plan and schedule will be provided. A proposed timeline training can be found in Exhibit D1.4-3.

Exhibit D1.4-3: Proposed Training Timeline

Training Type	Proposed Timeline
1. Four half-day regional trainings	To be determined with the Department, suggested each fall in preparation for summative assessments
2. User Role Reference Chart 3. Webinar for District Administrators 4. Webinar for Technology Coordinators	6–8 weeks before interim testing window opens
5. Online Test Administrator Certification Course	6–8 weeks before interim testing window opens
6. Online Practice Tests	6–8 weeks before interim testing window opens
7. Webinar for Test Administrators	6–8 weeks before interim testing window opens
8. Webinar on Online Reporting System and AIRWays	6–8 weeks before interim testing window opens

Training for Interim Assessments

Given that the same online testing system we propose for the summative assessments will be used for the interim assessments, systems training for the interim assessments will be the same as that outlined above. With the exception of the AIRWays reporting system, the same student registration system (TIDE), TDS, ORS, and minimum hardware requirements used to deliver the summative assessments will also be used to deliver the interim assessments. Therefore, once test administrators have been trained on the summative online systems, they will be prepared to use the same online systems to access the interim assessments. We propose to conduct additional training on the AIRWays system during the reporting webinars to help ensure that educators understand how to accurately interpret the results from the interim assessments in order to inform classroom instruction. We will also provide a separate user guide for the AIRWays reporting system using the processes indicated below. Similarly, all systems documentation, user guides, webinars, and training documentation described in this section will be used to train schools and districts on the interim assessment system. In addition, guides such as the *System Requirements for Online Testing* and *Technical Specifications for Online Testing Manual*, which detail the minimum hardware requirements to access the summative and interim assessments, will also be provided. All user guides and system manuals will be posted to the administrative portal for ease of access.

Training Documentation: User Guides and System Documentation

As discussed in Topic 12 User Manuals and Guides, AIR recognizes that easy-to-read system documentation is essential for successful training. Over the past five years, we have produced concise, effective, system-specific user guides explaining to novice and sophisticated users alike how to complete critical tasks in the TIDE, TDS, and ORS. All user guides are available in electronic format and PDF format as required. AIR will produce ADA-compliant documents as specified in the RFP. We will work closely with the Department to produce a set of user guides that include step-by-step instructions on how to use each system and how to address common issues encountered in each system. We commit to providing:

- system-specific user manuals and documentation written in clear, accessible language that include visual aids to each system in the form of screenshots describing key functions and the steps required to resolve common technical problems;
- a New Hampshire-specific accessibility and accommodations guide;
- recorded webinars for online trainings;
- a set of FAQs for educators and administrators that are housed in the administrative portal; and
- training modules for the field, as needed, to explain how to complete a specific process using a particular online system.

AIR will review and update all system documentation annually or as specified by the Department to ensure that information in the user guides is correct and up-to-date. The process used to produce accurate, concise, and accessible system documentation is discussed in Topic 12 User Manuals and Guides.

Topic 13.1 Teacher Directions

AIR, in collaboration with the Department will develop and electronically distribute TAMs that clearly explain all procedures relative to test administration for the New Hampshire Statewide Assessments, including teacher directions. For more information on teacher test directions and the contents of the TAM, please refer to Topic 12.2 Test Administration Manual.

Topic 14 Practice Tests and Student Materials

Each year, AIR will provide a practice test. AIR's TDS can deliver practice tests in two modes: *Guest Mode* for practice tests and *Secure Mode* for training tests. In Guest Mode, anyone can log on to the system anonymously and take any test offered in the system.

Practice Tests

Practice tests will be administered in Guest Mode. Users can experience all item types as presented in the Student Interface, including Performance Tasks. Because the tests are administered anonymously, there is no way to resume a paused test. At each sitting, users start a test anew. Anonymity also implies that users may take as many tests as they like, since the system cannot track how many have been taken. Many states enjoy the benefits of a freely, publicly available view of the assessment, which provides transparency to parents and other stakeholders. The practice environment allows students to configure their own settings. In Guest Mode, the user can select from among the available accommodations and supports at the beginning of the test.

Training Tests

Training tests will be administered with the same test delivery system used to administer operational tests, providing a full dress rehearsal of the operational test administration. Training tests will be delivered in Secure Mode. The training test site uses the same applications as the operational test site, such as the Test

Administrator Interface, the Student Interface, and the test management reports. This design ensures that students, educators, and teachers become familiar with the online testing system before operational testing begins.

The system delivering the training tests will be the same system delivering the summative assessment, so the same set of tools, accommodations, and embedded supports are supported. When delivered in Secure Mode, the system will support the same mechanisms for assigning test settings as the summative assessment. In addition, user guides and systems documentation will be made available when the online training tests open, to help educators prepare their materials prior to testing. The training tests will be available throughout the test window.

Other Student Materials

In addition to the practice and training tests, students will have access to item type tutorials, demo calculators, and various training modules. These materials will be made available on the administrative portal as well as any additional materials deemed necessary by the Department to ensure that students are prepared to function within the online testing environment.

Topic 15 Software Implementation Training

As described in Topic 12 User Manuals and Guides and Topic 13 Training Materials, much of the technology training will be covered in the trainings and easy-to-read user guides proposed. AIR's purely Internet-based system has the smallest possible technology footprint in the industry and low-end user maintenance, making it easy to set up and use. Our trainings support technology users on the minimal setup requirements they will need for test preparation and implementation.

In addition to user guides that will outline system and technology requirements and configurations, AIR can work together with the Department on developing a series of technology webinars to complement the trainings outlined in Topic 13 Training Materials. These webinars will go over requirements and specifications, along with tools, such as the Diagnostic Tool, to help users prepare for testing. AIR will work with the Department to develop any necessary training materials not described in Topic 12 User Manuals and Guides and Topic 13 Training Materials, but we think the Department will find our documentation clear and comprehensive. Any additional trainings may take place early in the school year to ensure technology coordinators are well-prepared.

Topic 15.1 Training and Professional Development

See Exhibit D1.4-3 for trainings offered and tentative timelines for their implementations. Upon award of the contract, we will determine exact dates for each training opportunity with the Department.

Topic 15.2 Technology Director Training

As described above, AIR will work with the Department to develop and carry out a webinar for technology directors to ensure that they are prepared for the assessments. In addition, the documentation we will provide, such as the *Secure Browser Installation Manual* and the *System Requirements for Online Testing*, will help familiarize technology directors with our systems.

Topic 15.3 Assessment Administration Training

AIR's assessment training plan is described in Topic 13 Training Materials. The webinars and modules will show district test coordinators and test administrators how to enroll students for testing both interim and summative tests (TIDE: Registering Users and Modifying Student Settings), and how to administer the tests (How to Start and Monitor Online Testing and Modify Test Settings). Each webinar and module will be developed in collaboration with the Department through the use of storyboards and/or PowerPoint presentations. AIR will also provide a TA Certification Course (online) to users to further familiarize them with the test administration process using our systems.

Topic 15.4 Assessment Results Training

AIRs user guides are developed with educators in mind. As described above, AIR will provide comprehensive user guides to educators that will show them how to explore the data made available after the assessments are taken. As described in Topic 13 Training Materials, AIR will develop both AIRWays and ORS webinars that will describe the reports and data made available to educators to help them make informed instructional decisions.

Topic 15.5 Online Training Support

In addition to the user guides and trainings outlined in Topic 13 Training Materials, educators will have comprehensive support from the AIR help desk. Our responsive and knowledgeable help desk is available to all users, and in fact, because of our test-monitoring processes alert system, we have contacted schools to help troubleshoot before they were even aware that they were having a problem. More information on our help desk can be found in Topic 38 Help Desk Support and Topic 39 Support Center.

D1.5 Assessment Scoring, Analysis, and Equating

Topic 16 Machine-Scored Items

AIR has developed a large suite of machine-scored item types that provide fast scoring across a range of selected- and constructed-response item types. Most machine-scored item types are scored using explicit, test developer-defined rubrics that are contained in the item, which ensures fidelity to scoring rules across all test administrations. The rule-based rubrics allow test developers to construct test items that measure reason-based problem-solving, and validity of test score interpretation is increased because the rubrics explicitly define how students must respond to achieve each score point rather than providing guidelines for the kinds of responses associated with each score point.

AIR has six scoring engines that use explicit rubrics:

1. Multiple-choice scoring engine, which takes an option identifier as the key
2. Hot-text scoring engine, in which students select or rearrange sentences or phrases in a passage
3. Graphic-response scoring engine, which has an explicit test developer-created rubric that describes the properties of correct responses and relates scores assigned to those properties
4. Equation scoring engine, which evaluates the characteristics of student-entered equation responses against an explicit test developer-created rubric
5. Proposition-response scoring engine, which uses a pattern-matching algorithm to recognize test developer-created formal propositions in text using varying words, grammar, etc. (The explicit rubric defines concepts and relations among them and the relationship between the presence or absence of propositions and scores assigned.)
6. Simulation-interaction scoring engine, which evaluates a sequences of trials in a simulation item against an explicit test developer-supplied rubric

We propose to use our Automated Essay Scoring (AES) engine, which uses statistical algorithms to score student essays for the writing assessment in real time. The statistical rubrics used to develop the scoring models measure a broad set of features, some of which may be item specific and "learned" from a training set. During training, these features are related to human scores through a statistical model. The resulting equation predicts how a human would score a response with the measured features.

Autoscore is currently being used to accurately and quickly score student essay responses in Arizona, Ohio, and Utah. Statistical scoring models have already been developed to score most AIRCore writing assessments, so automated scoring of essay responses for New Hampshire's statewide assessments can begin immediately. In addition, we note that there are additional AIRCore writing assessments for which scoring models have not yet been developed. Since reporting of assessment results for spring 2018 must necessarily await the outcome of standard setting workshops, the Department may also choose to administer writing tasks for which models have yet to be developed. Following the procedures described below, a sample of essay responses for training and cross validation will be identified, and for which optimally valid human scores will be obtained. From the responses, statistical scoring models will be developed allowing for immediate scoring and reporting of essay responses for those items in spring 2019. In the following paragraphs we describe how the AES scores student essays, and how we maintain the validity of all writing scores.

AES uses a statistical algorithm to score essay responses. The statistical scoring models also yield an indicator of score confidence based on (1) responses with unusual features and (2) responses scoring near rubric thresholds. For each model, a confidence threshold can be identified, and any scored response with a confidence value below the threshold will be automatically routed for verification scoring by a trained reader.

Scoring rubrics for all machine-scored items are highly secure, because they never leave our host servers. Although item scoring is nearly instantaneous, a modular, asynchronous scoring framework allows the adaptive algorithm to proceed with item selection even while awaiting completion of scoring of submitted item responses, both machine scored and handscored. This ensures that students never need to wait for items to be scored to proceed through the test, even if demands on the network slow the scoring of some machine-scored items or item responses require handscoring following test administration.

AIR proposes 100% machine scoring of constructed-response items for all New Hampshire assessments. In conjunction with pre-equated item parameter estimates used for both item selection in the adaptive algorithm and immediate reporting of summative and interim assessments in the ORS and additional interim assessment reporting in AIRWays, machine scoring allows for immediate reporting of all test results. Upon submission of the test, the student’s test record is passed to the Quality Monitor (QM), which performs a series of checks on the test record, including item and test scores. Records passing successfully through the QM (which are virtually all of them) are deposited to the database of record (DoR), where they are available to the Online Reporting System (ORS) and, for interim assessments, AIRWays.

As noted above, when assigned scores are near rubric thresholds, or when responses are dissimilar from those encountered in the training set, the confidence value assigned to the score becomes low. For accountability assessments, we propose to flag very low confidence scores automatically and route those responses for verification scoring by human raters. We propose to flag approximately 15% of all responses for verification reads.

As we describe in Section D1.6 Reporting, and Section D1.8 Reporting Portal, the ORS dynamically creates score reports from the contents of the DoR and the roster tracking system (RTS). As soon as an individual student’s test record is passed to the DoR, his or her assessment results will be available to the reporting portals (the ORS and AIRWays). Aggregate-level reports will also be available, with class, school, and district reports being updated as tests are completed and become available for reporting.

Topic 16.1 Automated Scoring of Student-Generated Responses

Creating and Refining Machine Rubrics for Scoring Engines with Explicit Rubrics

Rubrics for each explicit rubric scoring engine are first developed by item writers and reviewers. We note that our rubrics support true rule-based reasoning, which is much more flexible and powerful than the simple token-matching approaches that are more common. True rule-based items allow more flexible scorings and admit rubrics that are more tightly aligned to the constructs being measured. These types of rubrics can be used for true constructed-response items. As with any constructed-response items, examinees may generate responses not anticipated by the test developers. Therefore, each item goes through a process similar to rangefinding for human-scored items.

AIR has developed a process called *rubric validation* that efficiently reviews scoring rubrics for true rule-based scoring. This process is supported by our REVISE software.

AIR’s proven method for rubric validation is second-to-none in the industry. After items are field tested, AIR content specialists work with the Analysis Team to prepare for the rubric validation meetings. All student responses for non-selected response items are loaded into AIR’s REVISE system, and AIR content specialists use the REVISE system to generate student samples based on a stratified random sample: one-third of the samples generated reflect students who performed well on the overall assessment but poorly on the item in question, one-third reflect students who were less proficient overall on the assessment but performed relatively well on the item in question, and one-third reflect students who performed as expected on the item in question. This process allows AIR to identify any potential scoring concerns prior to the rubric validation meeting, such as unanticipated (but accurate) responses, equivalent responses that were not originally considered, and responses that are getting credit but should not (based on the content and the item rubric).

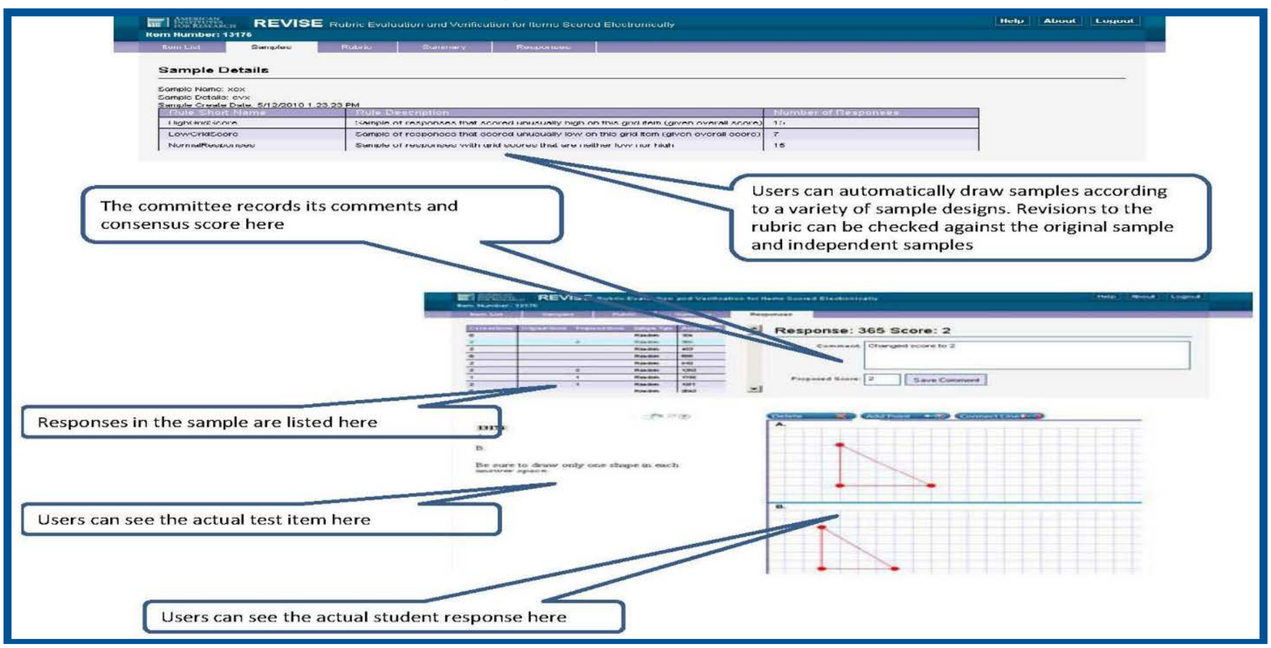
During the rubric validation meeting, AIR content specialists lead a training for participants to explain the rubric validation process, the importance of teacher participation and review of student responses, and how to review student responses in light of the current rubric. Participants are grouped by grade and content area, and AIR content specialists facilitate in each meeting room. Participants read the item/question, review the current rubric, and review the stratified random sample of student responses. If participants note responses that should be receiving credit but are not currently, they can recommend to Department that the rubrics be updated to include these additional correct responses. Similarly, if there are responses that participants feel should be excluded, they can be disallowed in the updated scoring. Finally, if a multi-point item is not performing well as a multi-point item, AIR may recommend to Department that the item be collapsed to a one-point item. (For example, if 48% of students previously earned 0 points, 48% of students earned 2 points, and only 4% of students earned 1 point, it could be argued that the item should truly be a 1-point item.)

Department content specialists will approve all changes to rubrics before the item is finally approved or rejected. AIR content specialists who are well-versed in item scoring and rubrics will apply the approved edits and generate all student responses whose scores have changed as a result of these rubric edits. These responses are reviewed to ensure that new responses deemed correct match the new rubric, and student responses that are no longer deemed correct are not receiving credit.

After the rubric validation meetings, any items rejected as part of the process (and approved by the Department) are rejected within ITS and are not moved forward. All items that are accepted at rubric validation are rescored with the updated rubrics by AIR's analysis team, and then AIR psychometric staff review the item statistics of the items (based on the updated rubrics and scoring) prior to data review.

Exhibit D1.5-1 highlights some of the features of the REVISE software.

Exhibit D1.5-1: Features of the Revise System



For the equation and proposition items, we bolster the rubric validation process with a validation study that compares the performance of the machine-scoring rubric with handscoring. To execute the validation study, a random sample of 500 cases from each item are handscored, and discrepancies from the machine score are reviewed and resolved. We will report a validity rate for each item.

Automated Essay Scoring Training

The engine employs a training set, a set of essays scored with optimally valid scores, which we obtain by having all responses double- scored by expert scorers and using an adjudication process for discrepant scores. The quality of the human-assigned scores is critical to the identification of a valid model and final performance of the scoring engine. The essay scoring engine is a statistical scoring engine. Measurable features of the essays are extracted and statistically analyzed. The measurable features include both syntactic features (e.g., misspellings, sentence length, grammar mistakes) and semantic features.

The semantic features reflect a summary of the words used in the documents. The summary is formed by first reducing words to their roots (e.g., “eating” becomes “eat”) and creating a matrix of words (or n-grams, which are sequences of words) by responses. This, of course, is a very large sparse matrix. The matrix is factor analyzed to reduce its dimensionality, and the factors become the semantic predictors.

During training, the first step separates a random sample of approximately 500 cases to reserve for independent validation of the final model.

The second step removes responses flagged for deterministic condition codes. The computer, of course, is quite good at identifying responses that are too short to score, in which the prompt is copied, or where the student simply copies the same text over and over. These cases are flagged and removed from the training. The remaining cases, which still include cases to which humans assigned other condition codes, such as “off topic,” are decomposed into the variables that comprise the predictors in the model. These predictors, or subsets of them, are included in ordered probit models predicting (1) whether a condition code would be applied, and (2) the score that would be applied. Each prediction comes with a confidence index. An algorithm takes this information into account in determining whether to assign a condition code or a score.

When our psychometricians are satisfied with the models, they are run against the validation sample to estimate the agreement between the scoring engine and the fully resolved human score.

The essay scoring system generates several confidence indices, which are then used to determine whether the paper should receive a score or a condition code. The mean and standard deviation of the confidence index are computed for each dimension, and the lowest value is selected for flagging responses with one or more low confidence scores. Any dimension score with a confidence index below this threshold is flagged for verification by a human rater.

We propose to route papers with the lowest confidence scores to human scorers. In the past this has amounted to approximately 20% of the papers with the expected turnaround being seven business days. This would mean that 80% of students get immediate scores in AIR’s reporting systems and the other 20% get their scores within seven business days.

Scoring of Paper Tests

Given the very small number of anticipated paper test administrations, we propose to score paper-administered tests using our data entry interface (DEI) in which test administrators enter student responses online for automated scoring and reporting. The DEI is fully integrated with AIR’s ORS. As a result, test administrators will receive preliminary scores for students participating in paper test administration immediately after submitting student responses. The DEI substantially reduces processing and scoring costs of paper test administrations.

Topic 17 Analysis and Psychometric Support

AIR has established, automated systems for designing and producing customized data files and score reports. Data processing begins with three key specifications documents:

- Analysis specifications
- Scoring and reporting specifications
- Return data layouts

These documents lay out, in reproducible detail, all calculations, algorithms, and business rules for processing the data, from item-level scoring through the generation of score reports. The analysis specifications provide the details for the classical and IRT item analyses, including, for example, handling of missing responses in the data, as well as procedures to reproduce all linking studies. The scoring specifications detail every rule for processing tests. These include scales to be calculated, subscales and reporting categories, transformations of these scales and subscales, attemptedness rules, participation rules, and every other calculation. The return data layouts appear in the form of a spreadsheet intended for both human and machine consumption. For humans, this document specifies the contract between our systems and yours, including what data will be exchanged, how the exchange is performed, and what format is used.

We strive to have all of our analysis programs configured and ready to go before the testing window opens. Configuration includes a substantial quality control (QC) process in which we generate data and run it through the system. The same data is processed by a team of psychometricians using statistical processing programs as an independent verification.

As we describe in Topic 40 on Technical Reporting, AIR will annually produce a technical report documenting all facets of New Hampshire's statewide assessment system. AIR is committed to working with the Department to provide all technical documentation necessary to successfully complete the peer review process. To support the Department and their TAC in the development and review of peer review evidence, AIR will provide the Department with a crosswalk between the peer review guidelines and the technical report sections.

Score Types

The RFP requires that the primary student scores will provide an overall achievement score and performance level, based on all operational items administered to the student. We propose to use the maximum likelihood method to compute scores based on the pattern of correct and incorrect student responses to the operational items. The likelihood function is defined by the item response theory model used for item calibration. Calibration models and the maximum likelihood method for obtaining student scores are discussed in detail in Topic 18 Calibration and Scaling.

Subscale performance based on subsets of items can provide actionable feedback to students, parents, and teachers, and they may provide information regarding a student's on relative strengths and weaknesses. Subscores will also be computed using the maximum likelihood method, based solely on the administered operational items contained in the particular reporting category. In general, reporting a subscore performance level at the student level requires about the equivalent of 10 dichotomous items for reasonable reliability. Therefore, we expect to be able to report scores at the content domain level for students but not at more fine-grained levels.

More fine-grained reporting categories can be used at aggregate levels (e.g., classroom, school, district) because the reliability of aggregated scores is higher than the reliability of individual scores. The administration of standards-based adaptive design for ELA and mathematics and of a matrix design in science is instrumental for the generalizability of the aggregated subscores. Each student may only see a few items pertaining to a reporting category, but through adaptive item selection for ELA and math, and by spiraling matrix blocks of items across students within a class in science, a much larger set of items pertaining to the reporting category is administered at each aggregate level. This enables the assessments to reflect the depth and breadth of the NH academic standards.

Prediction Model

The items comprising the Utah formative assessment bank, which is one of the available banks from which the New Hampshire interim assessments could be constructed, were originally part of Utah's SAGE summative item banks. The Utah formative banks are thus linked to the ICCR scale. In addition to block interim assessments reported in AIRWays, Utah offers comprehensive interim assessments based on a blueprint proportional to the summative assessments. Exhibit D1.5-2 shows mean performance gains for a sample of students taking the fall comprehensive interim assessment and the spring summative assessment, and it also shows the correlations between the test scores between occasions. As the correlations indicate, the interim subject area test scores based on a comprehensive blueprint are predictive of subsequent summative performance. We will also conduct analyses of the New Hampshire assessments to verify the predictive relationship, but note that if the interim assessments are administered in content homogenous blocks comprised of relatively few items, the magnitude of observed correlations will be reduced.

Exhibit D1.5-2: Stability and Growth between Fall 2015 Interim and Spring 2016 Summative Assessments: Full Reporting Interim Assessment

Assessment	N	Fall 2015 Full Reporting Interim		Spring 2016 Summative		Change From Interim to Summative		% Scoring Lower	Corr.
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev		
ELA									
ELA_G3	3607	285	62	329	61	44	43	15	0.75
ELA_G4	3389	320	70	357	69	36	42	20	0.81
ELA_G5	3747	362	71	395	66	33	40	21	0.83
ELA_G6	4677	397	72	426	69	30	40	22	0.83
ELA_G7	3515	416	70	435	70	20	39	30	0.84
ELA_G8	2897	420	80	447	78	26	44	27	0.84
ELA_G9	1234	425	92	449	88	24	48	30	0.86
ELA_G10	1224	452	88	474	94	22	52	32	0.84
ELA_G11	1166	463	96	479	93	16	56	38	0.82
Mathematics									
MATH_G3	4640	279	30	316	35	37	22	4	0.77
MATH_G4	4345	311	36	344	41	33	24	7	0.82
MATH_G5	4295	334	41	374	49	40	29	7	0.81
MATH_G6	4759	374	50	410	55	36	28	8	0.86
MATH_G7	3896	410	54	441	59	32	32	13	0.84
MATH_G8	3346	427	65	477	73	50	44	11	0.80
MATH_SM 1	3028	456	85	498	88	42	56	18	0.79
MATH_SM 2	2742	499	97	536	107	38	74	24	0.74
MATH_SM 3	1791	552	103	594	103	42	79	26	0.70

Exhibit D1.5-2: Stability and Growth between Fall 2015 Interim and Spring 2016 Summative Assessments: Full Reporting Interim Assessment (continued)

Assessment	N	Fall 2015 Full Reporting Interim		Spring 2016 Summative		Change From Interim to Summative		% Scoring Lower	Corr.
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev		
Science									
Science_G4	3471	810	24	835	30	26	20	10	0.75
Science_G5	3463	818	22	837	25	20	16	11	0.77
Science_G6	4524	809	32	844	39	34	24	6	0.79
Science_G7	4320	815	25	836	29	21	18	11	0.80
Science_G8	4134	821	24	837	28	16	17	16	0.80
Earth Science	2209	814	33	834	34	20	24	18	0.74
Biology	3781	817	27	833	30	16	18	17	0.80
Chemistry	1335	811	32	837	37	26	27	14	0.71
Physics	836	817	34	837	39	20	25	19	0.78

Topic 18 Calibration and Scaling

As described in Topic 1 Test Design, we propose to administer the New Hampshire statewide assessments in ELA and mathematics adaptively using AIR's ICCR item bank. The ICCR items have been developed in conjunction with state assessment programs, following all item review procedures prescribed by the participating state. The ICCR item bank has been used to administer state assessments in Arizona, Florida, Ohio, Tennessee and Utah. ICCR items have also been administered in the embedded field-test slots in the Oregon administration of the Smarter Balanced assessments.

A common-item linking design was implemented in spring 2015 to link ICCR item parameters to the scale of each participating state. In addition, most ICCR items are administered in more than one state, effectively providing a common-item linking design between the assessments of the states using the ICCR item bank.

For each state where ICCR items were administered (except Oregon), the ICCR items were calibrated concurrently with the state-specific items. A separate calibration was performed for each subject and grade. All items were calibrated under the three most prominent item response theory (IRT) models in use: the Rasch model, the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). Relying on the common ICCR items across states, all ICCR items for a given grade and subject were linked and expressed on a common scale regardless of where they were administered. In addition, by relying on the items common across grades in states that use a vertical scale, the ICCR items are also expressed on a vertical scale for each subject. Thus, the ICCR item bank is an extensive item bank with pre-equated item parameters for the three most common operationally used IRT models, for both within-grade and vertical scales.

Because item parameters estimated from each of the state assessments are linked to the common ICCR scale, the performance standards and other benchmarks established in each of the state assessments can be represented on the ICCR scale. These linkages allow New Hampshire to provide the locations of performance-standards and other benchmark locations to assist standard-setting panelists to identify performance standards corresponding to locations where multiple performance standards and benchmarks converge.

Topic 18.1 Calibration Plan

Calibration Models for ELA and Mathematics

As noted above, the ICCR items have been calibrated using Rasch, 2PL, and 3PL IRT models. AIR recommends using the more general 3PL/Generalized Partial Credit IRT model, which is the default model for the ICCR item bank. Traditional IRT models assume a single underlying trait, and they assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^{n} P(z|\theta),$$

where Z represents the pattern of item responses and θ represents a student's true proficiency.

For multiple-choice models, the three-parameter logistic (3PL) model takes the form

$$P(x_j = 1|\theta_{k,}a_j, b_j, c_j) = c_j + \frac{1-c_j}{1+e^{-1.7a_j(\theta_k - b_j)}} = P_{j1}(\theta_k).$$

The b parameter is called the *location* or *difficulty* parameter. The a parameter is referred to as the *slope* or *discrimination* parameter. The slope parameter is essentially the inverse of the standard deviation of the measurement error associated with the item. The third parameter, c, defines a lower asymptote. In the absence of the c parameter, the probability of a correct response approaches zero as proficiency decreases toward negative infinity. The c parameter allows the probability to approach some other lower bound. Given multiple-choice questions, a student with very little ability on the target trait could guess a correct answer. The c parameter captures the effect of such guessing.

For items that have multiple, ordered response categories (i.e., partial credit items), we again have the choice of a simple Rasch family model (Masters' 1982 partial credit model) or a more general variant such as Muraki's (1992) generalization of Samejima's (1972) graded response model. For tests with smaller sample size, such as state-specific alternate assessments, we recommend the Rasch-family variants because they can be reliably estimated with fewer cases. Under Masters' model, the probability of a response in category i for an item with m_j categories can be written as

$$P\left(x_j = i|\theta_k, b_{j0} \dots b_{jm_j-1}\right) = \frac{e^{\sum_{v=0}^{i} 1.7(\theta_k - b_{jv})}}{\sum_{g=0}^{m_j-1} e^{\sum_{v=0}^{g} 1.7(\theta_k - b_{jv})}}.$$

Muraki's generalization adds an item-dependent discrimination parameter as follows (again, Masters' formulation does not usually include the arbitrary constant 1.7):

$$P\left(x_j = i|\theta_k, b_{j0} \dots b_{jm_j-1}\right) = \frac{e^{\sum_{v=0}^{i} 1.7a_j(\theta_k - b_{jv})}}{\sum_{g=0}^{m_j-1} e^{\sum_{v=0}^{g} 1.7a_j(\theta_k - b_{jv})}}$$

Returning to the likelihood equation, the contribution of each item to the overall likelihood function remains independent of all other items, given θ. This is convenient for two reasons: mixing models within an analysis (e.g., one-parameter and partial credit items on the same scale) becomes no more complicated, and the likelihood of the response pattern may be calculated as the product of the likelihood of responses to individual items.

The number of items that a student answers correctly and the difficulty of the items presented are used to assign maximum likelihood estimates to student ability. In this approach, we maximize the log likelihood function, given as:

$$\ln\left(L(\theta_i)\right) = \sum_{j=1}^{k_i} \ln\left(p_j\left(z_{ji}|\theta_i;\beta_i\right)\right)$$

where θ_i is an estimate of the student score, z_{ji} is student i's response to item j, and $\beta_i$represents the item parameters, treated as fixed and known, associated with the items seen by student i. The function $p_j(.)$ represents the particular IRT model employed for item j. The summation is over the k_i operational items administered to student i. This likelihood function is maximized using Newton Raphson or Steepest Ascent iterations, each potentially with adaptive step sizes. Non-concave likelihoods are maximized starting from multiple starting points.

The student's performance in each content-area test is reported in an overall test score referred to as a *scaled score*. The scaled scores represent a linear transformation of the ability estimates (theta scores). AIR proposes to use MLE scoring based on pre-equated item parameters to score all New Hampshire tests, whether adaptively administered or fixed form.

Calibration Models for Science

AIR is working collaboratively with states to develop ICCR items and item clusters based on common item specifications to measure three dimensional science standards such as New Hampshire's college and career readiness standards. That support includes research on psychometric approaches that take into account some of the complexities inherent to the new item types that are being developed to mesure these standards. For example, we anticipate that there may be local dependencies between items pertaining to the same stimulus (item clusters). As we mentioned at the beginning of this topic, traditional IRT models assume that items are independent given a single underlying trait. The local dependencies between the items pertaining to the same stimulus constitute a violation of the local independence assumption.

Many current ELA assessments also contain clusters of items. For example, often several items share the same reading passage. Currently, the effects of item clusters and the resulting conditional dependencies are typically ignored (resulting in an underestimation of the standard error of measurement for estimated student proficiencies). This approach works reasonably well in practice, and this may be because individual ELA items within a cluster are often written so that the effects of item clusters are reduced; for example items may assess comprehension of different parts of the reading passage. However, for science item clusters, we expect that the conditional dependencies between the items of an item cluster may be too substantial to be ignored because those items are more intrinsically related to each other. For example, the items within an item cluster are organized around a single performance expectation.

The effects of item clusters can be accounted for by including additional dimensions in the IRT model. These dimensions are considered to be nuisance dimensions. Examples of IRT models that follow this approach are the bi-factor model (Gibbons & Hedeker, 1992) and the testlet model (Bradlow, Wainer, & Wang, 1999). The testlet model is a special case of the bi-factor model (Rijmen, 2010).

The 3PL version of the bi-factor model takes the form

$$P\left(x_{jc} = 1|\theta, u_c; a_j, b_j, c_j\right) = c_j + \frac{1-c_j}{1+e^{-a_j\left(\theta+u_c-b_j\right)}} = P_{j1}(\theta, u_c),$$

where u_c indicates the nuisance dimension for item cluster c. The item response function $P_{j1}(\theta, u_c)$ now becomes a response surface that is a function of two latent variables: the latent trait θ representing a student's proficiency in science (the underlying trait of interest), and the nuisance dimension u_c accounting for the conditional dependencies between items of the same item cluster. A Rasch version of the testlet model is obtained by setting the guessing parameters to zero and the discrimination parameters to one, as proposed by Wang and Wilson (2005).

Whereas traditional unidimensional IRT models assume that all items are independent given a single underlying trait θ, the testlet and bi-factor models assume conditional independence of items, given the underlying latent trait θ *and* all nuisance dimensions:

$$P(\mathbf{x}|\theta,\mathbf{u}) = \prod_{c=1}^{C}\prod_{j\in c} P\left(x_{jc}|\theta,u_c\right),$$

where $\mathbf{u}$ is the vector of all C nuisance dimensions. Even though every item cluster introduces an additional dimension, the testlet and bi-factor models do not suffer from dimensionality like other multidimensional IRT models because one can take advantage of special structure of these models during model calibration (Gibbons & Hedeker, 1992; Rijmen, 2010).

Currently, AIR is field testing science item clusters in several states. We will use the field-test data to compare parameter estimates and item fit between traditional unidimensional IRT models and multidimensional models that take item cluster effects into account. Specifically, we will compute fit statistics geared towards assessing conditional item independence and compare them across psychometric models (e.g., the Q_3 index; Yen, 1984, 1993). The empirical results, together with policy considerations, will enable us to select the most appropriate item calibration model for the NH science assessments in collaboration with the NH Department of Education and the technical advisory committee.

Appropriateness for the Type of Scores

The RFP requires the primary student scores to be an overall achievement score and performance level, based on all operational items administered to the student. Because an overall score is the primary score, it is appropriate to use calibration models that incorporate a single primary dimension of interest, like the models proposed. The scales will be constructed as a linear transformation of the underlying trait θ that is common to all items.

We propose to use the maximum likelihood method to compute scores based on the pattern of correct and incorrect student responses to operational items. For the unidimensional IRT models, the likelihood equals the conditional probability for the observed response pattern but considered as a function of theta:

$$L(\theta) = P(\mathbf{x}|\theta) = \prod_{j=1}^{n} P\left(x_j|\theta\right).$$

The likelihood function (or technically, its logarithm) is maximized with respect to θ using the Newton Raphson method.

For the bi-factor and (Rasch) testlet models, the conditional probability $P(\mathbf{x}|\theta,\mathbf{u})$ for an observed response pattern $\mathbf{x}$ contains the vector $\mathbf{u}$ of nuisance dimensions associated with item clusters, in addition to θ, the general proficiency. We propose to marginalize out the nuisance dimensions and maximizing the marginalized likelihood function for θ,

$$L(\theta)_{marginal} = \int_{\mathbf{u}} P(\mathbf{x}|\theta,\mathbf{u})p(\mathbf{u})d\mathbf{u},$$

where $P(\mathbf{x}|\theta,\mathbf{u})$ is defined above and $p(\mathbf{u})$ is the prior (i.e., normal) distribution for $\mathbf{u}$ in the population. The marginal likelihood is maximized with respect to θ using the Newton Raphson method.

Estimation of IRT Models

ICCR items are calibrated using software consistent with the participating ICCR states. Thus, ICCR items administered in multiple states may have item calibrations estimated using different software packages. AIR will use any software requested by the Department to calibrate item parameters. We prefer to use our own software, the IRT Plugin to our *AM* statistical software (*iAM*), which we distribute without charge (Cohen & AIR, 2002). *iAM* offers some capabilities that are not available from other packages, as we discuss below. We note, however, that AIR psychometricians currently use flexMIRT, IRTPRO, Multilog, Parscale, and other software to calibrate item parameters using the more general IRT models in other state assessment programs, and we are therefore prepared to use any software requested by the Department.

iAM offers a key advantage over other available software because it provides design-consistent standard errors and yields accurate estimates of linking errors under complex designs. Like Parscale, *iAM* includes a variety of constructed-response models (including the generalized partial credit model) that can be simultaneously estimated with two- and three-parameter logistic models on the same test. Like Bilog-MG, *iAM* can simultaneously estimate the proficiency distributions in multiple groups. Building on its design-consistent estimators, *iAM* provides fit statistics that reflect the complex sample design.

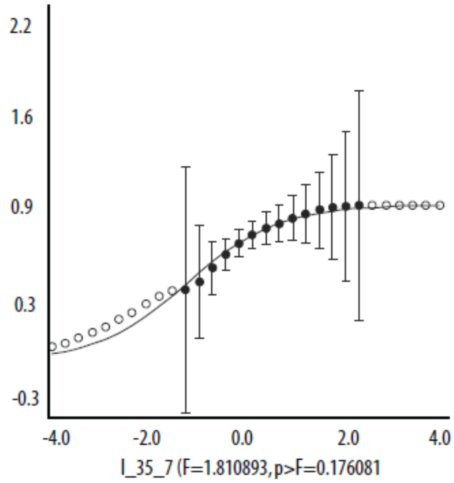
Like Parscale and Bilog-MG, *iAM* estimates item parameters by using marginal maximum likelihood (MML), which provides consistent estimates under the assumed proficiency distribution. MML estimation numerically integrates over the population proficiency distribution. In the absence of analytic solutions to the problem, finding the parameter values that maximize the likelihood of the data requires numerical optimization. The preferred method for finding the maximum likelihood estimate of such models is the use of the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). Bock and Aitkin (1981) propose an EM algorithm for this class of models, and this is the approach taken in popular IRT software.

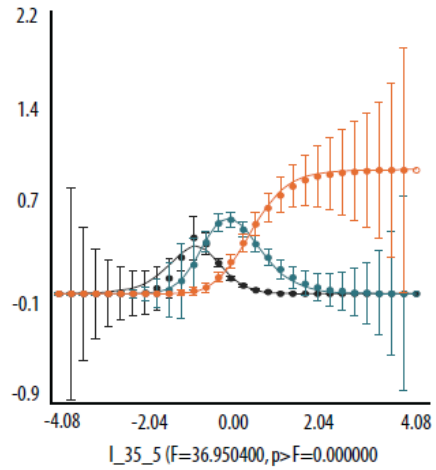
Evaluation of Model Fit

Psychometricians use a variety of measures to evaluate the fit between models based on IRT and observed item data (Yen, 1981; McKinley & Mills, 1985; Fitzpatrick et al., 1996). These statistics operate by evaluating the variance between observed and predicted item responses, a task made difficult by the fact that both the observed and the predicted item responses are contingent on the value of a latent trait (ϑ). Typically, some estimate of ϑ is used, and the discrepancy between the true and estimated values of ϑ is ignored. Fit statistics typically ignore a second source of error as well: the estimates of ϑ and the predictions both depend on a common set of item parameters that are themselves estimates. Hence, IRT fit statistics are notoriously conservative. This problem is compounded by the fact that item data almost always come from a complex, clustered sample. The end result is that typical IRT packages report significant χ^2 statistics for most items, regardless of actual fit.

AIR uses a fit statistic that overcomes these problems, incorporating uncertainty in the item parameter estimates and uncertainty in the estimates of the latent trait and taking into account uncertainty in the item statistics. When these sources of uncertainty are correctly addressed, the χ^2 becomes an F statistic. *iAM* reports fit statistics along with innovative graphics that characterize the item fit, as shown in Exhibit D1.5-3.

Exhibit D1.5-3: Item Fit Graphics





The top panel in Exhibit D1.5-3shows a sample fit graph for a well-fitting multiple-choice item, and the bottom panel shows a sample fit graph for a poor-fitting partial credit item. In both graphs, the x-axis represents the range of student ability values, the y-axis represents probability, the solid line represents the item characteristic curve (ICC), the dots represent the empirical estimates, and the bars indicate the standard error of the distance of the empirical estimates from the ICC. Note that the standard error bars are smaller near the center of the distribution and grow larger toward the tails. When the standard errors grow to be virtually infinite, the empirical estimates are represented as an empty circle.

The item in the first panel of Exhibit D1.5-3 shows a well-fitting item; misfit appears only at the tails of the distribution, where data is sparse and often not enough to draw valid conclusions. The second graph, however, shows misfit along the one-point line (represented as a black ICC). The empirical data points fall substantially off of the ICC line. Although no single point is outside of its confidence limit, the cumulative misfit across points is statistically significant, as depicted by the *F* printed at the bottom of the graph. These fit graphics were designed to provide clearer insights than the fit graphics typically produced by off-the-shelf programs. For example, some graphics use larger symbols to denote scores with more cases, a practice that is visually misleading. Estimates from large samples are precise and should appear at a precise point on the page, while estimates from small samples should appear more diffuse. Our graphics achieve this by graphing standard error bars. Near the center of the proficiency distributions, the estimates are precise, as shown by narrow error bars; they grow diffuse in the tails where little data appear, as shown by wide error bars.

References

Bradlow, E.T., Wainer, H., & Wang, X (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Gibbons, R., & Hedeker, D. (1992). Full-information item bi-factor analyses. *Psychometrika, 57*, 423–436.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. Journal of Educational Measurement, 47, 361–372.

Wang, W.C., & Wilson, M. (2005). The Rasch testlet model. Applied Psychological Measurement, 29, 126–149.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125–145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187–213.

Topic 19 Equating

Topic 19.1 Equating Plan

Equating is a statistical process used to adjust scores on test forms so that scores on different forms can be used interchangeably. Both pre-equating and post-equating methods are used in statewide assessment programs. The principle advantage of pre-equating methods is that they can be applied before the assessment takes place, allowing for immediate scoring and reporting of test results. In addition, they allow the possibility of computer adaptive item selection. However, pre-equating methods rely on the assumption that the item parameters are invariant across measurement occasions. This assumption may be violated when items were calibrated several years prior to their current use (item drift over time) or when the items were calibrated for a different population (differential item functioning).

Post-equating methods are applied after the operational test administration. They rely on a set of anchor items, a representative set of items that are in common between the current test and a previous test (or item bank). After the operational test administration, all items are (re)calibrated, and the anchor items are used to linearly transform the new item parameters so that they are expressed on the existing scale. Whereas pre-equating methods rely on the assumption that the psychometric properties of items did not change from one measurement occasion to another, it is sufficient for post-equating methods that the assumption of item parameter invariance holds for the common items. The item parameters of all other items are always based on the most current administration.

In practice, a mixed approach is often used. For example, the pre-equated item parameters are used by default for all items, but the items that show item drift are recalibrated and post-equated to the base scale using the remaining items as common items.

Adaptive test administrations require the use of pre-equated item parameters to enable adaptive item selection. More generally, one of the most consequential features of online test administration, coupled with automated scoring of test items, is the ability to score and report assessment results immediately. In fact, students and educators participating in online assessments expect to view assessment results immediately or with very little delay. For these reasons, the slight advantages accrued by post-equating may be outweighed by delays in score reporting. Thus, we propose a pre-equating design for scoring and reporting assessment results. However, because the science item clusters and items have yet to be calibrated and placed on a common scale, we also describe an operational test design for implementation in spring 2018 that delivers test forms that meet blueprint and also enact a common-item linking design to allow items to be calibrated on a common scale. We note that immediate reporting of test results will not be possible for spring 2018 in any event since test scoring and reporting will need to await the adoption of performance standards in summer 2018.

Monitoring Scale Drift

AIR will design a study investigating scale drift in the ICCR item banks and present the details of that study to the Department and its technical advisory committee (TAC). To support the Department in this area, there are two key areas of work for AIR to consider. To begin, it is important to understand linking and equating methods that are likely to mitigate the effects of scale drift over time (Harris & Kolen, 1994). AIR is cognizant of these recommendations, and our test development protocols used for large-scale testing programs build linking sets and use linking methods that are known to minimize the effects of scale drift.

Additionally, AIR will provide the Department with a complete research plan for its review and distribution to the TAC. That review will include a complete and comprehensive review of the literature on scale drift and methods for assessing its impact. Various methods for assessing drift have been proposed (Guo, Liu, Dorans, & Feigenbaum, 2011; von Davier, 2009). A technical report following the implementation of the study will be provided to the Department documenting the methods, results, and implications.

Test Design Variation for the First Year of the Science Assessment

To establish a healthy item bank for the duration of the contract and beyond, we propose a balanced incomplete block matrix design for the first year. This will allow for the inclusion of more items in the first year, which will be treated as an operational field test. The availability of statistics and parameters for a larger set of items will result in a better control of the test forms' psychometric properties in subsequent science assessments. Using a balanced incomplete block matrix design does not change our proposed equating plan; every form will be linked to multiple other forms because every block appears in multiple forms. All forms can be calibrated concurrently by treating the non-administered blocks as missing by design.

Concordance Table for ELA and Mathematics

New Hampshire currently uses the Smarter Balanced assessment system for ELA and mathematics.

A sample of ICCR items, proportional to the Smarter Balanced blueprint, were administered as part of the embedded field test pool of Oregon's Smarter Balanced test administration in spring 2015. Embedding these items in the Smarter Balanced test administration allowed for linkages to be established between the ICCR and Smarter Balanced scales. As we describe in Section D1.7 Standard Setting, we propose to identify the approximate location of important benchmarks, including the Smarter Balanced performance standards, in the standard setting ordered-item booklets. These benchmarks will help panelists identify OIB locations where important benchmarks converge, and which represent neighborhoods in which New Hampshire performance standards likely reside.

Topic 19.2 Equating Verification

The Department requires that a quality assurance (QA) contractor independently replicate all calibration, equating, and scoring activities, including analyses based on both simulated and operational test results. AIR works closely with multiple QA contractors across our state assessment projects to provide the highest quality psychometric services to our clients. In some cases, the QA contractor is considered an extension of the client and all deliverables are provided simultaneously to the state department of education and the QA contractor. Other clients prefer to provide the QA contractor with materials directly. We are prepared to adopt any strategy the Department wishes to ensure that all psychometric activities are thoroughly vetted and conducted without error.

Topic 19.3 Equating Report

AIR will produce an annual report documenting the equating process and results. The report will be available for use by the Department in evaluating and approving the results of the equating process prior to reporting. The report will also be part of the annual technical report.

References

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and practices. (2nd ed.) New York, NY: Springer.

Stocking ML, Lord FM (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201–210.

Topic 20 Assessment Evaluation – Item Evaluation

In this section, we describe the procedures AIR uses to evaluate the performance of field-tested ICCR items for inclusion in the ICCR item pools, as well as our procedures for monitoring item performance in operational test administrations. We propose to administer New Hampshire's statewide assessments in ELA and mathematics adaptively, with immediate reporting of test results following establishment of the New Hampshire score after year one. To support both adaptive test administration and immediate scoring and reporting of student assessment results, we propose scoring student records using pre-equated item parameter estimates. ICCR items are embedded in operational test administrations for field testing. Thus, student responses are obtained under operational test conditions so that resulting item parameter estimates are precise and stable. That said, the performance of items can change over time, so in Topic 20.2 Operational Test Item Evaluation we also describe AIR's tools and procedures for monitoring the performance of test items throughout each test administration.

We propose a balanced incomplete block matrix design for assessment of science in spring 2018 that ensures each student is administered a test conforming to a blueprint that ensures each student is assessed on at least 51% of grade band standards, as well as sufficient items to support reporting of subscale performance levels, but also ensures complete coverage of the standards at all aggregate levels of analysis. Under Topic 20.3 Test Construction Evaluation, we describe our procedures for configuring adaptive test administrations that meet all blueprint requirements, as well as our procedures for constructing equivalent fixed-form assessments to ensure psychometric equivalence of test forms within and between science test administrations.

Topic 20.1. Field-Test Item Evaluation

Following the close of the testing window each year, AIR will analyze field-test data in preparation for item data review meetings and the promotion of high-quality test items to the operational ICCR item pool. The item analyses include classical item statistics and item response theory (IRT) item calibrations. Classical item statistics are designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit a bias across subgroups (DIF analyses). The IRT item analyses allow us to examine the fit of items to the measurement model and provide the statistical foundation for operational form construction and test scoring and reporting. Our proposed approach to IRT calibration, equating, and scaling of test items is described in Topic 18 Calibration and Scaling and Topic 19 Equating. In this section, we describe the classical item analyses that will be instrumental in the evaluation of field-test item performance by the data review committees following field-test administration.

Classical Item Analysis

Classical item analyses ensure that the items function as intended with respect to the underlying scales. AIR's analysis program computes the required item and test statistics for each multiple-choice and constructed-response item to check the integrity of the item and to verify that the difficulty of the item is appropriate for the grade level. Key statistics that we compute and examine include the following:

- *Item Discrimination.* The item discrimination index indicates the extent to which each item differentiates between those examinees who possess the skills being measured and those who do not. In general, the higher the value, the better the item is able to differentiate between high- and low-achieving students. The discrimination index for multiple-choice items is calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we compute the mean total number correct for students scoring in each of the possible score categories. Items are flagged for subsequent reviews if the biserial correlation for the keyed (correct) response is less than .05.
- *Distractor Analysis.* Distractor analysis for the multiple-choice items is used to identify items that may have marginal distractors or ambiguous correct responses. In the distractor analysis, the correct response should be the most frequently selected option among high-scoring examinees. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items are flagged for subsequent reviews if the biserial correlation for the distractor response is greater than .05.
- *Item Difficulty*. Items that are either extremely difficult or extremely easy are flagged for review but are not necessarily deleted if they align with the test specifications. For multiple-choice items, we com¬pute the proportion of examinees in the sample selecting the correct answer (p-values), as well as those selecting the incor¬rect responses. For constructed-response items, item difficulty will be calculated both as the item's mean score and as the average proportion correct (this is analogous to p-value and indicates the ratio of an item's mean score divided by the number of points possible). Items are flagged for review if the p-value is less than .25 or greater than .95. We note that the general flagging values described here assume traditional, four-response option multiple-choice items. Items with a different number of response options or no guessing component will have different expected lower bound values.

Analysis of Differential Item Functioning

AIR conducts differential item functioning (DIF) analysis on all field tested items to detect potential item bias across major ethnic and gender groups. The following DIF groups will be included in this analysis:

- Ethnicity
  - African American
  - Hispanic
  - Native American
  - Asian
  - White
- Gender
- Special Education Status
- Limited English Proficient Status
- Free/Reduced Price Lunch Status
- Accommodation Status

Differential item functioning refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because sometimes it is a clue that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in low-income areas are less likely to offer geometry classes, students at those schools might perform more poorly on geometry items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but the curriculum. However, DIF can indicate bias, so all field-tested items are evaluated for DIF, and all items exhibiting DIF are flagged for further examination by a Fairness and Sensitivity Committee. Committee members are asked to reexamine each flagged item, using the statistics as a guide, and to make a final decision about whether the item should be excluded from the pool of potential items given its performance in field testing.

AIR typically uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. IRT ability estimates for each student on the test are used as the ability-matching variable. That score is divided into five intervals to compute the MH chi-square DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta for the multiple-choice items, the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the constructed-response items. The purification method described by Holland and Thayer (1988) is included in the DIF procedure.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention illustrated in Exhibit D1.5-4. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, or female), or negative DIF (i.e.,–A,–B, or–C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the “C” category for any group. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. These items are flagged regardless of whether the DIF statistic favors the focal or reference group. A Fairness and Sensitivity Committee reviews all items flagged on the basis of DIF statistics. Committee members are encouraged to discuss these items and are asked to decide whether each item should be excluded from the pool of potential items given its performance in field testing.

Exhibit D1.5-4: DIF Classification Rules for Items

Dichotomous Items	
Category	Rule
C	$MH\chi^2$ is significant and $\lvert\hat{\Delta}_{MH}\rvert \geq 1.5$
B	$MH\chi^2$ is significant and $\lvert\hat{\Delta}_{MH}\rvert < 1.5$
A	$MH\chi^2$ is not significant.
Polytomous Items	
Category	Rule
C	$MH\chi^2$ is significant and $\lvert SMD\rvert / \lvert SD\rvert \geq .25$.
B	$MH\chi^2$ is significant and $\lvert SMD\rvert / \lvert SD\rvert < .25$.
A	$MH\chi^2$ is not significant.

Item Statistic Flags

Following our analysis of field-test items, items are submitted to a second round of external reviews before they are eligible for selection into the operational test item bank. As a first step in the field-test item review, field-test items that did not perform as expected are flagged for additional review. Flagging rules for items are configurable and are defined in the Analysis Specifications document. The flagging criteria presented here are for illustration only but do represent values typical for statewide assessment programs:

- *Item Discrimination Flags.* Multiple-choice items are flagged for subsequent review if the biserial correlation for the item is less than .25 for the keyed (correct) response and greater than zero for distractors. For constructed-response items, items are flagged if the polyserial correlation is less than .25.
- *Item Difficulty Flags.*
  - Multiple-choice items are flagged for review if the p-value is less than .25 or greater than .95, but flagging rules are configurable. Multiple-choice items are also flagged when the keyed response is not the modal response. In conjunction with low or negative biserial correlations, non-modal responding to the keyed response may indicate that the item has been miskeyed. These general flagging values assume traditional, four-response option multiple-choice items. Items with a different number of response options or no guessing component will have quite different expected lower bound values.
  - Constructed-response items are flagged when the proportion of students assigned any score point category is greater than .95. A very high proportion of students in any single score point category may suggest that the other score points are not useful or, if the score point is in the minimum or maximum score point category, that the item may not be grade-appropriate. In addition, constructed-response items are flagged if the average ability estimate of students in a score-point category is lower than the average ability estimate of students in the next lower score-point category. For example, if students who receive three points on a constructed-response item score lower, on average, on the total test than students who received only two points on the item, the item will be flagged for review.
- *DIF Flags.* A DIF classification of C means that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. Items in the C category for any group are flagged and reviewed by the Fairness Data Review Committee. Exhibit D1.5-4 details the DIF classification rules.

The flagging rules presented above are used to identify items for external review committees but do not constitute the extent of AIR's review of test items. As described in Topic 18, AIR also examines the fit of IRT parameters to the underlying measurement model. The information provided by classical and IRT item analyses is in many ways redundant, but each approach provides a different perspective on item functioning. However, because classical item statistics are more intuitive for external item committees to understand, AIR recommends presenting only those statistics for item data review.

Item Statistics Review

Items flagged for review on the basis of any of these criteria typically must pass through an additional series of reviews to be included in the final item pool from which operational forms are created. First, a team of AIR psychometricians reviews all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct, and there are no other obvious problems with the items. AIR content staff then review each of the flagged items to ensure that the items are accurate and administered as intended. In addition, Department assessment and curriculum staff, as well as its Content Review Committee and Fairness Review Committee, meet to re-evaluate flagged field-test items in the context of each item's statistical performance.

Effectively evaluating the quality of test items, including the alignment of the knowledge and skill requirements of test items to New Hampshire's college career ready standards, requires reviewers to be able not only to view but to operate on test items using the same interface and available tools administered to students. The Web Preview feature in the Item Tracking System (ITS) uses the same software to render test items that is used by the test delivery system, so that AIR and Department test development staff and reviewers for the Content Review Committee and Parent Review Committee can fully evaluate the knowledge and skill requirements of test items as administered online. Even items that are nominally consistent across test administration modes may be rendered quite differently in paper and online. An example is the display of reading passages or other stimulus-bound items, where stimuli and items are displayed simultaneously to students in split screens. For these reasons, electronic review of test items by external review committees is critical.

Topic 20.2. Operational Test Item Evaluation

To support adaptive test administration and immediate reporting of assessment results, AIR proposes to score student assessment results using pre-equated item parameters. Because item parameter estimates are based on student responses to field-test items embedded in operational test administrations, bank item parameters are precise and stable over time. Nevertheless, the behavior of items may change over time, so AIR psychometricians actively monitor the performance of test administrations throughout the testing window using a set of quality assurance (QA) reports generated from the Quality Monitor (QM) system. Throughout the testing window, QA reports will be routinely generated and evaluated to ensure the quality of New Hampshire's statewide assessments. The QA reports provide information on item behavior, blueprint match rates, and item exposure rates. Additional reports, detailed in Topic 8.2, include a forensic analysis report that flags unlikely patterns of behavior in testing administrations aggregated at the test administration, test administrator, and school levels. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to blueprint and items are performing as anticipated.

Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Exhibit D1.5-5 presents an overview of the QA reports.

Exhibit D1.5-5: Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology items)
Blueprint Match Rates	To monitor unexpected low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

Item Analysis Report

The item analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as IRT-based item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

Item p-Value. For multiple-choice items, the proportion of students selecting each response option is computed. For constructed-response, performance, and technology items, the proportion of student responses classified at each score point is computed. For multiple-choice items, if the keyed response is not the modal response, the item is also flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

Item Discrimination. Biserial correlations for the keyed response for selected-response items and polyserial correlations for polytomous constructed-response, performance, and technology items are computed. AIR psychometric staff evaluates all items with biserial correlations below a target level, even if the obtained values are consistent with past item performance.

Item Fit. In addition to the item difficulty and item discrimination indices, an item fit index is produced for each item. For each student, a residual between the observed and the expected score given the student's ability is computed for each item. Each residual is averaged across all students, and the average residual is used to flag an item.

We begin by defining $p_{ij} = pr(z_{ij} = 1)$, representing the probability that student *i* responds correctly to item *j* (z_{ij} represents the student's score on the item). For selected-response items, we use the 3PL IRT model to calculate the expected score on item *j* for student *i* with estimated ability $\hat{\theta}$ as

$$E(z_{ij}) = c_j + (1 - c_j)\frac{\exp\left(Da_j(\hat{\theta}_i - b_j)\right)}{1+\exp\left(Da_j(\hat{\theta}_i - b_j)\right)}.$$

For constructed-response, performance, or technology items, using the generalized partial credit model, the expected score for student *i* with estimated ability $\hat{\theta}$ on an item *j* with a maximum possible score of K_j is calculated as

$$E(z_{ij}) = \sum_{l=1}^{K_j} \frac{l\exp\left(Da_j \sum_{k=1}^{l}(\hat{\theta}_i - b_{j,k})\right)}{1 + \sum_{m=1}^{K_j} \exp\left(Da_j \sum_{k=1}^{m}(\hat{\theta}_i - b_{j,k})\right)}$$

.

For each item *j*, the residual between observed and expected score for each student is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij})\ .$$

The statistic δ is aggregated across students of different abilities for each item

$$\bar{\delta}_j = \frac{1}{n}\sum_{i=1}^{n}(\delta_{ij})\ .$$

The report can be configured to report all items or to flag and report only those items where the fit index is above a given threshold. For example, items could be flagged when

$$\frac{\bar{\delta}_j}{se(\bar{\delta}_j)} > 1.96$$

$$\text{where } \bar{\delta}_j = \frac{SD(\delta_{ij})}{\sqrt{n}}.$$

Blueprint Match Report

The QA system generates two blueprint match reports for each assessment, one based on evaluation of match to content standards and a second for evaluating cross-cutting specifications such as item type and depth of knowledge (DoK) level. For each blueprint element (e.g., strand, standard, benchmark), the report indicates the minimum and maximum number of items specified in the blueprint, the number of test administrations in which those specifications were met, the number of administrations in which the blueprint requirements were not met, and, for administrations in which specifications were not met, the number of items by which the requirement was not met.

Item Exposure Rates

This report allows test items to be monitored for unexpectedly large exposure rates or unusually low item pool usage throughout the testing window. As with other reports, it is possible to examine the exposure rate for all items or flag items with exposure rates that exceed an acceptable range. Often, item overexposure indicates a blueprint element or combination of blueprint elements that are underrepresented in the item pool and should be targeted for future item development. Such item overexposure is also usually anticipated in the simulation studies used to configure the adaptive algorithm.

Cheating Detection

As we detail in Topic 8.2, the QA system also provides a forensics report to identify possible irregularities in test administration for further investigation. Unusual patterns of responding at the student level are aggregated to the test session, test administrator, and school levels to identify possible group-level testing anomalies. AIR psychometricians monitor testing anomalies throughout the test administration window. Evidence evaluated includes changes in test scores across administrations, item response times, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. The analyses used to detect the testing anomalies can be run anytime in the testing window.

Topic 20.3. Test Construction Evaluation

Test construction proceeds somewhat differently in the context of adaptive versus fixed-form assessments. In a fixed-form test, the ideal test information function is decided once. Each year, a team of content experts and psychometricians makes a sequence of decisions about items to include. The test construction team may allow some deviation from the ideal information function to get a better match to the blueprint. The team may allow reuse of some items from a prior administration to achieve the desired balance. The number and types of compromises made often depend on the items available in the operational item bank. Participants draw on their experience and professional judgment to make these decisions but rarely articulate the guiding principles.

In an adaptive, criterion-referenced test, these decisions take on a different character. The system does not construct a test but rather implements an algorithm for constructing many tests. Calculations and algorithmic logic take the place of professional judgment. AIR's adaptive algorithm is highly configurable. This allows variable weighting of different blueprint constraints and provides several mechanisms for controlling the balance between blueprint match and the precision of measurement. In addition, the adaptive algorithm provides the ability to limit item exposure. In a fixed-form test, all items are overexposed each year.

Blueprint Construction

Sample blueprints are provided in Topic 1. Distribution of item ranges is designed to accomplish several important goals. First, as criterion-referenced tests, New Hampshire's statewide assessments must reflect a representative sample of the knowledge and skills that students are expected to achieve by the end of each school year or course. In addition, to support student-level reporting of domain scores, each domain must be assessed with sufficient items to support reliable subscale performance-level reporting. Whether New Hampshire chooses adaptive or fixed-form tests, the blueprints are designed to accomplish both goals, with items targeting the full breadth of the academic content standards but constrained to ensure a sufficient number of items in each domain to achieve reliable subscale reporting. Moreover, item ranges for DoK levels and item types ensure that students, regardless of overall ability, are administered test items representing the full range of cognitive complexity and provided the opportunity to demonstrate their knowledge and skills through the range of item types.

Configuration of Adaptive Tests

Constructing forms for adaptive tests consists of two steps: establishing the item pool and optimizing the item selection parameters for the selected pool. Our adaptive engine is robust to imbalances in the item pool as long as there is a sufficient number of items in the various blueprint categories. Therefore, the selection of items for inclusion in adaptive pools is less critical. When items are clustered into groups, as with reading tests, we must also pay attention to the composition of the item sets and the degrees of freedom available to the algorithm to select from among them. To configure and evaluate the performance of the adaptive algorithm, AIR psychometricians use our simulation tool.

Item Pool. As long as there are sufficient items to meet each of the blueprint specifications, the adaptive algorithm is robust with respect to item pool composition. It is therefore not necessary that the distribution of content standards in the item pool be proportional to the test blueprint. Nevertheless, AIR test development and psychometric staff inventory the item pools before each test administration to identify any gaps in the pool that should be especially targeted for item development.

Item development in a fixed-form environment typically aims to concentrate test information near the critical proficient cut score. When all students are administered the same form, the goal is to minimize measurement error near the location of important cut scores. In adaptive environments, of course, the goal is to minimize error near each student's ability level. This places a greater premium on developing items that more effectively measure students near the ends of the ability distribution, increasing the proportion of very easy and very difficult items. As part of our item pool inventory, we also evaluate the distribution of item difficulties in content domains to guide our item development strategy.

With respect to item pool adequacy, we note that in an adaptive assessment, the evaluation of items and their selection into the operational pool proceed somewhat differently than they do in a fixed-form assessment. In a fixed-form environment, the criteria for promoting items to the operational bank may be somewhat relaxed, allowing the inclusion of items that may not have optimal statistical characteristics but that are intended only to be available for placement in a test form when no other suitable items are available. Thus, item statistics are generally evaluated twice, once during data review for inclusion in the operational pool and again at form construction for placement in operational test forms. In an adaptive assessment, all items in the pool are available for test administration. For this reason, the criteria for including items in the operational bank should be more stringently applied than might be the case in a fixed-form environment. All items in the adaptive item pool must meet minimum requirements for item discrimination, be within the acceptable range of item difficulties, and demonstrate appropriate fit statistics. The adaptive algorithm does not recognize items as being for use only in emergencies.

Simulation Tool. The simulator allows psychometricians to adjust parameters in the adaptive algorithm to optimize it for use with a particular item pool. The simulator outputs a host of indicators including, but not limited to,

- estimates of bias at the overall and reporting category levels;

- standard errors and distributions of standard errors;
- match to blueprint;
- number of unique benchmarks administered to each student;
- item exposure;
- number of item groups administered; and
- item reuse across opportunities.

Psychometricians adjust configuration parameters to achieve the optimal mix of outcomes. For new assessment systems, they may recommend changes to the blueprint, extending or restricting the range of items on each content standard or with each affinity group. Sometimes small changes can improve outcomes, and we encourage our psychometricians to inform our clients when this is the case. Of course, any substantive blueprint change remains the Department's decision. Psychometricians interact with the simulator by modifying the configuration parameters in a web-based environment. When the configuration parameters are finalized, this same configuration is uploaded to ITS to generate a new version of the configuration file. Simulations are run on the new file as a final check that nothing has changed and that all procedures were followed. The Department will receive a final report on all simulations containing the detailed statistical output.

Fixed-Form Test Construction

We propose to administer the science assessments using a matrix design of fixed forms that ensures (1) that individual students are administered test forms that measure more than 50% of all science standards and conform to blueprint specifications for reporting categories and DoK and (2) that at all aggregate levels, the full range of content standards is assessed. In this section, we describe the process AIR uses to build fixed-form assessments that meet all blueprint specifications and measure equivalently across forms.

As described above, form construction proceeds from the test blueprint. This blueprint describes the content to be covered, the DoK with which it will be covered, the type of items that will measure the constructs, and every other content-relevant aspect of the test. In addition, statistical targets ensure that students will receive scores of similar precision, regardless of which form of the test they are administered.

AIR's FormBuilder software assists our test developers as they construct operational forms. FormBuilder interfaces with ITS to extract item and test information and interactively evaluates the form's match to blueprint to ensure that test forms meet all blueprint specifications. Moreover, as items are added to each form, FormBuilder automatically updates test characteristics curves (TCCs) and standard error of measurement curves (SEMCs) to ensure that test forms are statistically parallel.

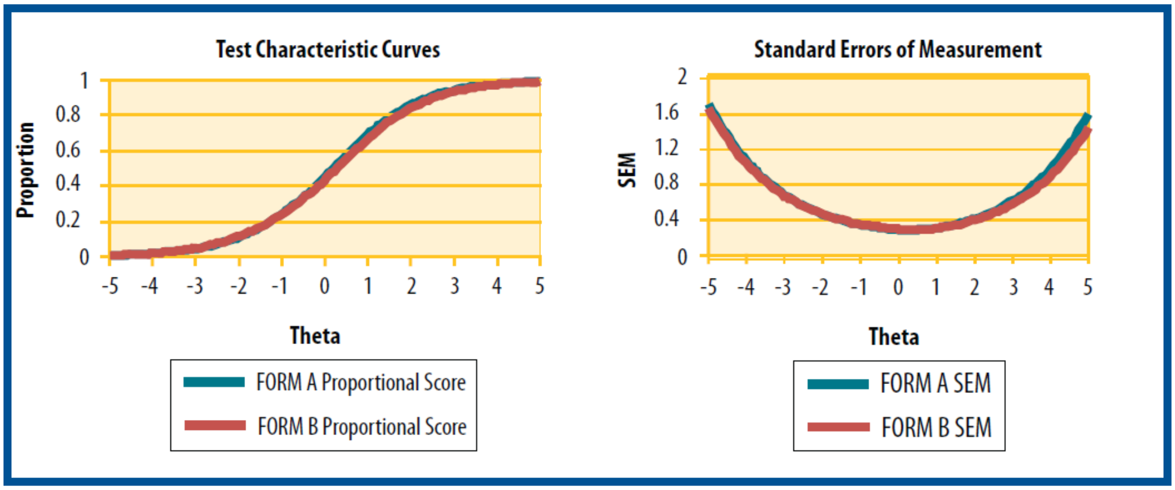
All test forms will be constructed so that they are strongly parallel. This means that not only that test content will be parallel across test forms, but also that the test scales and performance levels will be equivalent from form to form and that the raw score to scale score comparisons for each form will be as similar as possible, given the characteristics of the item bank.

Using FormBuilder, our content specialists select test items that match the blueprint and are of appropriate difficulty. Beginning with content considerations and supplementing those considerations with statistical considerations, AIR will create alternate, parallel test forms by comparing TCCs for the form being created with TCCs from the reference form. To the degree that the TCC for the total test is the same as for previous tests, the raw score required for meeting any performance standard will remain as close to the same as it was on the reference form.

Not only will AIR construct alternate, parallel forms at the total test level, we will also attempt to create alternate forms that have similar subscale psychometric structures, as well. This goal may be more difficult to attain than equivalence at the total test level because of the smaller number of items from which to select.

As test developers build forms, FormBuilder generates TCCs and SEMCs for the reference (previously used) form and the target (new) form(s) on the screen. The TCCs and SEMCs are plotted using a different color trace line for each prototype form. Sample TCCs and SEMCs are provided in Exhibit D1.5-6.

Exhibit D1.5-6: Test Characteristic Curves and Standard Errors of Measurement



At each point in the form construction process, the test developer can see the exact difficulty relationship between the reference and target forms. If the difficulty and error structure for the target forms is virtually identical to the reference form, as in the sample TCCs and SEMCs, then the item selection process concludes with multiple, parallel test forms. Once the goal of parallel forms is achieved, the information is entered into ITS, which tracks item usage and generates bookmaps (test maps) for use in scoring, form development, and other processes.

Upon construction of test forms that meet blueprint specifications and are psychometrically equivalent to the reference form, the test developer submits the draft form to AIR psychometricians for review and approval. AIR test development and psychometric staff work closely to identify possible form revisions that will result in greater equivalence of test scores with respect to both the subject area test score and subscales. Upon sign-off by psychometric staff, proposed test forms are submitted to the state for their review. When submitting test forms to the state for review, AIR will produce a form evaluation workbook that includes an evaluation summary checklist as well as summary statistics and test characteristic graphs.

Topic 20.4. Additional Assessment Analyses

We begin this section with a review of the procedures we implement to ensure that all items are scored accurately. We then describe our procedures for scoring tests, including our simulation procedures to ensure that test records are accurately scored, and our processes for monitoring test scoring throughout the test administration window.

As we describe in Topic 16 Machine Scored Items, AIR has developed a large suite of machine-scored item types that provide very fast scoring across a range of selected- and constructed-response item types. Most machine-scored item types are scored using explicit, test developer-defined rubrics contained in the item, ensuring fidelity to scoring rules across all test administrations. The rule-based rubrics allow test developers to construct test items that measure reason-based problem solving, and validity of test score interpretation is increased because the rubrics explicitly define how students must respond to achieve each score point rather than providing guidelines for the kinds of responses associated with each score point.

We propose to use our Automated Essay Scoring (AES) engine, which uses statistical algorithms to score student essays for the writing assessment in real time. The statistical rubrics used to develop the scoring models measure a broad set of features, some of which may be item specific and “learned” from a training set. During training, these features are related to human scores through a statistical model. The resulting equation predicts how a human would score a response with the measured features.

Autoscore is currently being used to accurately and quickly score student essay responses in Arizona, Ohio, and Utah. Statistical scoring models have already been developed to score most ICCR writing assessments, so automated scoring of essay responses for New Hampshire’s statewide assessments can begin immediately. In addition, we note that there are additional ICCR writing assessments for which scoring models have not yet been developed. Since reporting of assessment results for spring 2018 must necessarily await the outcome of standard setting workshops, the Department may also choose to administer writing tasks for which models have yet to be developed. Following the procedures described below, a sample of essay responses for training and cross validation will be identified, and for which optimally valid human scores will be obtained. From the responses, statistical scoring models will be developed allowing for immediate scoring and reporting of essay responses for those items in spring 2019. In the following paragraphs we describe how the AES scores student essays, and how we maintain the validity of all writing scores.

Scoring rubrics for all machine-scored items are highly secure because they never leave our host servers. Although item scoring is nearly instantaneous, a modular, asynchronous scoring framework allows the adaptive algorithm to proceed with item selection even before completion of scoring of submitted item responses, whether machine- or hand-scored. This ensures that students never need to wait for items to be scored to proceed through the test, even if demands on the network slow the scoring of some machine-scored items, or if item responses require hand-scoring following test administration.

Creating and Refining Machine Rubrics for Scoring Engines with Explicit Rubrics

Item writers and reviewers first develop rubrics for each explicit rubric scoring engine. Our rubrics support true rule-based reasoning, which is much more flexible and powerful than the simple token-matching approaches that are more common. True rule-based items allow more flexible scoring and admit rubrics that are more tightly aligned to the constructs being measured. These types of rubrics can be used for true constructed-response items. As with any constructed-response items, examinees may generate responses not anticipated by the test developers. Therefore, each item goes through a process similar to rangefinding for human-scored items.

AIR has developed a process called *rubric validation* that efficiently reviews scoring rubrics for true rule-based scoring. This process is supported by our REVISE software.

The process of creating and refining machine rubrics for scoring engines with explicit rubrics is elaborated upon in Topic 16.1 Automated Scoring of Student-Generated Responses. The rubric validation process begins with the selection of student responses to review. The sample is selected to over-represent anomalous records (e.g., over-sampling responses for records in which students performance on MC items is high, but performance on target item is poor). A review committee then reviews the student responses, notes observations about each response, designates a consensus score for each response, and selects additional samples to review according to a variety of sampling schemes. The entire process is facilitated by our REVISE software, which is a secure web-based application that selects and presents responses, gathers committee input, and updates ITS with the results.

For equation and proposition items, the rubric validation process bolstered with a validation study that compares the performance of the machine-scoring rubric with handscoring. To execute the validation study, a random sample of 500 cases from each item are handscored, and discrepancies from the machine score are reviewed and resolved. Human- and machine-scoring agreement rates are reported for each item.

Automated Essay Scoring Training

AES uses a statistical algorithm to score essay responses. The statistical scoring models also yield an indicator of score confidence based on (1) responses with unusual features and (2) responses scoring near rubric thresholds. For each model, a confidence threshold can be identified, and any scored response with a confidence value below the threshold will be automatically routed for verification scoring by a trained reader.

The engine employs a training set, a set of essays scored with optimally valid scores, which we obtain by having all responses double- scored by expert scorers and using an adjudication process for discrepant scores. The quality of the human-assigned scores is critical to the identification of a valid model and final performance of the scoring engine. The essay scoring engine is a statistical scoring engine. Measurable features of the essays are extracted and statistically analyzed. The measurable features include both syntactic features (e.g., misspellings, sentence length, grammar mistakes) and semantic features.

The semantic features reflect a summary of the words used in the documents. The summary is formed by first reducing words to their roots (e.g., “eating” becomes “eat”) and creating a matrix of words (or n-grams, which are sequences of words) by responses. This, of course, is a very large sparse matrix. The matrix is factor analyzed to reduce its dimensionality, and the factors become the semantic predictors.

During training, the first step separates a random sample of approximately 500 cases to reserve for independent validation of the final model.

The second step removes responses flagged for deterministic condition codes. The computer, of course, is quite good at identifying responses that are too short to score, in which the prompt is copied, or where the student simply copies the same text over and over. These cases are flagged and removed from the training. The remaining cases, which still include cases to which humans assigned other condition codes, such as “off topic,” are decomposed into the variables that comprise the predictors in the model. These predictors, or subsets of them, are included in ordered probit models predicting (1) whether a condition code would be applied, and (2) the score that would be applied. Each prediction comes with a confidence index. An algorithm takes this information into account in determining whether to assign a condition code or a score.

When our psychometricians are satisfied with the models, they are run against the validation sample to estimate the agreement between the scoring engine and the fully resolved human score.

The essay scoring system generates several confidence indices, which are then used to determine whether the paper should receive a score or a condition code. The mean and standard deviation of the confidence index are computed for each dimension, and the lowest value is selected for flagging responses with one or more low confidence scores. Any dimension score with a confidence index below this threshold is flagged for verification by a human rater.

We propose to route papers with the lowest confidence scores to human scorers. In the past this has amounted to approximately 20% of the papers with the expected turnaround being seven business days. This would mean that 80% of students get immediate scores in AIR’s reporting systems and the other 20% get their scores within seven business days.

The process of automated essay scoring training is elaborated upon in Topic 16.1 Automated Scoring of Student-Generated Responses.

Exhibit D1.5-7 presents agreement indicators for the two initial human raters and between the resolved human and statistical rubric score. Indicators include percent exact agreement, Pearson’s correlation, a quadratic weighted kappa statistic, and the standardized mean difference between the scores. Although absolute values for evaluating statistics have been advanced (Condon, 2013; Higgins, 2013), the focus of these comparisons is degradation of agreement when moving from human-human agreement to machine-human agreement. Agreement between human raters is an indicator of how reliably the responses can be scored by human raters. Because the statistical rubrics attempt to reproduce human-assigned scores, evaluation of machine-human agreement is with respect to observed human-human agreement. Responses with poor human-human agreement will not be reliably scored by either humans or machines.

Scoring Quality Assurance

Before the opening of the testing window for each administration, the testing system and content are deployed to a staging server for user acceptance testing (UAT). UAT of the TDS serves both a software evaluation and content approval role. The UAT period provides the Department with an opportunity to interact with the exact test with which the students will interact. As we describe in Topic 16 Machine Scored Items, because all attributes of the item, including the machine-scoring rubrics, travel with the item, and because our Web Preview system uses the same display software used by the TDS, the

Exhibit D1.5-7: Summary of Human and Machine Scores for AIRCore Writing Prompts

Grade	ITS ID	Dimensions	Score Point	*N*	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	%z Exact	Pearson *r*	Weighted *κ**	SMD*	% Exact	Pearson *r*	Weighted *κ**	SMD*
3	13021	Conventions	2	2092	1.43	1.55	0.75	0.71	0.69	0.65	0.65	0.03	0.72	0.71	0.70	0.16
		Evidence	4		1.93	1.90	0.78	0.61	0.65	0.64	0.64	0.02	0.65	0.65	0.63	0.05
		Organization	4		1.93	2.00	0.76	0.66	0.66	0.67	0.67	0.00	0.67	0.66	0.65	0.10
3	13022	Conventions	2	2093	1.47	1.61	0.69	0.64	0.71	0.67	0.67	0.02	0.72	0.67	0.66	0.20
		Evidence	4		2.02	2.02	0.75	0.63	0.60	0.64	0.64	0.00	0.67	0.66	0.65	0.00
		Organization	4		2.12	2.10	0.74	0.66	0.64	0.68	0.68	0.00	0.68	0.65	0.64	0.03
3	13023	Conventions	2	2090	1.51	1.57	0.72	0.65	0.70	0.65	0.65	0.01	0.73	0.67	0.66	0.09
		Evidence	4		1.89	1.93	0.74	0.65	0.62	0.60	0.60	0.03	0.68	0.68	0.67	0.05
		Organization	4		1.95	1.92	0.77	0.61	0.64	0.66	0.66	0.02	0.68	0.66	0.64	0.04
3	13024	Conventions	2	2096	1.44	1.53	0.70	0.67	0.71	0.66	0.66	0.01	0.76	0.68	0.67	0.13
		Evidence	4		1.93	1.90	0.76	0.64	0.63	0.64	0.64	0.04	0.64	0.63	0.62	0.04
		Organization	4		1.96	1.96	0.80	0.65	0.63	0.66	0.66	0.05	0.64	0.64	0.63	0.00
3	13025	Conventions	2	2093	1.37	1.48	0.76	0.70	0.66	0.58	0.58	0.01	0.71	0.67	0.66	0.15
		Evidence	4		1.94	1.97	0.72	0.61	0.63	0.65	0.65	0.00	0.69	0.65	0.64	0.04
		Organization	4		1.92	1.86	0.82	0.71	0.61	0.64	0.64	0.01	0.64	0.68	0.68	0.08
3	13026	Conventions	2	2090	1.45	1.55	0.73	0.67	0.71	0.66	0.66	0.00	0.75	0.68	0.67	0.15
		Evidence	4		1.94	1.94	0.74	0.68	0.66	0.68	0.68	0.04	0.72	0.71	0.71	0.01
		Organization	4		2.04	2.02	0.80	0.71	0.64	0.68	0.68	0.05	0.63	0.65	0.65	0.03
4	13094	Conventions	2	2095	0.95	0.95	0.75	0.68	0.66	0.67	0.67	0.00	0.65	0.65	0.65	0.01
		Evidence	4		1.30	1.27	0.47	0.47	0.77	0.52	0.52	0.00	0.82	0.58	0.58	0.08
		Organization	4		1.40	1.34	0.51	0.49	0.74	0.56	0.56	0.01	0.83	0.66	0.66	0.11
4	13095	Conventions	2	2096	1.17	1.17	0.67	0.63	0.64	0.62	0.62	0.01	0.67	0.59	0.59	0.01
		Evidence	4		1.35	1.24	0.53	0.45	0.75	0.57	0.57	0.00	0.81	0.63	0.60	0.22
		Organization	4		1.54	1.51	0.59	0.54	0.71	0.59	0.59	0.03	0.73	0.56	0.56	0.06
4	13118	Conventions	2	2096	1.15	1.16	0.71	0.65	0.64	0.60	0.60	0.01	0.67	0.63	0.63	0.01
		Evidence	4		1.33	1.29	0.49	0.48	0.76	0.55	0.55	0.01	0.84	0.64	0.64	0.07
		Organization	4		1.56	1.53	0.61	0.56	0.71	0.59	0.59	0.03	0.77	0.67	0.67	0.04

Exhibit D1.5-7: Summary of Human and Machine Scores for AIRCore Writing Prompts (continued)

Grade	ITS ID	Dimensions	Score Point	*N*	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	%z Exact	Pearson *r*	Weighted *κ**	SMD*	% Exact	Pearson *r*	Weighted *κ**	SMD*
4	13119	Conventions	2	2094	1.15	1.19	0.72	0.63	0.66	0.64	0.64	0.02	0.66	0.63	0.63	0.06
		Evidence	4*		1.38	1.30	0.54	0.49	0.73	0.53	0.53	0.05	0.77	0.57	0.56	0.14
		Organization	4		1.51	1.46	0.60	0.52	0.72	0.60	0.60	0.01	0.75	0.60	0.60	0.10
4	13120	Conventions	2	2091	1.05	1.09	0.70	0.67	0.67	0.66	0.66	0.02	0.68	0.64	0.64	0.06
		Evidence	4		1.28	1.20	0.49	0.42	0.77	0.54	0.54	0.04	0.85	0.65	0.63	0.17
		Organization	4		1.49	1.43	0.58	0.53	0.74	0.63	0.63	0.03	0.79	0.65	0.64	0.11
4	13121	Conventions	2	2096	1.10	1.08	0.69	0.59	0.67	0.65	0.65	0.03	0.68	0.61	0.60	0.02
		Evidence	4*		1.34	1.27	0.54	0.49	0.77	0.60	0.60	0.03	0.81	0.65	0.64	0.14
		Organization	4*		1.53	1.45	0.58	0.54	0.72	0.61	0.61	0.03	0.74	0.59	0.59	0.13
5	13236	Conventions	2	2099	1.41	1.57	0.67	0.62	0.74	0.69	0.69	0.02	0.76	0.69	0.67	0.25
		Evidence	4		1.81	1.79	0.58	0.52	0.71	0.59	0.59	0.01	0.79	0.64	0.64	0.03
		Organization	4		1.92	1.88	0.68	0.58	0.70	0.65	0.65	0.03	0.73	0.67	0.66	0.05
5	13237	Conventions	2	2095	1.30	1.40	0.74	0.67	0.73	0.72	0.71	0.04	0.73	0.69	0.68	0.13
		Evidence	4		1.59	1.53	0.60	0.53	0.73	0.61	0.61	0.04	0.76	0.62	0.62	0.09
		Organization	4		1.75	1.75	0.66	0.57	0.72	0.66	0.66	0.01	0.72	0.64	0.64	0.01
5	13238	Conventions	2	2099	1.47	1.51	0.62	0.61	0.72	0.65	0.65	0.00	0.75	0.65	0.64	0.06
		Evidence	4		1.87	1.88	0.64	0.53	0.69	0.63	0.63	0.01	0.75	0.63	0.62	0.02
		Organization	4		1.95	1.99	0.68	0.56	0.70	0.65	0.65	0.01	0.74	0.62	0.61	0.06
5	13239	Conventions	2	2095	1.41	1.51	0.69	0.60	0.73	0.66	0.66	0.02	0.75	0.68	0.67	0.15
		Evidence	4		1.67	1.67	0.62	0.56	0.65	0.56	0.56	0.02	0.74	0.63	0.63	0.01
		Organization	4		1.92	1.93	0.64	0.52	0.71	0.65	0.65	0.02	0.76	0.63	0.61	0.03
5	13246	Conventions	2	2093	1.36	1.45	0.68	0.65	0.72	0.68	0.68	0.01	0.73	0.69	0.69	0.13
		Evidence	4		1.54	1.58	0.57	0.54	0.72	0.59	0.59	0.03	0.77	0.61	0.60	0.07
		Organization	4		1.81	1.82	0.66	0.57	0.71	0.65	0.65	0.01	0.73	0.64	0.64	0.02
5	13247	Conventions	2	2097	1.38	1.43	0.68	0.63	0.72	0.67	0.67	0.01	0.75	0.71	0.71	0.07
		Evidence	4		1.77	1.80	0.67	0.59	0.65	0.60	0.60	0.02	0.73	0.65	0.65	0.05
		Organization	4		2.00	1.97	0.69	0.57	0.69	0.66	0.66	0.02	0.72	0.63	0.62	0.04

Exhibit D1.5-7: Summary of Human and Machine Scores for AIRCore Writing Prompts (continued)

Grade	ITS ID	Dimensions	Score Point	*N*	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	%z Exact	Pearson *r*	Weighted *κ**	SMD*	% Exact	Pearson *r*	Weighted *κ**	SMD*
6	13304	Conventions	2	2097	1.43	1.52	0.67	0.62	0.67	0.57	0.57	0.02	0.76	0.72	0.71	0.14
		Evidence	4		1.74	1.75	0.65	0.61	0.63	0.57	0.56	0.04	0.74	0.66	0.66	0.02
		Organization	4		1.89	1.86	0.74	0.65	0.62	0.61	0.61	0.01	0.69	0.68	0.67	0.04
6	13305	Conventions	2	2095	1.45	1.59	0.68	0.61	0.66	0.57	0.57	0.00	0.76	0.69	0.67	0.22
		Evidence	4		1.53	1.43	0.60	0.55	0.70	0.58	0.58	0.01	0.74	0.61	0.60	0.17
		Organization	4		1.62	1.60	0.68	0.62	0.65	0.59	0.59	0.01	0.70	0.62	0.62	0.02
6	13306	Conventions	2	2097	1.47	1.54	0.69	0.64	0.71	0.64	0.64	0.02	0.75	0.68	0.67	0.11
		Evidence	4		1.67	1.63	0.64	0.57	0.65	0.55	0.55	0.00	0.71	0.61	0.60	0.08
		Organization	4		1.85	1.80	0.69	0.61	0.64	0.62	0.61	0.04	0.71	0.67	0.66	0.07
6	13307	Conventions	2	2095	1.36	1.42	0.69	0.65	0.66	0.64	0.64	0.04	0.72	0.68	0.68	0.09
		Evidence	4		1.54	1.52	0.68	0.65	0.67	0.62	0.62	0.04	0.72	0.64	0.64	0.03
		Organization	4		1.78	1.80	0.74	0.63	0.62	0.61	0.61	0.02	0.68	0.65	0.64	0.02
6	13308	Conventions	2	2097	1.41	1.50	0.67	0.62	0.62	0.54	0.54	0.06	0.74	0.68	0.67	0.13
		Evidence	4		1.46	1.37	0.62	0.57	0.69	0.57	0.57	0.03	0.71	0.59	0.58	0.15
		Organization	4		1.64	1.57	0.69	0.62	0.63	0.60	0.60	0.03	0.71	0.66	0.65	0.10
6	13309	Conventions	2	2093	1.39	1.48	0.65	0.56	0.68	0.58	0.58	0.06	0.76	0.68	0.67	0.15
		Evidence	4		1.69	1.60	0.73	0.67	0.65	0.59	0.59	0.02	0.72	0.71	0.70	0.13
		Organization	4		1.84	1.83	0.78	0.69	0.61	0.62	0.62	0.01	0.70	0.71	0.71	0.01
7	13400	Conventions	2	2082	1.35	1.45	0.66	0.63	0.70	0.67	0.67	0.02	0.74	0.70	0.69	0.14
		Evidence	4		1.84	1.83	0.61	0.53	0.66	0.60	0.60	0.07	0.77	0.65	0.65	0.03
		Organization	4		1.92	1.90	0.64	0.54	0.65	0.62	0.62	0.02	0.74	0.61	0.60	0.03
7	13401	Conventions	2	2084	1.65	1.72	0.56	0.49	0.79	0.62	0.62	0.04	0.80	0.64	0.63	0.14
		Evidence	4		1.86	1.87	0.58	0.50	0.72	0.63	0.63	0.01	0.79	0.64	0.63	0.03
		Organization	4		2.00	2.02	0.54	0.48	0.73	0.59	0.59	0.02	0.83	0.66	0.65	0.05
7	13402	Conventions	2	2088	1.49	1.55	0.63	0.62	0.69	0.60	0.60	0.03	0.75	0.67	0.67	0.10
		Evidence	4		1.83	1.87	0.51	0.43	0.73	0.59	0.59	0.04	0.88	0.74	0.72	0.08
		Organization	4		1.91	1.93	0.59	0.50	0.70	0.61	0.61	0.01	0.80	0.66	0.65	0.02

Exhibit D1.5-7: Summary of Human and Machine Scores for AIRCore Writing Prompts (continued)

Grade	ITS ID	Dimensions	Score Point	*N*	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	%z Exact	Pearson *r*	Weighted *κ**	SMD*	% Exact	Pearson *r*	Weighted *κ**	SMD*
7	13403	Conventions	2	2085	1.56	1.62	0.57	0.55	0.77	0.61	0.61	0.03	0.81	0.70	0.70	0.11
		Evidence	4		1.65	1.58	0.60	0.57	0.73	0.66	0.66	0.00	0.80	0.72	0.72	0.12
		Organization	4		1.75	1.75	0.64	0.56	0.68	0.61	0.61	0.02	0.78	0.69	0.69	0.00
7	13405	Conventions	2	2093	1.46	1.48	0.61	0.62	0.75	0.63	0.63	0.03	0.77	0.68	0.68	0.02
		Evidence	4		1.63	1.66	0.59	0.62	0.74	0.70	0.70	0.03	0.79	0.72	0.72	0.04
		Organization	4		1.83	1.80	0.62	0.56	0.73	0.68	0.68	0.03	0.79	0.70	0.69	0.05
7	13406	Conventions	2	2090	1.44	1.47	0.62	0.58	0.72	0.63	0.63	0.01	0.76	0.67	0.67	0.05
		Evidence	4		1.80	1.81	0.54	0.46	0.73	0.58	0.58	0.03	0.79	0.58	0.57	0.01
		Organization	4		1.92	1.89	0.56	0.49	0.71	0.59	0.59	0.01	0.79	0.62	0.62	0.06
8	13437	Conventions	2	2391	1.47	1.53	0.68	0.63	0.74	0.69	0.69	0.03	0.77	0.72	0.72	0.09
		Evidence	4		1.91	1.89	0.68	0.61	0.76	0.75	0.75	0.01	0.75	0.70	0.70	0.02
		Organization	4		2.05	2.01	0.77	0.69	0.73	0.75	0.75	0.01	0.75	0.75	0.74	0.06
8	13438	Conventions	2	2631	2.01	1.95	0.77	0.71	0.79	0.70	0.70	0.01	0.73	0.75	0.74	0.08
		Evidence	4		2.11	2.08	0.80	0.76	0.77	0.78	0.78	0.00	0.71	0.75	0.74	0.04
		Organization	4		1.55	1.57	0.63	0.59	0.73	0.76	0.76	0.01	0.79	0.72	0.72	0.03
8	13439	Conventions	2	2548	1.57	1.67	0.63	0.55	0.78	0.67	0.67	0.02	0.83	0.76	0.74	0.17
		Evidence	4		2.05	2.07	0.72	0.60	0.73	0.74	0.74	0.01	0.73	0.70	0.69	0.03
		Organization	4		2.15	2.16	0.81	0.73	0.71	0.74	0.74	0.01	0.69	0.72	0.71	0.01
8	13452	Conventions	2	2491	1.61	1.65	0.58	0.54	0.79	0.67	0.67	0.02	0.80	0.68	0.68	0.06
		Evidence	4		2.07	2.06	0.75	0.64	0.77	0.77	0.77	0.01	0.74	0.73	0.72	0.02
		Organization	4		2.20	2.18	0.76	0.67	0.74	0.75	0.75	0.01	0.75	0.76	0.76	0.03
8	13453	Conventions	2	2538	1.53	1.57	0.64	0.60	0.76	0.68	0.68	0.01	0.78	0.71	0.71	0.06
		Evidence	4		1.99	1.99	0.78	0.74	0.76	0.78	0.78	0.01	0.73	0.76	0.76	0.00
		Organization	4		2.14	2.12	0.79	0.73	0.75	0.79	0.79	0.02	0.74	0.77	0.77	0.03
8	13454	Conventions	2	2544	1.56	1.58	0.61	0.56	0.77	0.68	0.68	0.01	0.78	0.68	0.68	0.03
		Evidence	4		1.99	1.91	0.74	0.67	0.77	0.77	0.77	0.01	0.74	0.73	0.73	0.10
		Organization	4		2.04	2.06	0.76	0.74	0.75	0.77	0.77	0.01	0.74	0.76	0.76	0.02

Exhibit D1.5-7: Summary of Human and Machine Scores for AIRCore Writing Prompts (continued)

Grade	ITS ID	Dimensions	Score Point	*N*	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	%z Exact	Pearson *r*	Weighted *κ**	SMD*	% Exact	Pearson *r*	Weighted *κ**	SMD*
9	13554	Conventions	2	2751	1.61	1.68	0.59	0.55	0.81	0.71	0.71	0.02	0.80	0.69	0.68	0.13
		Evidence	4		1.89	1.92	0.62	0.53	0.82	0.76	0.76	0.01	0.79	0.68	0.67	0.04
		Organization	4		2.02	2.03	0.65	0.60	0.79	0.76	0.76	0.02	0.80	0.74	0.73	0.01
9	13555	Conventions	2	2853	1.58	1.68	0.63	0.57	0.81	0.74	0.74	0.02	0.81	0.76	0.74	0.17
		Evidence	4		1.88	1.90	0.66	0.59	0.81	0.77	0.77	0.02	0.79	0.72	0.72	0.03
		Organization	4		2.02	1.99	0.70	0.66	0.80	0.79	0.79	0.01	0.80	0.78	0.78	0.05
9	13556	Conventions	2	1469	1.66	1.72	0.57	0.56	0.80	0.69	0.69	0.02	0.81	0.71	0.71	0.11
		Evidence	4		1.86	1.90	0.64	0.60	0.79	0.76	0.76	0.01	0.78	0.71	0.71	0.08
		Organization	4		2.00	2.01	0.70	0.62	0.77	0.77	0.77	0.02	0.77	0.74	0.73	0.01
9	13557	Conventions	2	2815	1.54	1.58	0.65	0.60	0.79	0.73	0.73	0.00	0.78	0.71	0.70	0.06
		Evidence	4		1.82	1.85	0.57	0.53	0.83	0.78	0.78	0.01	0.83	0.72	0.72	0.05
		Organization	4		1.99	1.97	0.69	0.65	0.79	0.78	0.78	0.01	0.80	0.77	0.77	0.02
9	13565	Conventions	2	2869	1.52	1.56	0.62	0.61	0.80	0.75	0.75	0.02	0.78	0.71	0.71	0.06
		Evidence	4		1.92	1.92	0.67	0.60	0.81	0.80	0.80	0.03	0.79	0.74	0.73	0.01
		Organization	4		2.11	2.11	0.72	0.66	0.79	0.80	0.80	0.01	0.78	0.76	0.76	0.00
9	13566	Conventions	2	2852	1.54	1.59	0.63	0.60	0.81	0.76	0.76	0.03	0.81	0.76	0.76	0.09
		Evidence	4		1.93	1.93	0.62	0.54	0.84	0.80	0.80	0.00	0.82	0.74	0.73	0.01
		Organization	4		2.08	2.09	0.67	0.63	0.79	0.79	0.79	0.02	0.82	0.77	0.77	0.00
10	13635	Conventions	2	2436	1.61	1.65	0.55	0.53	0.71	0.60	0.60	0.02	0.77	0.61	0.61	0.07
		Evidence	4		2.04	2.08	0.77	0.71	0.69	0.73	0.73	0.01	0.75	0.76	0.76	0.05
		Organization	4		2.25	2.26	0.76	0.69	0.70	0.73	0.73	0.04	0.72	0.73	0.72	0.02
10	13636	Conventions	2	2344	1.69	1.78	0.49	0.45	0.72	0.58	0.57	0.01	0.83	0.63	0.62	0.19
		Evidence	4		1.99	1.96	0.74	0.66	0.74	0.73	0.73	0.00	0.76	0.76	0.76	0.04
		Organization	4		2.06	2.08	0.75	0.72	0.72	0.74	0.74	0.00	0.79	0.81	0.81	0.03
10	13637	Conventions	2	1314	1.58	1.65	0.60	0.53	0.70	0.59	0.59	0.02	0.76	0.62	0.61	0.11
		Evidence	4		1.89	1.88	0.70	0.66	0.76	0.72	0.72	0.01	0.77	0.75	0.75	0.02
		Organization	4		2.06	2.04	0.68	0.60	0.75	0.72	0.72	0.03	0.76	0.69	0.69	0.03

Exhibit D1.5-7: Summary of Human and Machine Scores for AIRCore Writing Prompts (continued)

Grade	ITS ID	Dimensions	Score Point	*N*	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	%z Exact	Pearson *r*	Weighted *κ**	SMD*	% Exact	Pearson *r*	Weighted *κ**	SMD*
10	13638	Conventions	2	2475	1.62	1.69	0.54	0.48	0.70	0.57	0.56	0.03	0.79	0.60	0.59	0.14
		Evidence	4		1.99	2.00	0.74	0.64	0.71	0.71	0.71	0.02	0.74	0.74	0.73	0.02
		Organization	4		2.12	2.14	0.76	0.70	0.69	0.71	0.71	0.01	0.77	0.79	0.78	0.03
10	13639	Conventions	2	2306	1.66	1.73	0.53	0.47	0.70	0.56	0.56	0.03	0.80	0.60	0.59	0.14
		Evidence	4		1.97	1.96	0.72	0.62	0.72	0.69	0.69	0.04	0.74	0.72	0.71	0.01
		Organization	4		2.10	2.14	0.71	0.69	0.70	0.70	0.70	0.03	0.78	0.75	0.75	0.05
10	13640	Conventions	2	2399	1.68	1.71	0.52	0.51	0.74	0.61	0.61	0.03	0.81	0.65	0.65	0.05
		Evidence	4		2.14	2.11	0.81	0.71	0.69	0.71	0.71	0.02	0.71	0.75	0.74	0.04
		Organization	4		2.26	2.29	0.79	0.72	0.67	0.71	0.71	0.04	0.73	0.76	0.76	0.04
11	13720	Conventions	2	2091	1.56	1.56	0.61	0.58	0.74	0.65	0.65	0.03	0.72	0.58	0.58	0.00
		Evidence	4		1.91	1.88	0.77	0.72	0.62	0.67	0.67	0.00	0.70	0.72	0.72	0.05
		Organization	4		2.15	2.15	0.77	0.68	0.67	0.71	0.71	0.01	0.73	0.73	0.72	0.00
11	13721	Conventions	2	2090	1.56	1.62	0.61	0.57	0.74	0.62	0.62	0.01	0.74	0.62	0.61	0.10
		Evidence	4		2.18	2.16	0.77	0.66	0.60	0.66	0.66	0.03	0.68	0.69	0.68	0.03
		Organization	4		2.36	2.35	0.71	0.64	0.66	0.67	0.66	0.03	0.76	0.74	0.73	0.01
11	13722	Conventions	2	2090	1.58	1.65	0.60	0.57	0.78	0.68	0.68	0.01	0.81	0.74	0.73	0.12
		Evidence	4		2.18	2.20	0.80	0.70	0.62	0.67	0.67	0.02	0.70	0.74	0.73	0.03
		Organization	4		2.38	2.36	0.73	0.68	0.66	0.70	0.70	0.04	0.76	0.75	0.75	0.03
11	13723	Conventions	2	2095	1.60	1.63	0.59	0.57	0.75	0.61	0.61	0.01	0.77	0.65	0.65	0.05
		Evidence	4		2.24	2.24	0.83	0.74	0.62	0.70	0.70	0.02	0.70	0.74	0.73	0.00
		Organization	4		2.47	2.47	0.74	0.68	0.64	0.69	0.69	0.00	0.74	0.74	0.73	0.01
11	13724	Conventions	2	2089	1.60	1.61	0.57	0.54	0.73	0.61	0.61	0.03	0.79	0.66	0.66	0.03
		Evidence	4		2.24	2.27	0.79	0.74	0.63	0.68	0.68	0.00	0.74	0.76	0.75	0.04
		Organization	4		2.28	2.32	0.75	0.68	0.64	0.66	0.66	0.01	0.75	0.76	0.75	0.06
11	13725	Conventions	2	2085	1.45	1.50	0.67	0.63	0.71	0.64	0.64	0.04	0.78	0.74	0.73	0.08
		Evidence	4		2.21	2.26	0.83	0.78	0.64	0.72	0.72	0.02	0.74	0.79	0.79	0.05
		Organization	4		2.36	2.34	0.81	0.71	0.66	0.73	0.73	0.00	0.74	0.78	0.77	0.02

Note. Weighted K = Quadratic weighted kappa; SMD = Standardized Mean Difference

*For asterisked items, no 4-point responses were identified in the training set, so at present statistical models for these items can only assign up to three points.

presentation and scoring of all items is completed well before the assessment system is launched. Nevertheless, it is important to ensure a final set of system checks before the testing window opens.

Online Test Decks. Online tests are different than scanned tests because the items are represented as data, and the programming is constant across items. This minimizes the risk that a bug will exist for a single item or for an item on a single form.

The test decks are implemented on the production TDS so that final verification is performed on the operational test platform. As with test decks used to calibrate document imaging systems, we implement a series of test administrations designed to test all possible responses to selected-response items, as well as a range of possible responses to constructed-response items, to ensure that student responses are captured accurately and that item responses are scored correctly. Although AIR is solely responsible for identifying any issues related to item response capture, the Department may also wish to implement its own online test decks during the UAT window. AIR will provide the Department with data files containing all Department-entered UAT test records to verify that captured test records match data entry records implemented by the Department.

Simulated Test Administrations. While we do verify the scoring of item responses and test scores and performance levels using the UAT input records, test deck data are limited with respect to scoring of test records. This is especially true with pattern scoring, where there are thousands of possible test scores. Therefore, we verify the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the state. The ability of each simulated student is used to generate a sequence of item response scores consistent with the underlying ability. These simulations provide a rigorous test of the adaptive algorithm for adaptively administered tests but also provide a check of the full range of item and test scores in fixed-form tests. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the ORS, we merge item response data with the demographic information taken from previous-year assessment data, or if current year enrollment data is available by the time simulated data files are created, we can verify online reporting using current-year testing information. By populating the simulated data files with real school information, it is possible to verify that special school types and special districts are being handled properly in the reporting system.

Accommodated Test Administrations. Test accommodations, including Braille, large print, text-to-speech (TTS), American Sign Language (ASL), and a host of other available accommodations such as foreign language versions described in Topic 1.4 Accessibility and Fairness, are available for the online tests. In the online system, accommodations are provided as tags to items in the item bank. Because accommodations are attributes of items, rather than adaptations of specific test forms, the accommodations travel with the items and eliminate the need to produce special accommodated test forms. Moreover, because ITS uses the same item display software used by the TDS, as soon as items are tagged with an accommodation, the Department and AIR can verify that the accommodated version of the test item is being displayed correctly. This allows accommodated items to be verified well before the beginning of the test administration window.

In addition, because data capture is identical for accommodated and non-accommodated items, a separate process for verifying correct capture of accommodated responses is not necessary. Test cases for accommodated test administrations are designed into the online test deck and can be incorporated into the Department's UAT. For accommodations provided via paper test forms, test administrators will use the data entry interface (DEI) to submit student responses. The DEI provides a user-friendly environment for data entry that mirrors the student's test booklet to reduce the likelihood of data entry errors. The functionality of the DEI is also verified using the online test decks.

Specifications for generating simulated data files, including the test deck designs for both paper-pencil and online assessments, and the design of simulation runs to verify scoring and reporting, will be included in the scoring and reporting specifications document submitted to the Department each year. UAT processes, procedures, and support is described in detail in Topic 37 User Acceptance Testing.

D1.6 Reporting

Introduction

The RFP envisions a reporting system that supports New Hampshire teachers, students, and administrators as they transition to the new summative and interim assessments. We have successfully delivered online reports to New Hampshire educators and parents for the past three years. We have seen that it is critical for teachers to understand the results of the assessments to make instructional decisions and pinpoint strengths and weaknesses in their classes. We have also learned that teachers need an easy way to integrate their understanding of the results from multiple assessments, specifically the summative and interim assessments. Providing reporting systems where relevant enables actionable data to arrive quickly at the school and support instruction and success. As a result, AIR is uniquely positioned to meet the reporting objectives outlined in the RFP and offers New Hampshire the following benefits:

- We have successfully reported the performance of more than 76,000 student assessment results yearly on 163 different New Hampshire Smarter Balanced ELA and mathematics summative, interim comprehensive, and interim block assessments to New Hampshire educators, administrators, and parents for the past three years.
- New Hampshire teachers and administrators are already familiar with our reporting systems, which minimizes the need for intensive training on a new system.
- AIR understands that results from the first operational administration will be reported following standard setting, but for the subsequent operational administrations, our test delivery system (TDS) and online reporting system (ORS) provide real-time scores as students submit their tests, even for most students with writing and extended-response options in the assessment. Only a small percentage of the students taking writing assessments will have their scores sent for verification by human readers.
- Our ORS provides longitudinal data of roster, teacher, school, district, and state for the purpose of measuring growth and setting academic goals. Please note that the students must have unique identification across years in order to report longitudinal data at any level of aggregation.
- AIRWays, our benchmark reporting system, provides teachers with better ways of tracking student progress throughout the year. Teachers and other users can now see the items from the non-secure, fixed-form assessments in addition to actual student responses, allowing them to pinpoint students' strengths and weaknesses and gaps in student understanding.
- All reporting systems provide secure reporting by restricting access to authorized users.
- Our paper reports offer unparalleled customizability, allowing any text, shape, or data display on any part of the page. AIR successfully delivers approximately 25 million pages of customized, full-color score reports each spring to families in twelve states, and we mail half-a-million paper reports directly to student homes in Delaware.

Should AIR be awarded this contract, AIR commits to do the following:

- Provide ORS reporting of Summative and Interim Comprehensive results at the individual student, class, teacher, school, district, and state levels.
- Provide AIRWays reporting of interim block results at the individual student, class, teacher, school, and district levels.
- Produce and deliver hard copy reports of summative results at the individual student level.
- Produce and deliver state data files for NH DOE to import into NH DOE reporting systems.

We explain the features of our online reporting systems, hard copy student reports, and data files in greater detail below and have divided the discussion that follows into eight sections:

1. Proposed Reporting Measures at individual and aggregate levels (Topic 21)
2. An overview of our Online Reporting System and process for designing mockups and specifications (Topic 22)
3. Summary of the reports for summative and interim assessments (Topic 22.1)
4. Description of the process for producing and printing hard copy student reports (Topic 22.2)
5. Reports available in ORS, including screen captures and a summary of AIR's online features and capabilities (Topic 22.3)
6. Summary of our capabilities for reporting interim block scores through AIRWays (Topic 22.3)
7. Description of the process for producing data files (Topic 23)
8. Description of the process for producing interpretive materials

Topic 21 Assessment Scores

We will work with NH DOE to configure the online reports and identify the measures the Department would like to provide to meet the objectives outlined in the RFP. We are committed to collaborating with NH DOE to continuously develop better measures and more meaningful ways of reporting. For example, in 2016 we developed new item-level reports for the interim tests, new target-level reports for the summative tests, and writing dimension measures for both the Smarter Balanced interim and summative tests. At a minimum, AIR commits to reporting the following measures:

- Overall subject area scores and proficiency level classifications for students.
- Average subject area scores and proficiency level classification rates for classrooms, schools, and districts.
- Subscale performance level classifications for students, including for writing in ELA (Writing includes both essay dimension scores and language items.).
- Subscale performance level classification rates for classrooms, schools, and districts.
- Benchmark strengths and weaknesses at the class, school, and district level.
- Rubric dimension scores for student essay responses.
- Individual items and students' responses to each item for non-secure, block interim assessments.

AIR will provide subscale or reporting category scores and standard errors for all aggregate reports. However, we do not recommend providing subscale scores for individual student reports since those scores are highly unreliable and thus volatile. Moreover, simply reporting standard errors for subscale scores does not solve the reporting problem, because the vast majority of intended users will not understand how to interpret test scores in light of the standard errors and tend to ignore them in any event, since they infer that any scores reported must be meaningful. AIR recommends subscale performance reporting that explicitly takes into account the standard error of the reporting category score in classifying student achievement. In this approach, the student's reporting category score and standard error are evaluated with respect to the performance standard. If the confidence band defined by plus and minus one standard error about the subscale score is above or below the proficient level performance standard, the student is classified as scoring above or below proficient within the reporting category. Where the error band encompasses the proficient standard, the performance is classified as uncertain or near the standard. AIR's reporting system can be configured to show the relative location of the student's score with respect to the standard to provide additional context for interpretation of reporting category level performance.

The adaptive test design for ELA and mathematics, as well as the matrix design for science, allow for more in-depth reporting at aggregate levels, so that in addition to reporting overall and reporting category performance, we propose to report on benchmark level performance for aggregated units. AIR proposes to report aggregate benchmark scores in two ways: (1) benchmark scores relative to a student's overall estimated ability, and (2) benchmark scores relative to the proficiency standard (Level 3 cut). The benchmark-level reports are not possible to produce for a fixed-form test because the number of items included per benchmark is too few to produce a reliable score at the benchmark level. A typical fixed-form test includes only one or two items per benchmark. Even when aggregated, these data reflect the benchmark only narrowly because they reflect only one or two ways of measuring the benchmark. However, an adaptive test or matrix-designed test offers a tremendous opportunity for benchmark-level data at the class, school, and district area level. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given benchmark. A benchmark score is an aggregate of the differences in student overall proficiency and the differences in the difficulty of the items measuring a benchmark in a class, school, or district area. Benchmark scores are computed for attempted tests based on the responded items. Benchmark scores are computed within each reporting category.

Proposed Reporting Measures at Individual Levels

AIR commits to meeting all requirements regarding student reports and will report all of the student-level measures outlined in the RFP. Although reports for each grade and subject may have differing templates, formats, and content, we understand that the data for student-level reports will include, but not be limited to, the following measures:

- Scale score
- Probable range of scores (standard error of measurement)
- Comparisons of student performance with school, district, and state
- Reporting category performance category (above, at/near, below standard)
- Reporting category level reporting for writing in ELA
- Rubric scores assigned for each writing dimension in the student essays
- Actual student responses on interim block assessment available through AIRWays

All student reports will be available in English. Individual student reports can also be translated into other languages with additional cost. Please see Topic 22.3 School, District, and State Reports for screen captures of the online student reports and see Exhibit D1.6-2 for screen captures of the proposed paper report features. Students' responses to each item are provided through AIRWays. For more information, please see Section D1.8 Reporting Portal.

Proposed Reporting Measures at Aggregate School, District, and State Levels

As required in the RFP, AIR will provide online class-, teacher-, school-, district-, and state-level reports that include information to help educators and administrators evaluate the efficacy of school-level academic programs or curricular decisions or evaluate large-scale programmatic efficacy through ORS and AIRWays. Our district-, school-, and class-level strengths and weaknesses can help both educators and school and district administrators pinpoint class-level and programmatic weaknesses. For example, this data can indicate to teachers which reporting categories the class has performed better on relative to the class performance on the test as a whole. These are the reporting categories that a teacher may want to spend more instructional time on. Similarly, a principal can analyze the strengths and weaknesses across classes. If the strengths and weaknesses for each grade 4 ELA class are the same, this could suggest a school-wide instructional approach is not working or that the instructional materials the school is using to teach these benchmarks are not adequate. In contrast, if different classes have different strengths and weaknesses, this likely indicates specific teachers need professional development support in their own respective areas.

To help educators and administrators evaluate the effectiveness of their programs, we will provide online educator reports for both the summative and interim assessments that include the following information:

- Overall proficiency rates by grade and subject for classes, teachers, schools, districts, and state
- Overall average scale score and standard error of mean
- Percentage of students performing in each performance level
- Reporting category-level percentage of students performing in each achievement category
- Reporting category-level average scale score and standard error of mean
- Benchmark level strengths and weaknesses indicators
- Aggregate level data for the teachers' current and past students
- Longitudinal trend across time

All aggregate reports are dynamic. Aggregate reports can be accessed for current and past students and can be disaggregated by the demographic sub-groups that NH DOE requires. All reports will be designed with the needs of stakeholders in mind. Educators will be able to conduct analysis of achievement and growth gaps by using sub-group filters to view disaggregated data. For example, district administrators can see growth gaps by sex, race/ethnicity, economic disadvantage, special education, and English language learner (ELL). District users can use the trend reports in the ORS to track the changes in achievement gaps over time. For example, a district can see its students' performance over the last three years on the summative to identify growth gaps.

Topic 22 Reports

Overview of the Online Reporting System

AIR proposes to provide New Hampshire teachers and administrators with an ORS that is familiar to the educators. The system is intuitive and user-friendly and provides immediate access to assessment data. Educators can use the system to find and interpret data for multiple assessments.

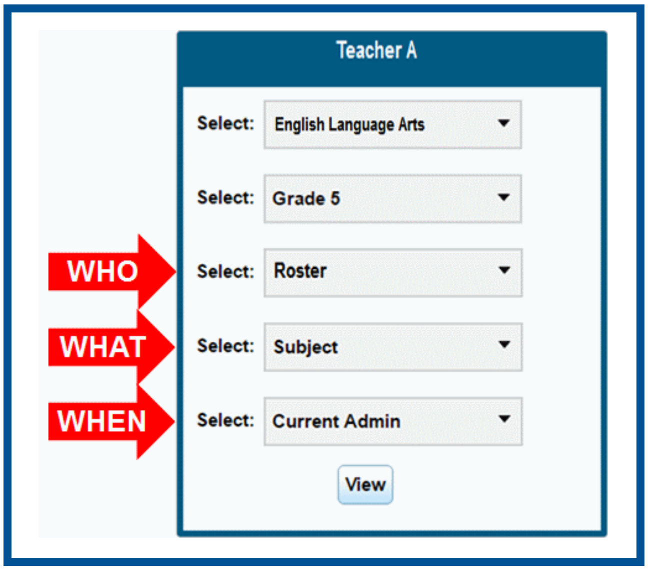
The system, which was built using feedback from stakeholders, is designed to support educators as they evaluate the needs of their students and reflect on their own curricula and practice. Navigation in the system mirrors the instructional decision-making process. The user can intuitively navigate in any of the three dimensions inherent in the data, and these three dimensions parallel the three kinds of questions that the data can help the user answer:

1. Who? The data can be displayed at levels of aggregation anywhere from the individual level for a specific student up to the entire state. Demographic breakdowns are immediately available at any level of aggregation.
2. What? The subject-area data can be broken down into finer or coarser chunks of content. Navigating this dimension allows the user to travel from subject to reporting category to benchmark and back.
3. When? When data are available over time, the system allows the user to view a data trend over time or toggle to a fixed point in time.

The interface design encourages users to think about the substantive, educational questions to which they need answers and access information from that perspective. In the following paragraphs we discuss each of these selections in more detail in context of the exploration menu. The menu demonstrated in Exhibit D1.6-1 is designed to help users analyze data by allowing them to move between data for different groups of students, subjects, and periods of time. When a user opens the exploration menu, he or she will see five drop-down menus. These selections guide the user to answering questions that will lead him or her to the most relevant data. The first two selections, Test and Grade, are straightforward. Next, the user must specify whether to view data for a sub-group of students in his or her class (called a *roster*) or for all of his or her students. Then the user will specify whether to see the data by overall subject, reporting category, or benchmark. Finally the user must specify whether to view data for the tests his or her students have taken this year or review the students' performance over the years.

By making these selections on the exploration menu drop-down menus, users can easily move to higher-level data or more granular data. In comparison, principals and district-level personnel will have different options in some of the drop-down menus. A school principal will be able to view data for different teachers by selecting Teacher from the third drop-down list. A district-level user will be able to view data for different schools in the district by selecting School from the third drop-down list.

Exhibit D1.6-1: Exploration Menu



Steps for Producing Online Reports

Our proposed plan for producing online reports involves three steps:

1. Working with NH DOE to create the look and feel of the ORS and online reports
2. Gathering requirements and specifications with NH DOE to configure the ORS
3. Conducting rigorous quality assurance checks on all scores before they are released

The first step in producing real-time online reports for teachers, schools, and administrators involves the initial formatting of the ORS to finalize the look and feel of all online reports. AIR will work with NH DOE to select a skin for the reporting system composed of colors, a logo, headers, and footers that will give each report and data display a unified style.

Next, we will work with NH DOE to gather requirements and configure the system. To ensure that the online reports meet stakeholder needs, AIR will also work closely with NH DOE to document requirements and configure the online reports. We will begin by drafting a reporting elements document that describes stakeholders' reporting needs and the available measures that can be used to meet those needs. When the reporting elements document is approved, we will work with NH DOE to complete the online reporting specifications, which define everything in the system, including table layouts, user role access, definitions, reporting rules, and content hierarchies. To facilitate this process, we propose conducting a working session with NH DOE's staff to capture requirements for each part of the reporting system and complete all specifications.

Each school year, AIR will provide mockups of online report pages for NH DOE staff review. Once NH DOE reviews and approves the mockups and specifications, AIR programmers will use these documents to configure the ORS and conduct numerous quality assurance (QA) checks to ensure that the system is functioning as required.

Quality Assurance

The final step involves conducting QA before scores are released. Scores for online assessments are assigned by automated systems in real time. For machine-scored portions of assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field testing. The review process locks down the item and rubric when the item is approved for web display (web approval). During operational testing, actual item responses are compared to the expected item responses, given the item response theory (IRT) parameters, which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

After scores have passed the QA checks and are uploaded to the Database of Record (DoR), they are passed to the ORS, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all QA system validation checks. All of the above processes take milliseconds to complete; within less than one second after handscores are received by AIR and pass the QA validation checks, the score results will be available in the ORS.

Topic 22.1 Types of Reports

AIR commits to providing online reports for all summative and interim assessments outlined in the RFP. Access to timely data is essential for a good assessment system. One of the most powerful benefits of our proposal for an adaptive assessment system with automated essay scoring is that we can easily support immediate reporting of assessment results for most students. Scores for all tests that successfully pass the scoring checks will be populated in the ORS and AIRWays as students submit them. This allows educators and administrators to view results on a real-time basis throughout the year and use these results as factors in important programmatic and instructional decisions. For summative assessments, approximately 15% of students will have writing responses automatically forwarded for verification by human readers, so test results will not be reported for those students until receipt of the handscoring results. AIR commits to reporting test results for records requiring verification reads within 10 business days. Note that results from the first operational administration will be reported following standard setting.

AIR’s ORS and AIRWays allow users to download on-demand standard and custom PDFs, spreadsheets, and data files for each assessment. They also provide individual student results and class, teacher, school, district, and state aggregate results as appropriate based on Family Educational Rights and Privacy Act (FERPA. The ORS allows schools and districts to batch print PDFs of hard copy student reports for delivery to families.

Results for all assessments except the interim block assessment will be reported in ORS, as shown:

- Grades 3−8 ELA summative assessments
- Grades 3−8 mathematics summative assessments
- Grades 5, 8, and 11 science summative assessments
- Grades 3–8 ELA interim comprehensive assessments (if NH DOE decides to go with the interim comprehensive test model)
- Grades 3–8 mathematics interim comprehensive assessments (if the Department decides to go with the interim comprehensive test model)
- Grades 5, 8, and 11 science interim comprehensive assessments (if the Department decides to go with the interim comprehensive test model)

We propose to report results for interim block assessments through AIRWays, because AIRWays provides individual items and students' response to each item for non-secure, fixed-form assessments.

AIR proposes to provide additional reporting capabilities for the interim block assessments in the AIRWays system. In order to guide teachers to pinpoint gaps in conceptual understanding and areas for further instruction based on the benchmark assessments, AIR proposes that the following measures be reported:

- Item-level data that provide teachers with detailed information about which items a student answered correctly or incorrectly
- Actual student responses to items on the benchmark assessment to help teachers pinpoint gaps in conceptual understanding
- The content area or standards to which the item is aligned
- The scoring rubric for each item so teachers can better understand how items are scored
- Strengths and weaknesses at the class, school, and district levels by identifying the five items with the highest and lowest scores
- Easy-to-read reports that allow teachers to quickly see all of the benchmark scores for a student or a class in a single report

All reports in AIRWays can be printed for classroom use or downloaded into a CSV file so schools and districts can load files into their own systems.

Topic 22.2 Student Reports

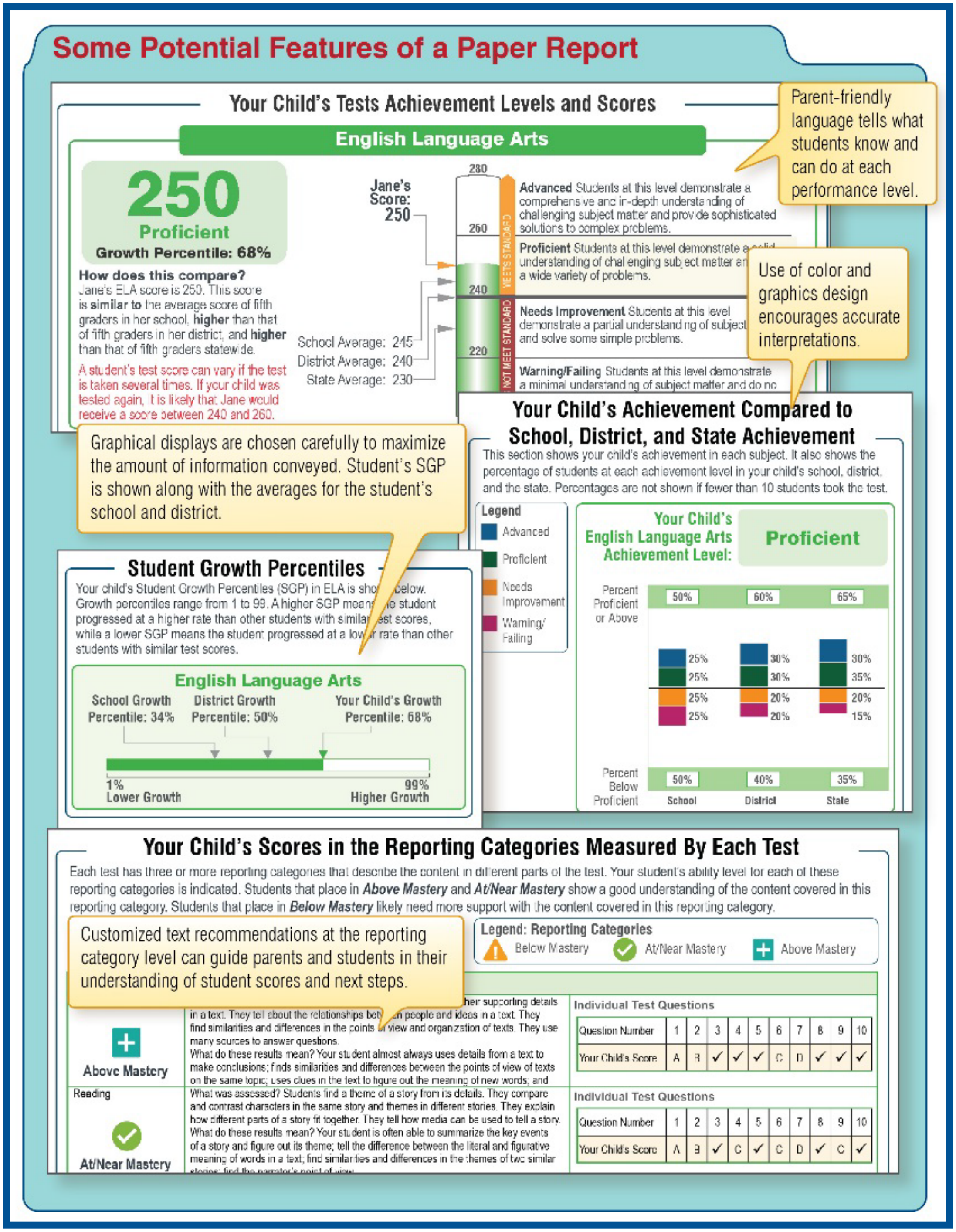
In addition to online student reports, we propose to work with NH DOE to design and deliver hard copy student reports following the close of the testing window. AIR's paper reporting capabilities allow the Department virtually infinite flexibility in designing new paper reports that meet the needs of students, teachers, and administrators. Our plan for assisting the Department in developing hard copy student reports offers several benefits:

- AIR successfully delivers approximately 25 million pages of customized, color score reports each spring to families in twelve states and we mail half-a-million paper reports directly to student homes in Delaware.
- Our hard copy student reports offer unparalleled customizability allowing any text, shape, or data display on any part of the page. This flexibility enables our designers and psychometricians to experiment with different ways to visualize the data and ultimately, to propose designs that can be accurately understood by parents.
- Our reports are grounded in cognitive and focus group research with end-users to help ensure parents who receive our reports understand the information as intended.
- Most importantly, an effective reporting system helps improve education by changing the behavior of stakeholders. AIR's hard copy student reports are designed to encourage meaningful change by helping parents draw accurate inferences about what their child knows and can do and by providing rich, text recommendations giving parents specific guidance on how to help their child improve.

If awarded this contract, AIR will work with the Department to design and develop customized, full-color reports, and revise them to eliminate sources of misunderstanding. The following pages highlight some of the design elements that we have developed for past score reporting endeavors. We have also included sample score reports for review in Appendix P.

Exhibit D1.6-2 highlights some of the features of paper reports that we have developed for our clients. We include these as samples as we expect to work closely with the Department to determine the design and format of all mockups.

Exhibit D1.6-2: Features of Paper Reports



To assist the Department in providing the hard copy student reports, we propose to design and develop 100% customizable, enhanced color paper reports for NH families. AIR's paper capabilities are the most sophisticated in the industry, allowing the Department virtually infinite flexibility in designing new paper reports including the ability to

- place any text, graphic, or color on any part of the page;
- develop customized text and graphics for different students based on their achievement level or for different tests; and
- test proposed designs with NH parents in two focus groups and revise the final design to help ensure that all text and data elements are accurately understood by parents.

Below we describe the five-step development process for our paper reports which includes

- designing the mockups;
- completing specifications;
- developing data-driven text recommendations;
- conducting quality assurance; and
- printing, packaging, and distributing the reports.

Every stage, particularly the research and design phase, requires close collaboration with the Department. All AIR reporting processes adhere to industry best practices and to the *Standards for Educational and Psychological Testing.*

Designing Mockups. The initial design phase for paper reports begins with the consideration of the inferences, actions, and choices that the data should encourage and then works backward to the inferences that will lead to these actions, the data displays that will guide readers to the appropriate inferences, and finally to the data and measures to be displayed.

We use our Report Elements Matrix to guide the discussion and identify the desired behaviors the Department would like to encourage. Our reporting team will work closely with Department staff to understand the broader goals and policy considerations around the new assessments and reports and identify measures and data displays that meet these objectives. Once the report elements matrix is complete, our graphic designers provide different design options. The design and data displays will be systematically evaluated and reviewed by AIR and the Department. We also propose to conduct two focus groups with New Hampshire parents to review the proposed designs. After the focus groups, AIR will work with the Department to select the final design and data displays to be included in the reports.

Complete Specifications. Once mockups are approved by the Department, the next step is to finalize the Reporting Specifications document which specifies rules for aggregating and merging data, handling special populations and discrepant records. All the data displayed on the paper reports will conform to the specifications agreed upon between the Department and AIR psychometricians and statisticians.

Data-driven text Recommendations. Once the Department approves the final report design, AIR content experts draft the report text, including performance level descriptions, interpretive texts for graphics, and customized recommendations triggered by test data. All text will be vetted with the Department.

Quality Assurance. After mockups and report specifications have been reviewed and approved by the Department, AIR will test the programs used to produce them. Each program is extensively tested on test-decks (control district) and real data from past administrations. The final programs are reviewed by two senior statisticians and one senior member of the Software team to ensure agreed-upon procedures are accurately implemented.

Once we receive final data, AIR's Score Reporting team will review proofs that contain actual data and verify that the data follows approved specifications and appears as it should on the reports. In addition, we will compare data independently calculated by AIR psychometricians with data on the reports. A large sample of each type of report will be reviewed by several AIR staff to make sure that all data are correctly placed on reports. This rigorous review typically is conducted over several days and takes place in a secure location in the AIR building.

Packaging and Distribution. Paper reports for each testing program will be distributed to schools according to the Department's specifications. Once all tests are scored, AIR will produce data files and print and ship reports for all students who tested during the school year.

Digital Version of the Student Report. Electronic version of the student report will also be available in the Online Reporting System, which is discussed further in the next section.

Topic 22.3 School, District, and State Reports

Reports Available in Online Reporting System and Summary of AIR's Online Features and Capabilities

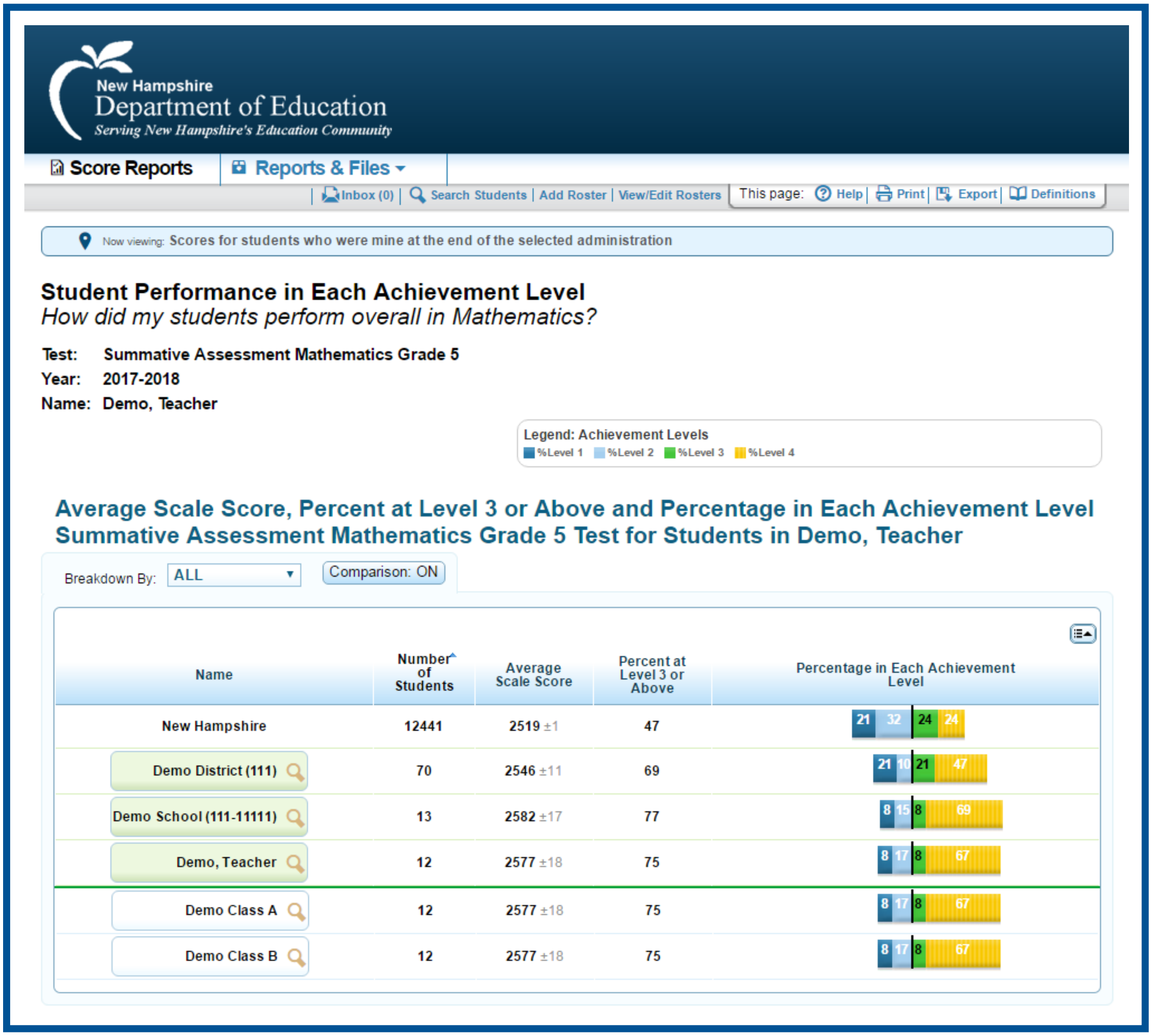
In this section, we summarize some current assessment reports and include screen captures, to demonstrate the kinds of data that educators and administrators may access in the ORS and how these displays can be customized.

It is important to note that these reports are just samples and can change based on NH DOE policy decisions regarding test design.

Scope of Aggregate Reports. We show a single example of each report type, but all aggregate reports are available at every level of aggregation. The levels of aggregation in New Hampshire are currently district, school, teacher, and roster. A roster is any grouping of students. It can be a classroom, the ELL students in a classroom, the students in the band or on the football team, or any other grouping that helps educators differentiate instruction as needed.

Subject Summary Reports. Aggregated subject reports may show average performance for the state, districts, schools, teachers, and classes. Bar-chart displays show the distribution of students' performance levels. Columns can be sorted on any table in the system to allow for easy comparison, and groups can be disaggregated to report by demographic sub-group (e.g., gender) or test condition event (e.g., accommodation). Exhibit D1.2-3 shows a sample subject summary report.

Exhibit D1.2-3: Subject Summary Report for a Teacher's Class


New Hampshire Department of Education
Serving New Hampshire's Education Community

Score Reports | Reports & Files

Inbox (0) | Search Students | Add Roster | View/Edit Rosters | This page: Help | Print | Export | Definitions

Now viewing: Scores for students who were mine at the end of the selected administration

Student Performance in Each Achievement Level
How did my students perform overall in Mathematics?

Test: Summative Assessment Mathematics Grade 5
Year: 2017-2018
Name: Demo, Teacher

Legend: Achievement Levels
%Level 1 %Level 2 %Level 3 %Level 4

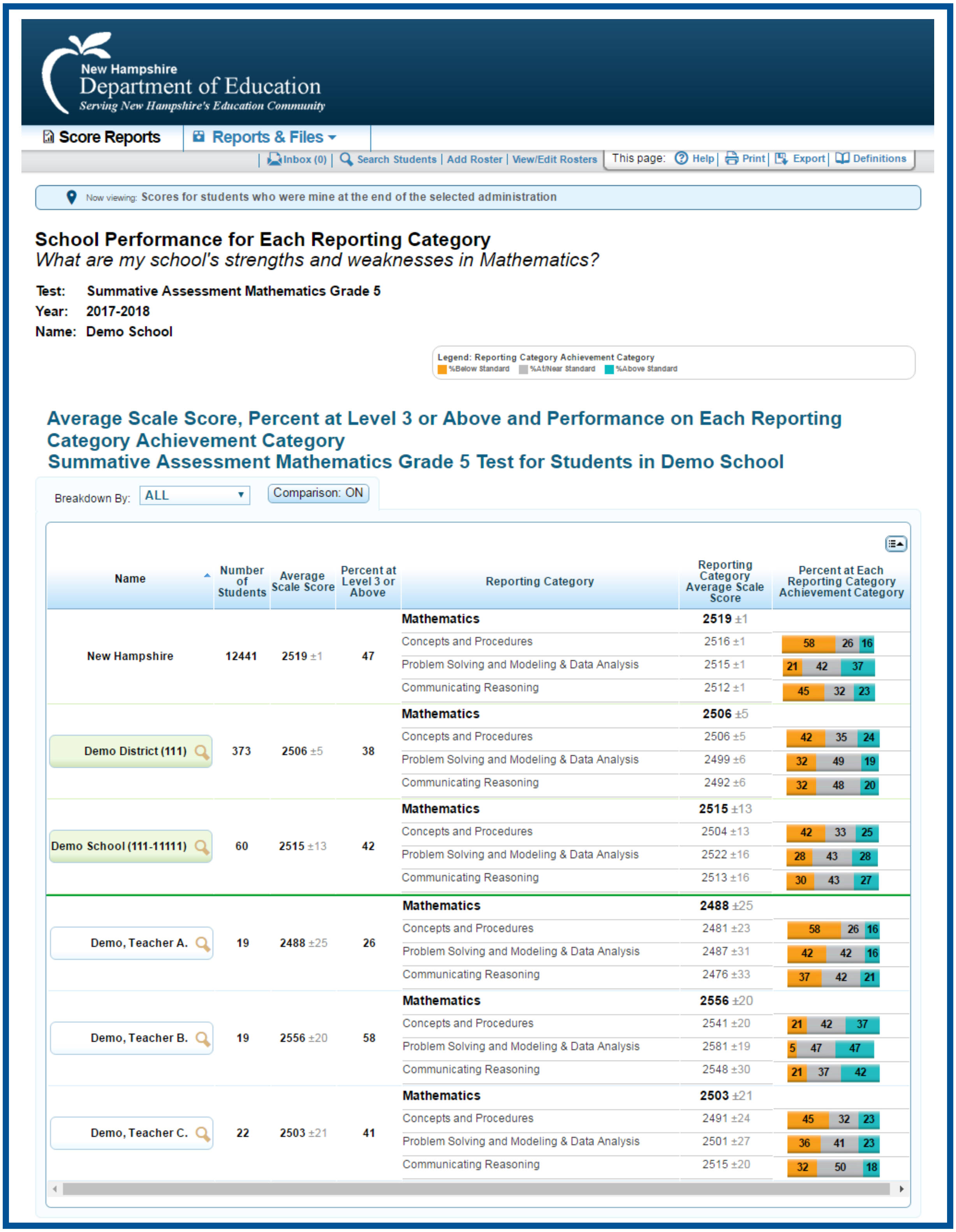
Average Scale Score, Percent at Level 3 or Above and Percentage in Each Achievement Level Summative Assessment Mathematics Grade 5 Test for Students in Demo, Teacher

Breakdown By: ALL Comparison: ON

Name	Number of Students	Average Scale Score	Percent at Level 3 or Above	Percentage in Each Achievement Level
New Hampshire	12441	2519 ±1	47	21 32 24 24
Demo District (111)	70	2546 ±11	69	21 10 21 47
Demo School (111-11111)	13	2582 ±17	77	8 15 8 69
Demo, Teacher	12	2577 ±18	75	8 17 8 67
Demo Class A	12	2577 ±18	75	8 17 8 67
Demo Class B	12	2577 ±18	75	8 17 8 67

Reporting Category Reports. Aggregated reporting category reports follow the layout of the subject summary reports, displaying performance data for the state, districts, schools, teachers, and classes. Exhibit D1.2-4 shows a sample school summary report for grade 5 ELA reporting categories. In addition to reporting the overall average scale score for the test, we report the average reporting category scale scores and percent at each reporting category achievement category.

Exhibit D1.2-4: Reporting Category Summary Report for a School


New Hampshire Department of Education
Serving New Hampshire's Education Community

Score Reports | Reports & Files

Inbox (0) | Search Students | Add Roster | View/Edit Rosters | This page: Help | Print | Export | Definitions

Now viewing: Scores for students who were mine at the end of the selected administration

School Performance for Each Reporting Category
What are my school's strengths and weaknesses in Mathematics?

Test: Summative Assessment Mathematics Grade 5
Year: 2017-2018
Name: Demo School

Legend: Reporting Category Achievement Category
%Below Standard | %At/Near Standard | %Above Standard

Average Scale Score, Percent at Level 3 or Above and Performance on Each Reporting Category Achievement Category
Summative Assessment Mathematics Grade 5 Test for Students in Demo School

Breakdown By: ALL | Comparison: ON

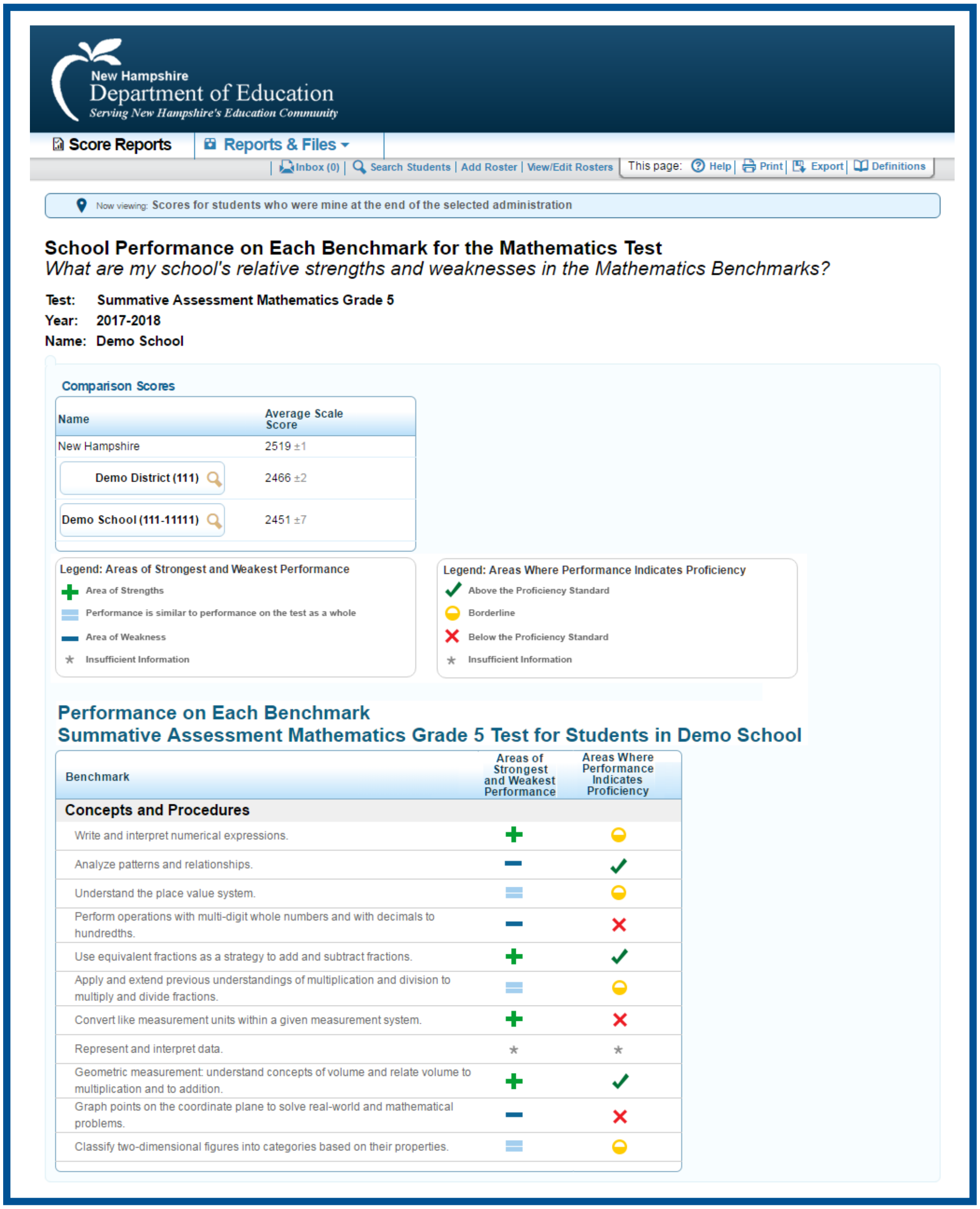
Name	Number of Students	Average Scale Score	Percent at Level 3 or Above	Reporting Category	Reporting Category Average Scale Score	Percent at Each Reporting Category Achievement Category
New Hampshire	12441	2519 ±1	47	**Mathematics**	**2519** ±1	
				Concepts and Procedures	2516 ±1	58 26 16
				Problem Solving and Modeling & Data Analysis	2515 ±1	21 42 37
				Communicating Reasoning	2512 ±1	45 32 23
Demo District (111)	373	2506 ±5	38	**Mathematics**	**2506** ±5	
				Concepts and Procedures	2506 ±5	42 35 24
				Problem Solving and Modeling & Data Analysis	2499 ±6	32 49 19
				Communicating Reasoning	2492 ±6	32 48 20
Demo School (111-11111)	60	2515 ±13	42	**Mathematics**	**2515** ±13	
				Concepts and Procedures	2504 ±13	42 33 25
				Problem Solving and Modeling & Data Analysis	2522 ±16	28 43 28
				Communicating Reasoning	2513 ±16	30 43 27
Demo, Teacher A.	19	2488 ±25	26	**Mathematics**	**2488** ±25	
				Concepts and Procedures	2481 ±23	58 26 16
				Problem Solving and Modeling & Data Analysis	2487 ±31	42 42 16
				Communicating Reasoning	2476 ±33	37 42 21
Demo, Teacher B.	19	2556 ±20	58	**Mathematics**	**2556** ±20	
				Concepts and Procedures	2541 ±20	21 42 37
				Problem Solving and Modeling & Data Analysis	2581 ±19	5 47 47
				Communicating Reasoning	2548 ±30	21 37 42
Demo, Teacher C.	22	2503 ±21	41	**Mathematics**	**2503** ±21	
				Concepts and Procedures	2491 ±24	45 32 23
				Problem Solving and Modeling & Data Analysis	2501 ±27	36 41 23
				Communicating Reasoning	2515 ±20	32 50 18

Benchmark-Level Reports. AIR leverages the adaptive environment to provide highly reliable, standard-level reports of strengths and weaknesses at the class, school, or district level. Educators want very precise information about the knowledge or skills that students have or have not mastered. In the benchmark-level report, strengths and weaknesses are reported for groups of students based on whether there is a statistically significant difference between that group's performance on each benchmark and its performance on a test as a whole. Another strengths and weaknesses indicator shows whether the performance is above, below, or near the performance level cut score.

This sort of fine-grained information is actionable; it tells educators what to teach and lets them know where they are succeeding. Exhibit D1.2-5 shows a sample benchmark-level report for a school where relative strengths and weaknesses are shown for each benchmark.

We note that this information is available under any design, but shorter tests will lead to more standards indicating insufficient information.

Exhibit D1.2-5: Benchmark Summary Report for a School



New Hampshire Department of Education
Serving New Hampshire's Education Community

Score Reports | Reports & Files

Inbox (0) | Search Students | Add Roster | View/Edit Rosters | This page: Help | Print | Export | Definitions

Now viewing: Scores for students who were mine at the end of the selected administration

School Performance on Each Benchmark for the Mathematics Test
What are my school's relative strengths and weaknesses in the Mathematics Benchmarks?

Test: Summative Assessment Mathematics Grade 5
Year: 2017-2018
Name: Demo School

Comparison Scores

Name	Average Scale Score
New Hampshire	2519 ±1
Demo District (111)	2466 ±2
Demo School (111-11111)	2451 ±7

Legend: Areas of Strongest and Weakest Performance
- ✚ Area of Strengths
- ═ Performance is similar to performance on the test as a whole
- ━ Area of Weakness
- ★ Insufficient Information

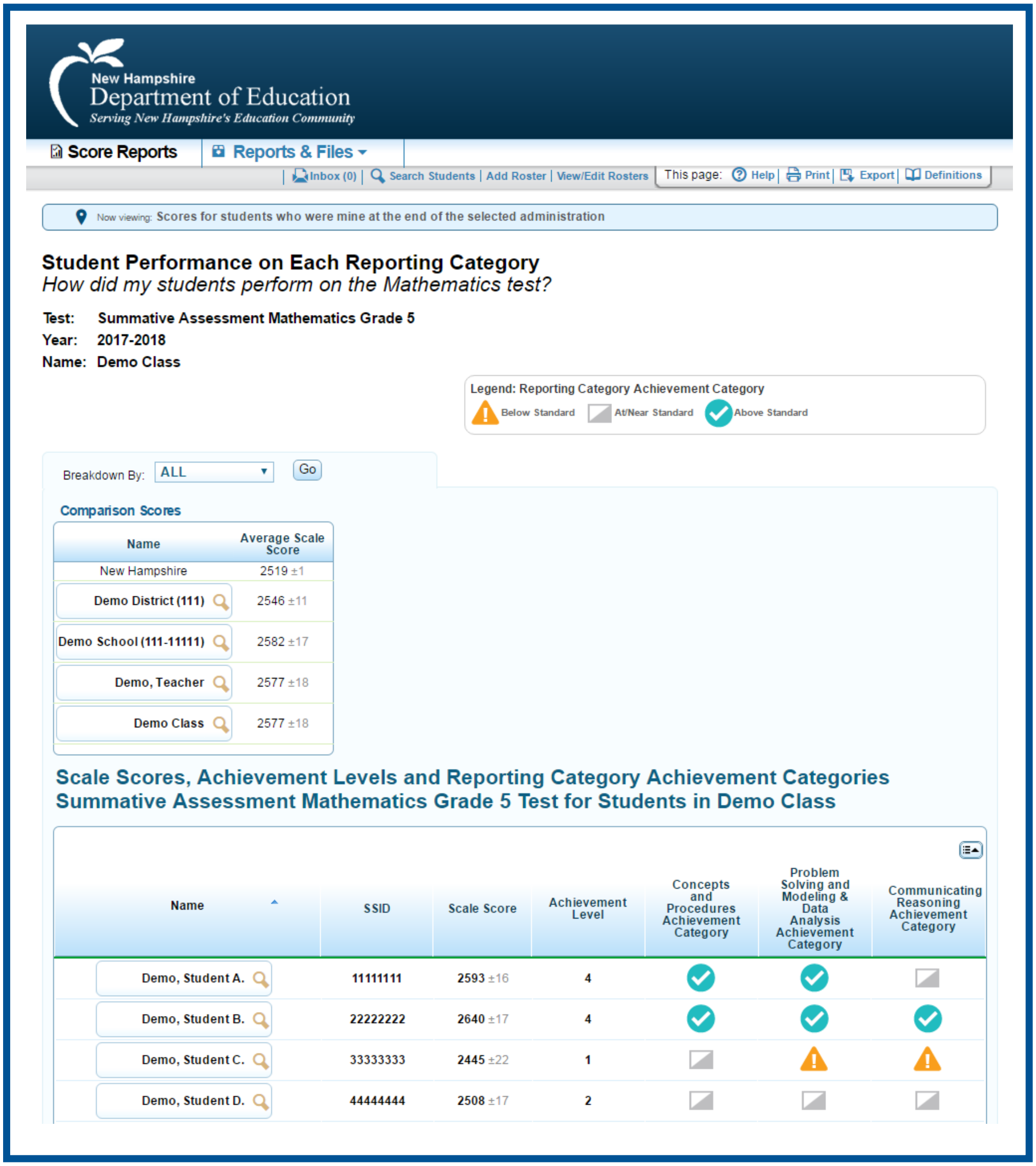
Legend: Areas Where Performance Indicates Proficiency
- ✓ Above the Proficiency Standard
- ◒ Borderline
- ✗ Below the Proficiency Standard
- ★ Insufficient Information

Performance on Each Benchmark
Summative Assessment Mathematics Grade 5 Test for Students in Demo School

Benchmark	Areas of Strongest and Weakest Performance	Areas Where Performance Indicates Proficiency
Concepts and Procedures		
Write and interpret numerical expressions.	✚	◒
Analyze patterns and relationships.	━	✓
Understand the place value system.	═	◒
Perform operations with multi-digit whole numbers and with decimals to hundredths.	━	✗
Use equivalent fractions as a strategy to add and subtract fractions.	✚	✓
Apply and extend previous understandings of multiplication and division to multiply and divide fractions.	═	◒
Convert like measurement units within a given measurement system.	✚	✗
Represent and interpret data.	★	★
Geometric measurement: understand concepts of volume and relate volume to multiplication and to addition.	✚	✓
Graph points on the coordinate plane to solve real-world and mathematical problems.	━	✗
Classify two-dimensional figures into categories based on their properties.	═	◒

Class Performance Rosters. Class performance rosters provide users with performance data for a group of students belonging to a system-defined or user-defined class. The report typically includes each student's unique state ID number, overall subject score, standard error of measurement, and, if generated from the reporting category summary page, reporting category scores or indicators. Exhibit D1.2-6 shows a sample class performance roster that indicates which reporting categories students are succeeding in and which ones students may need more help in understanding.

Exhibit D1.2-6: Reporting Category Report for Teacher's Roster

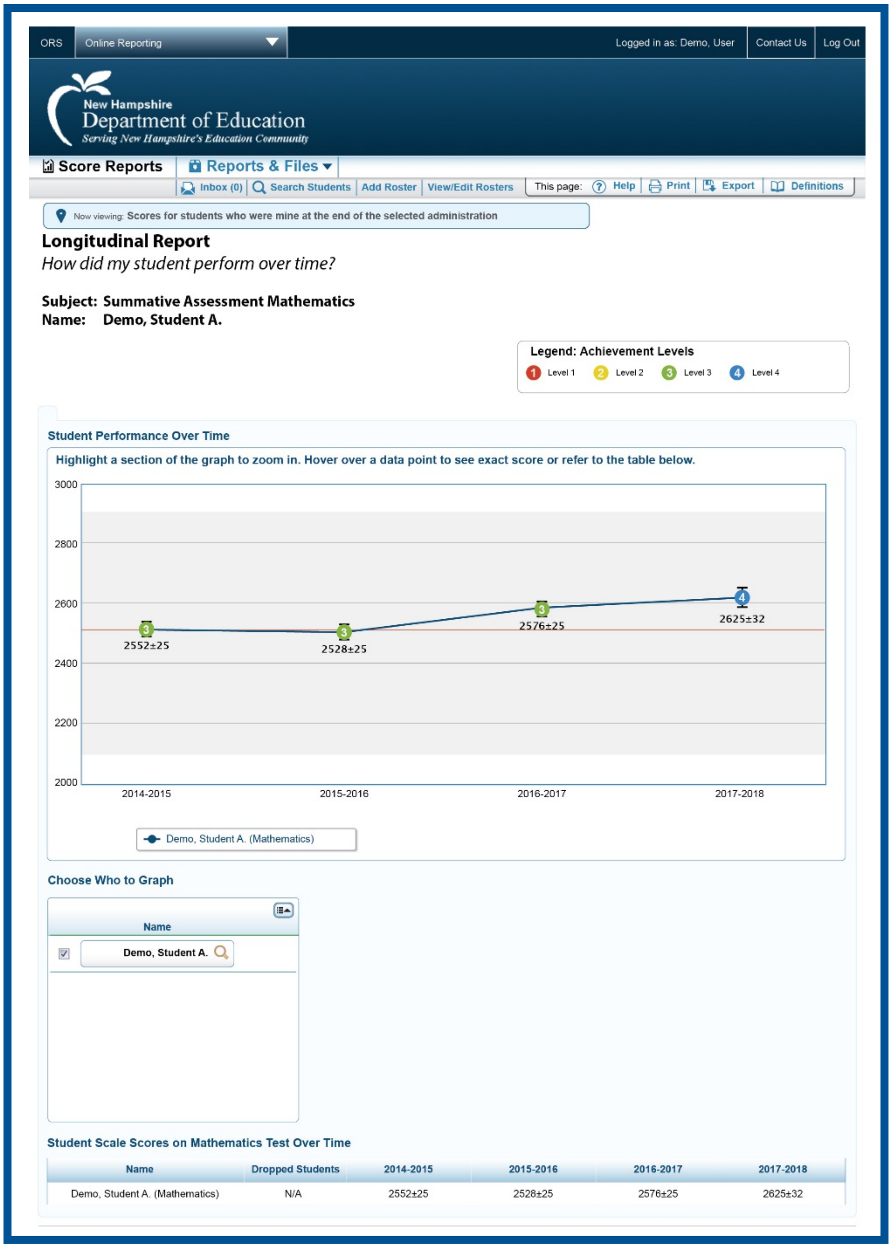


Individual Student Reports. The online individual student report typically includes scale scores, achievement levels, achievement-level descriptors, standard error of measurement around scale scores, and reporting category performance. AIR proposes to show reporting-category achievement level along with graphical representation of student's reporting-category score and standard error of measurement. Text for what these results mean and next steps will be variably placed, based on performance on each reporting category, so parents and educators can help the student improve on the particular areas of strengths and weaknesses. Student performance on the writing section offers detailed description of how a student performed on the writing portion of the ELA test. Exhibit D1.2-7 shows a sample individual student report.

Exhibit D1.2-7: Individual Student Report

Trend Reports. The state, districts, schools, teachers, classes, and students can be plotted on a trend report to illustrate how performance has changed over time. Scores can be across years (e.g., from spring administration to spring administration) and within a year (e.g., from interim comprehensive to summative). The longitudinal reports allow educators and administrators to analyze student growth when comparing performance result assessments. Multiple trend lines can be plotted at once to see how student growth compares between students or groups of students. Reporting category trends can also be added to see whether there are differences in performance within a subject. Exhibit D1.2-8 shows a sample trend report for a student across administrations in ELA assessments.

Exhibit D1.2-8: Trend Report for a Student



Summary of AIR Online Reporting System Capabilities

Next we describe the specific features and capabilities of the ORS.

Batch Printing of Individual Student Reports. AIR can deliver to each district a collection of student reports in PDF format, accompanied by a machine-readable manifest. This allows users to batch print multiple students' reports simultaneously. It also allows districts to upload the reports into their student information systems and, ultimately, deliver them electronically to parents. We believe that this approach saves the districts the time and costs associated with paper reports and that our world is now sufficiently connected to make this a viable delivery option.

Customizability. All text and measures in the reporting tool are configurable, and AIR reporting staff will work closely with NH DOE to define all text and measures as part of the reporting specifications development. Users can also choose to show or hide individual columns in a table or sort columns in ascending or descending order.

Delivery of District Data Files. AIR's ORS includes a retrieve student results tool that allows users to export a raw data file for a district, school, teacher, or roster as the data become available. The files can be exported in Microsoft Excel or any other delimited file format so that they can be easily loaded into other software platforms. Users can export any table on any report in the reporting system. Printed tables carry with them a footnote indicating the date on which the report was generated.

Mobile-Compatible Reports. All reports and features of the ORS are tablet-accessible. All tables and data displays adjust to the tablet screen size, making it easy to view rosters, request downloadable data files, or view any number of custom-created reports.

Report Security and Control of Access. The reporting tool belongs to AIR's suite of programs and can be accessed through the Common Login System (CLS). It allows educators and administrators to move seamlessly between authorized programs without needing to reenter login information. For example, a teacher who is moving between TIDE and the ORS will not need to sign in to different applications. When a user signs in, he or she will only see the reports and student data to which he or she has access. Besides ensuring that users do not have access to unauthorized reports, this also helps to improve the user experience by only displaying data and reports that are relevant to the user.

As with all our systems, student data are protected by the same physical, network, and software security as our TDS. If required, specialized rules can be applied to protect anonymity. For example, the system can be configured to display only student data above a minimum cell size threshold.

Reporting on Multi-Year Student Data. When a user logs in to the reporting system, he or she is automatically taken to his or her reports for the current test administration. Educators can also view their current students' previous-year scores, a capability that helps teachers understand where students may have faced challenges or exhibited strengths over time. When reflecting on his or her own practice, a teacher can review reports for students from past administrations but only for the years leading up to and including the year he or she taught those students.

Student Search. Sometimes a user simply wants an individual student's report and wants it quickly. The student search feature lets users search by student ID number, student name, or any partial student ID number or name. Users can only see student records to which they have access, though users can also search students from previous years. This capability allows a user to see an individual student's results with a minimal number of mouse clicks.

Drill Down Options. The report navigation is designed to allow the user to drill down through the different dimensions of data. For example, a user can move from the school reports to the teacher reports to the roster reports to the individual student reports. At any point, the user may also switch from subject-level reports and drill down into the reporting category reports or standard reports. The user can also switch from a specific administration to view a trend report over time. This flexibility allows the user to drill down to any desired level of granularity in the data and then modify dimensions of the report with a few clicks without needing to return to the beginning and drill down again.

Aggregation and Disaggregation. In addition to the expected aggregated reports for a district, school, teacher, or class, the reporting tool allows users to create custom rosters of students so that groups of students in a class can be tracked and reported separately. All aggregated reports are calculated from the student level up; as new students test, the reports for the roster, teacher, school, and district are updated instantaneously. All aggregate reports in the reporting system can also be disaggregated by sub-group or test condition event as defined by NH DOE. For example, reports can be disaggregated by demographic characteristics such as gender, ethnicity, or ELL status. The reporting system can also be configured to disaggregate by multiple characteristics at once (e.g., gender by ethnicity).

Interpretation of the Reports. AIR knows that a report is not useful if it is not interpreted validly. The reporting tool's design, layout, graphical displays, and use of color have all been implemented based on rigorous focus groups conducted with educators, administrators, and parents. Navigation emphasizes the context of a student's performance by relating it to aggregate performance and trends over time. Colors associated with performance are used consistently throughout the report, and graphical displays are chosen carefully to maximize the information conveyed. For example, bar charts can illustrate the percentage of a teacher's students at each performance level, but the relative positioning of each teacher's bar chart can also efficiently communicate which teachers have a higher performing group of students, even if there are differences by performance level.

The reporting system comes with an intuitive user guide (interpretive guide) that provides step-by-step instructions for every report, information about each measure included in the reporting system, and screen captures to illustrate processes and outputs. The user guide is accessible from any page in the reporting system. The reporting system also has a configurable Definitions tab that provides a quick reference for users who want more information about how to interpret a report.

In the past few years, AIR has collaborated with NH DOE to conduct webinars of the ORS and AIRWays. If awarded with the contract, AIR will continue to work closely with NH DOE to help parents and educators understand and interpret the results accurately.

Overview of AIRWays for Interim Block Assessments

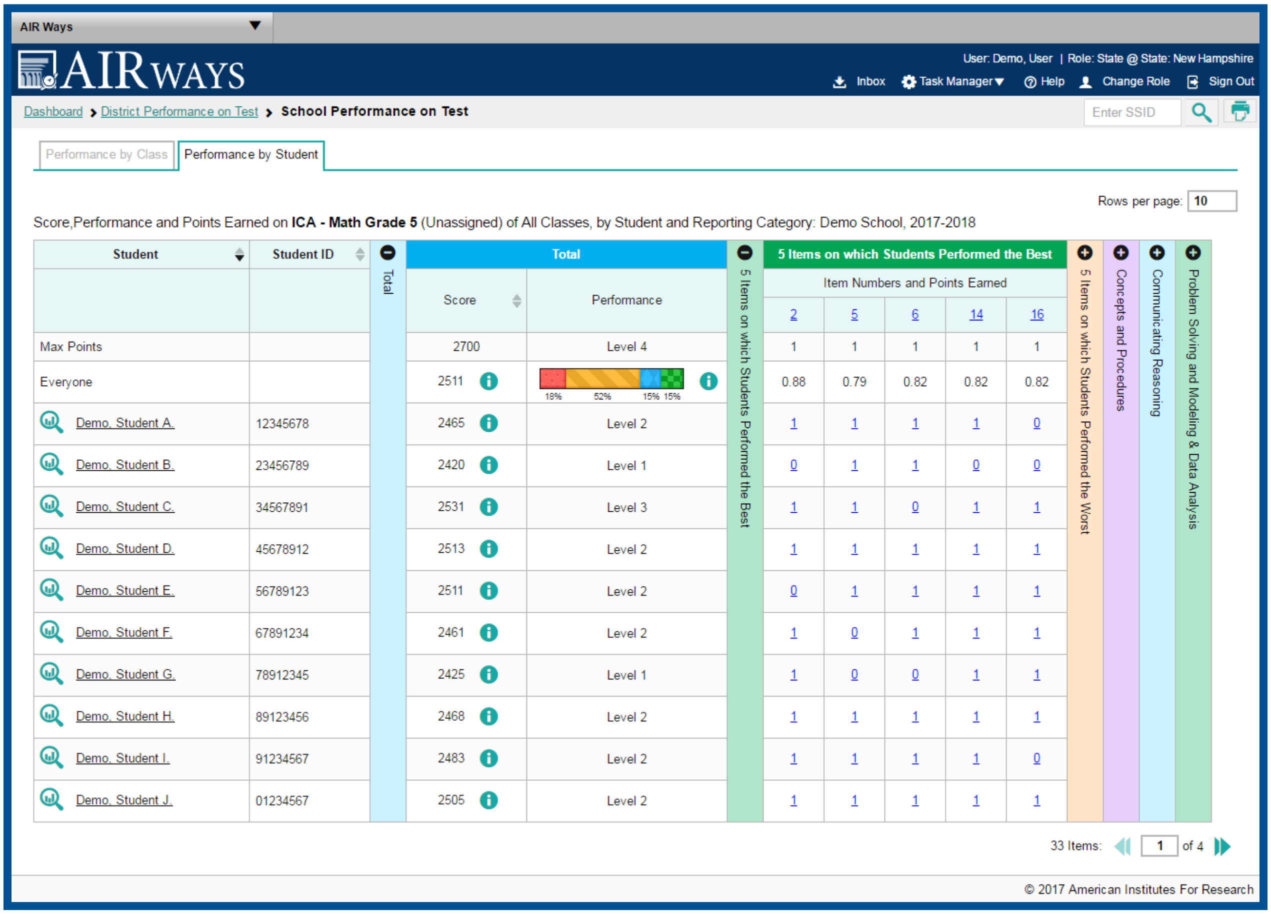
In addition to the reports in the ORS, AIRWays will provide New Hampshire educators with robust, timely reports for the interim block assessments. Results for interim block assessments are reported through AIRWays only, because individual items and students' response to each item are provided through AIRWays.

AIRWays is designed to leverage testing events to drive students and teachers to interact. The system reports only non-secure test results and shows the teacher not only the item, but each student's actual response. This provides a platform and opportunity for the teacher and student to begin exploring gaps in knowledge and misconceptions and otherwise to support instruction.

In New Hampshire, approximately 115 interim tests had been administered and reported through AIRWays as of this writing this school year. Feedback from the field has been extremely positive.

The teacher assessment report included in Exhibit D1.2-9 provides teachers with their interim scores, performance, and/or points earned for the selected assessment of all students. This report can be used to determine how individual students have performed compared to the rest of the students. It also displays the five assessment items with the highest or lowest average score for students, which helps teachers analyze students' strengths and weaknesses.

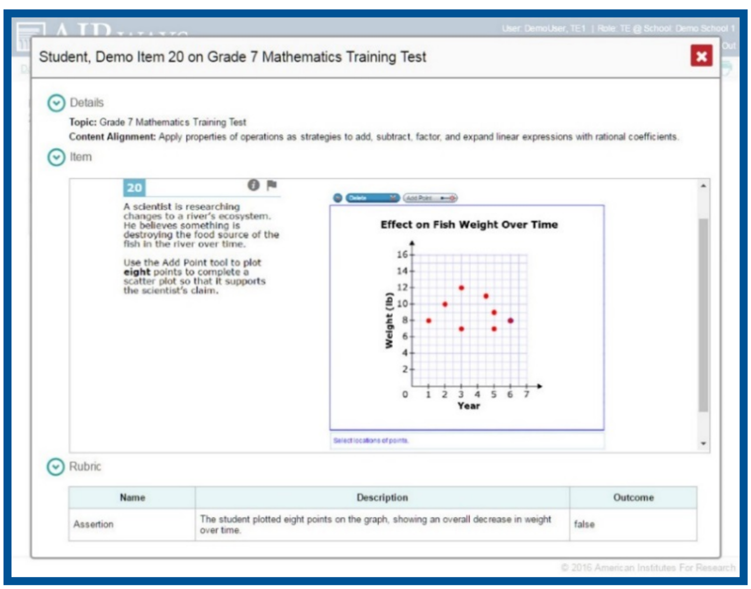
Exhibit D1.2-9: Teacher Assessment Report



Student	Student ID	Total Score	Total Performance	5 Items on which Students Performed the Best — Item Numbers and Points Earned: 2	5	6	14	16
Max Points		2700	Level 4	1	1	1	1	1
Everyone		2511	18% 52% 15% 15%	0.88	0.79	0.82	0.82	0.82
Demo, Student A.	12345678	2465	Level 2	1	1	1	1	0
Demo, Student B.	23456789	2420	Level 1	0	1	1	0	0
Demo, Student C.	34567891	2531	Level 3	1	1	0	1	1
Demo, Student D.	45678912	2513	Level 2	1	1	1	1	1
Demo, Student E.	56789123	2511	Level 2	0	1	1	1	1
Demo, Student F.	67891234	2461	Level 2	1	0	1	1	1
Demo, Student G.	78912345	2425	Level 1	1	0	0	1	1
Demo, Student H.	89123456	2468	Level 2	1	1	1	1	1
Demo, Student I.	91234567	2483	Level 2	1	1	1	1	0
Demo, Student J.	01234567	2505	Level 2	1	1	1	1	1

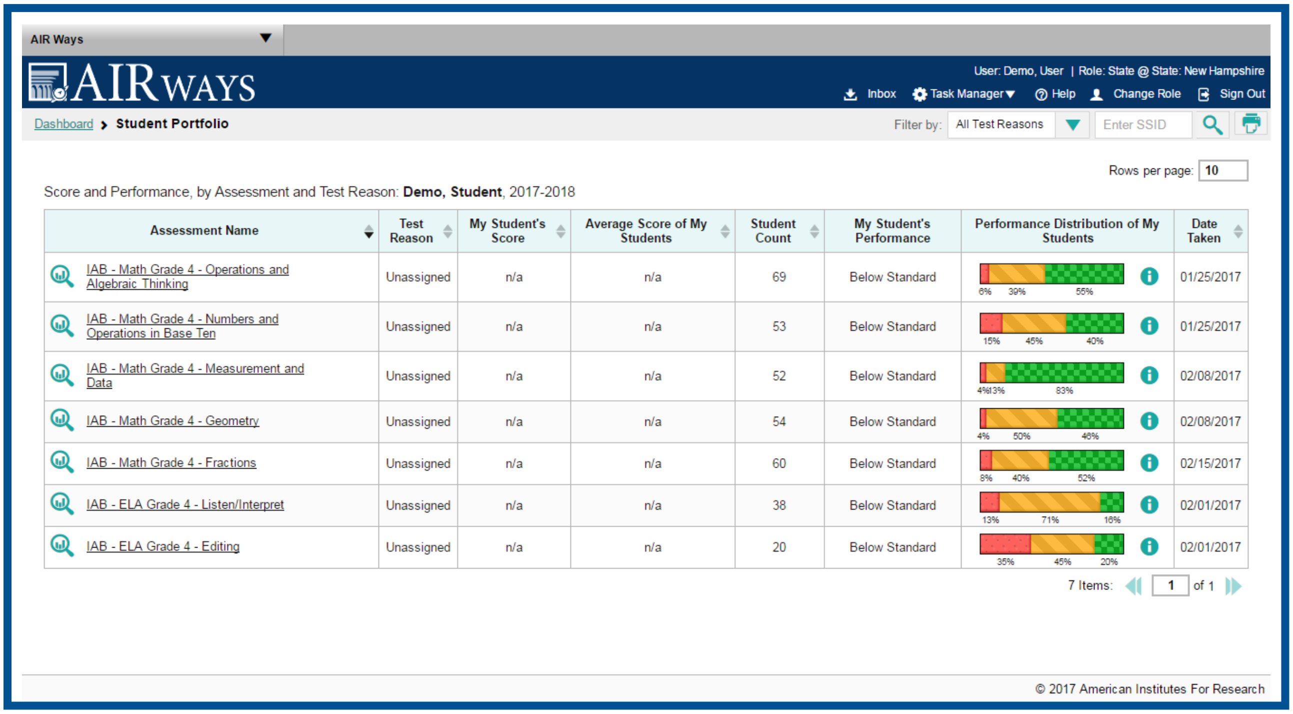
The AIRWays system also allows teachers to see exactly how a student responded to an item along with the actual test item (see Exhibit D1.2-10). By seeing actual student responses, teachers can more easily identify potential gaps in understanding and use that information to provide targeted instruction.

Exhibit D1.2-10: Student Item Response Report



In addition to providing teachers with actual student responses, AIRWays provides useful summary views of all interim block scores. For example, the student portfolio report shown in Exhibit D1.2-11 allows a teacher to see every interim block assessment taken by a specific student and quickly pinpoint that student's strengths or weaknesses.

Exhibit D1.2-11: Student Portfolio Report


AIR Ways

AIRWAYS

User: Demo, User | Role: State @ State: New Hampshire

Inbox Task Manager Help Change Role Sign Out

Dashboard > Student Portfolio

Filter by: All Test Reasons Enter SSID

Rows per page: 10

Score and Performance, by Assessment and Test Reason: **Demo, Student**, 2017-2018

Assessment Name	Test Reason	My Student's Score	Average Score of My Students	Student Count	My Student's Performance	Performance Distribution of My Students	Date Taken
IAB - Math Grade 4 - Operations and Algebraic Thinking	Unassigned	n/a	n/a	69	Below Standard	6% 39% 55%	01/25/2017
IAB - Math Grade 4 - Numbers and Operations in Base Ten	Unassigned	n/a	n/a	53	Below Standard	15% 45% 40%	01/25/2017
IAB - Math Grade 4 - Measurement and Data	Unassigned	n/a	n/a	52	Below Standard	4% 13% 83%	02/08/2017
IAB - Math Grade 4 - Geometry	Unassigned	n/a	n/a	54	Below Standard	4% 50% 46%	02/08/2017
IAB - Math Grade 4 - Fractions	Unassigned	n/a	n/a	60	Below Standard	8% 40% 52%	02/15/2017
IAB - ELA Grade 4 - Listen/Interpret	Unassigned	n/a	n/a	38	Below Standard	13% 71% 16%	02/01/2017
IAB - ELA Grade 4 - Editing	Unassigned	n/a	n/a	20	Below Standard	35% 45% 20%	02/01/2017

7 Items: 1 of 1

© 2017 American Institutes For Research

In summary, the AIRWays system will allow New Hampshire teachers to use data from the interim block assessments to

- determine a baseline of their class's strengths and weaknesses at the beginning of the year;
- track their class's and students' progress throughout the year;
- determine whether students understand the concepts that are being taught right after receiving the relevant instruction by looking at actual student responses;
- identify the areas of the test that students may be struggling with during the year and which require additional instruction;
- provide additional instruction and administer the interim block assessment again to see if particular students or groups have improved their understanding;
- create custom groups or rosters to track the performance of learning groups and confirm whether the composition of learning groups needs to be redefined; and
- view multiple interim block scores to identify student growth during the year.

Topic 23 Data Files

Process for Producing Data Files

AIR's automated data intake, processing, and extract systems are designed for high-volume efficiency and accuracy. In this section we describe how data flows through our systems as students are testing, along with the capabilities of the Data Extract Generator (DEG).

Each time a student completes an online test, the test delivery system (TDS) transmits data from that test event to our Quality Monitor (QM) system. Transmitted data includes

- student, school, and district identifiers;
- raw and scored student item responses;
- accommodation records; and
- initial scores (i.e., scores based only on machine-scored responses).

Upon receipt of the student record, the QM system

- verifies scoring of student responses;
- validates scores presented to the student;
- calculates any other scores that will be used in the reporting systems along with standard errors of measurement;
- calculates derivative measures from those scores (e.g., proficiency designations); and
- performs customized validations of incoming data.

In virtually all cases, these checks are all passed. When they are not, an automated system alerts project team members, who investigate and resolve the issue.

The QM system's quality checks are typically completed within a second of receipt of the data. Once complete, the data are transmitted to the Database of Record (DoR) and the ORS.

The QM system's quality reporting interface accesses the DoR to provide reports on a variety of ongoing quality monitoring issues. These statistical reports include monitoring of item exposure, item drift, and match to blueprint.

With the exception of item drift (which simulations cannot predict), the content of these reports is well predicted by the simulations that we conduct prior to deployment, so surprises are rare. AIR's data analytics and forensic procedures are described in detail in Topic 8.2 Data Forensics.

The DEG merges final testing data from the DoR with student data that comes into the system through TIDE into the student result files that are delivered to the NH DOE. States typically choose to transmit testing results with student biographic data using a single layout.

AIR can produce General Research Files and Biographic Files using separate layouts for students who have tested, depending on the needs of the NH DOE.

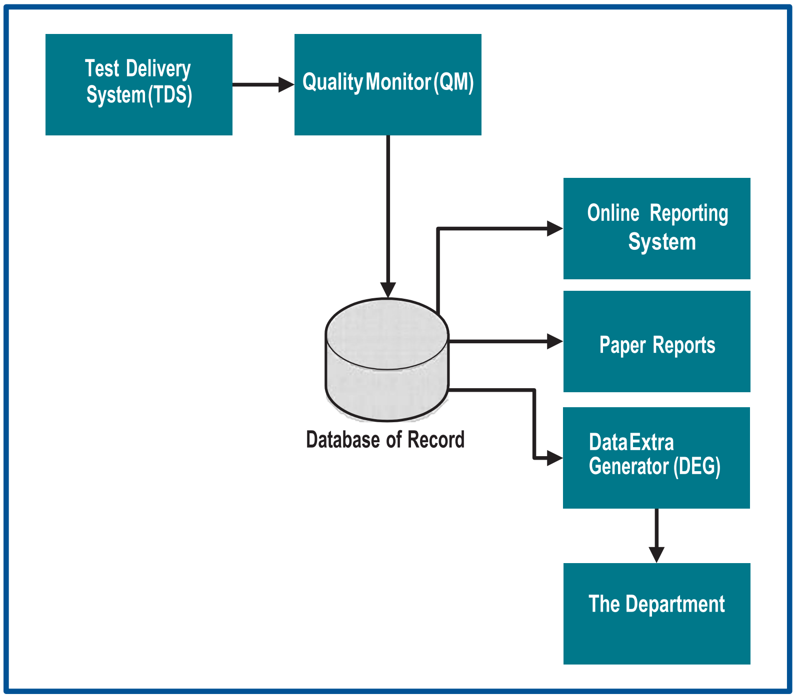
Data from these systems can be extracted and delivered with whatever frequency the NH DOE desires, up to once per day. We will work with the NH DOE to establish

- a customized data delivery format;
- a set of data validation rules for the files sent to the NH DOE; and
- a mutually agreeable transfer protocol (we typically use secure FTP to monitored directories).

These data delivery files are automatically validated against the format, using the established business rules prior to delivery. States receiving nightly updates typically retrieve the data files from the FTP server programmatically and automatically update the state data warehouse or other data store.

NH DOE personnel and educators throughout the state have immediate access to the data through the ORS. The data flow from the test delivery system all the way through to the reporting system is depicted in Exhibit D1.2-12 .

Exhibit D1.2-12: Data Flow Process for Reporting Results



Process for Producing Interpretive Materials

AIR will collaborate with NH DOE on development and production of interpretive products for reporting. Our experience assisting state Departments of Education in transitioning to new standards, new assessments, and online tests has taught us that clear, detailed, interpretive products are critical to building stakeholder success for any new assessment system. Understanding how new item types assess the new standards and how they are scored is essential if teachers are to provide effective classroom instruction. Similarly, both parents and educators need to understand how to interpret their student's test results so they can identify strengths and weaknesses early on. AIR will provide a full suite of interpretive products to ensure that New Hampshire parents, students, and educators understand the new standards, assessments, and test results. This suite includes the following:

- A Parent Brochure for New Hampshire's hard copy student score reports;
- An Interpretive Guide to help New Hampshire educators, schools, and districts understand New Hampshire's online reports

The section below describes the proposed content for the guides in more detail.

Proposed Content for Parent Brochure for New Hampshire's Hard Copy Student Score Reports

We propose to include the following content in this parent brochure and will work with the Department to obtain approval on detailed outlines before developing the guide:

- Full color screen captures of the ELA, mathematics, and science score reports to students and families along with explanatory text
- Plain-language explanations about the content areas and strands tested
- An explanation of how to interpret scores and Standard Error of Measurement

- A glossary of terms
- Detailed explanations of how released performance tasks are scored, including the released item, alignment to the standards, and annotated student responses explaining why the response was awarded the points received
- Links to more in-depth training materials

Overview of the ELA, mathematics, and science summative tests will provide a brief description of the assessment; the purpose it serves in improving New Hampshire's education system; and a Frequently Asked Questions section that will provide basic information about why students are taking the assessment, what happens if they do not score at proficient level, and who develops the assessment.

Performance Level Descriptors provides an explanation of the expectations of students for each subject and performance level. This section will allow parents and educators to know what is required for a student to perform at a particular proficiency level and will also provide plain-language explanations that include specific examples of what typical students can do at each level.

Interpretation of Student Scores includes background information about the scoring process for the ELA, mathematics, and science assessments. This section should define the terms used in the assessment student reports, such as scale score and subscale performance. In addition a description of the relationship between raw cut scores, which represent a student's overall score in a subject area, and sub-scores, which indicate a student's strengths and weaknesses in that subject's reporting category should be included. We propose that this section also include a list and definitions of all possible non-scorable codes that may appear on a student report.

Understanding Your Student's Test Scores includes information about the scores. We propose to include screen captures of each page of the printed parent report and provide detailed explanations of each element. For example, using non-technical language, we can explain the information presented in graphic displays, describe how to interpret a student's strengths and weaknesses using a sample report, and provide a glossary of jargon-free definitions of any technical terms that may be included in the reports (e.g., constructed-response question and score levels in the family guide). Should the Department choose to include Next Steps text in the parent reports, this section may also contain an explanation of how parents can use the customized recommendations to help their child improve his or her understanding of the standards. This section might also contain a detailed explanation of how to use the longitudinal data charts to compare student performance from year to year.

Proposed Content for the Online Reporting System Interpretive Guide Publication

ORS contains volumes of online score data. In addition to absorbing volumes of online score data, educators must also be able to understand and take appropriate actions based on these data. AIR will develop a interpretive guide for the ORS that describes the various reports available and provides a conceptual framework to help users quickly locate the right level of data, interpret its impact, and identify the concrete actions they can take to help students improve.

Over the past years, we have produced concise, effective interpretive guides for the NH Smarter Balanced Online Reporting System explaining to novice and experienced users alike how to access reports, create custom rosters, investigate trends in the data, and generate PDF reports. A similar guide will be produced for the new assessments and, as required by the RFP, made available in an electronic PDF for posting to the NH DOE website and AIR reporting systems. We will work closely with the Department staff to produce an interpretive guide that includes step-by-step instructions on how to use the Online Reporting System and AIRWays.

For educators, the ORS is used to establish a baseline of what their incoming students know and can do based on their performance from the previous year; track a group's performance on the interim assessments, which can help teachers adjust their instruction to the different levels of understanding in each group; and analyze their instructional methods and what worked and did not work well.

This framework is used to organize the information in this guide, which is divided into three main sections:

1. Understanding what your current students know by analyzing their data from the previous school year.
2. Determining whether your students are improving their understanding by analyzing their data from the interim assessments.
3. Reflecting on your own teaching practices by analyzing student data from the end of the school year.

This guide will also include the following sections to help users identify the reports and data most helpful to educators:

- A summary of the assessments reported in the ORS and when the data will be available
- Definitions of terms to help users understand students' performance
- Explanation of how to access reports and data
- Appendix that explains detailed functions such as printing reports and downloading data

The ORS Interpretive Guide will walk educators and administrators through the ORS. It will begin by describing how the data in the system can support the educators and administrators as they make instruction decisions. Then it will walk them through scenarios demonstrating appropriate uses of the data. In introducing each set of reporting capabilities, the system will describe why those reports are available, how they can help at the school or in the classroom, and what the limitations of the data are. The manual will show users how to create classes and other rosters of students and how to use these rosters to create aggregate statistics or track groups of students over time.

Production of Interpretive Guides

AIR will follow the same general process for the development of all materials. We divide the process into two phases:

- A drafting phase, during which we obtain agreement on the content and design of the document
- A production phase, during which we produce the document to the specifications developed in the first phase, accommodate any final adjustments, and ensure that the camera-ready proofs are error free

The drafting process begins with discussions with NH DOE, during which we discuss issues of both design and content. Once drafts are approved, we proceed to production. The first blackline version of the document, which results from the first iteration through our production process, is designed to reflect the final layout and formatting.

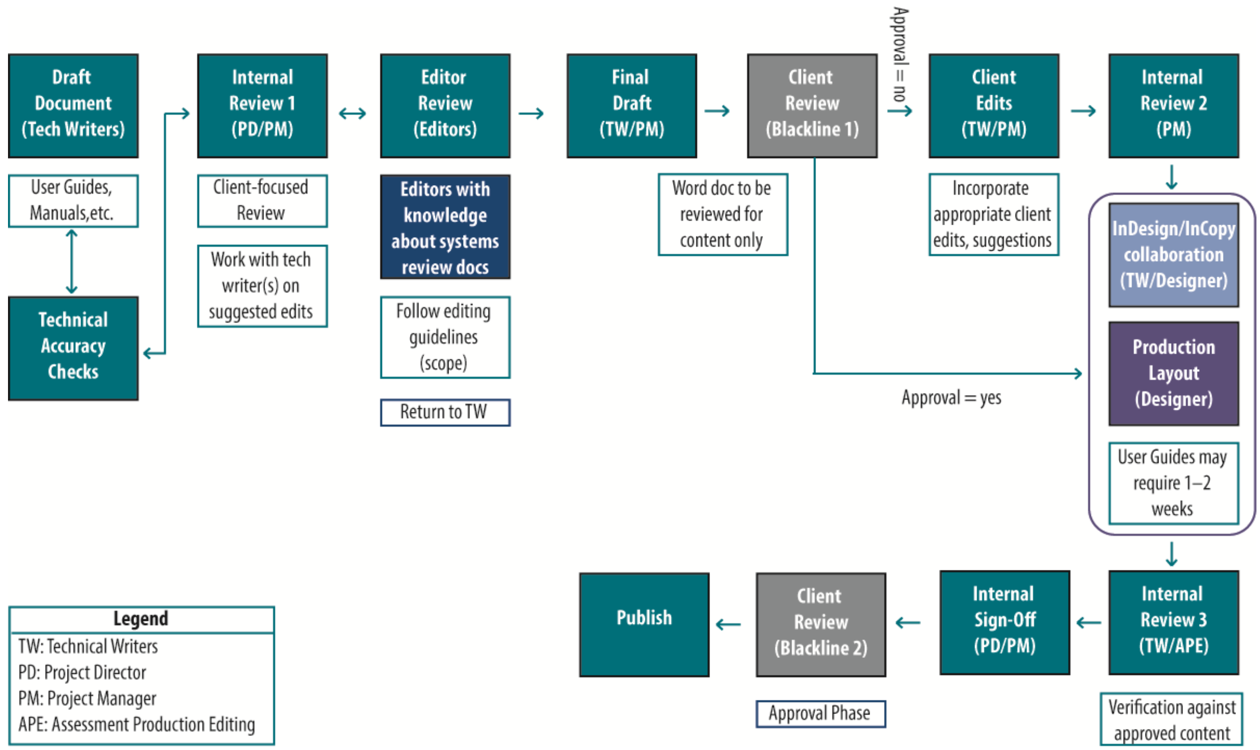
Typically, only format and layout changes are made at blackline. A second blackline is produced after incorporating these changes to confirm that they were made correctly.

Our production processes include multiple editorial reviews, and each document must receive senior staff sign-off. Exhibit D1.2-13 provides an overview of our editorial review process. Each document's content is fully drafted before the document enters the production system.

Camera-ready drafts are called blackline proofs, and we typically produce up to three of these, with final edits made on blackline 1, any missed changes corrected on blackline 2, and a clean version delivered at blackline 3. The final blackline proof must have no changes, and it serves as the master document against which the blueline (printer proof) is evaluated.

AIR will update the manuals and guides annually to reflect changes in the system.

Exhibit D1.2-13: Overview of Our Editorial Review and Revision Process



D1.7 Standard Setting

Interpretation of test scores rests fundamentally on how test scores relate to performance standards that define the extent to which students have achieved the expectations defined in the academic standards. For example, the currently administered Smarter Balanced test scores are reported with respect to four proficiency levels, demarcating the degree to which New Hampshire students have achieved the learning expectations defined by the Common Core State Standards. With the development of new statewide assessments, the Department will need to adopt a similar system of performance standards to determine whether students have met the learning goals defined by the New Hampshire College and Career Ready Standards. The cut score establishing the Proficient level of performance is the most critical, because it indicates that students are meeting grade-level expectations for achievement of New Hampshire College and Career Ready Standards, they are prepared to benefit from instruction at the next grade level, and they are on track to pursue post-secondary education or enter the workforce. Thus, the procedures used to adopt performance standards for the new science assessments are central to the validity of test score interpretations.

Topic 24 Performance Levels

AIR will work with the Department to identify the number of performance levels to be reported for New Hampshire's statewide assessments and their associated labels. We recommend that the Department identify no more than the four levels required to meet federal assessment reporting requirements. To meaningfully support reporting of more proficiency levels requires a test length that is greater than what the Department is looking for with respect to test administration time.

Performance Level Descriptors (PLDs) define the content area knowledge, skills, and processes that students at a specific performance level are expected to possess. The standard-setting panelists will base their judgments about the location of the achievement levels by using the PLDs as well as the subject area content standards. AIR will provide the Department with draft PLDs for review. The Department may wish to consider posting the PLDs for review by educators to incorporate important feedback from the field in the final set of PLDs. This procedure also ensures that panelists are already familiar with the PLDs before the standard-setting workshop, which greatly facilitates their performance.

Topic 25 Cut Scores

New Hampshire educators, serving as standard-setting panelists, will follow a standardized and rigorous procedure to recommend performance-level cut scores. We propose that the workshops employ the Bookmark procedure, a widely used method in which standard-setting panelists use their expert knowledge of New Hampshire College and Career Ready Standards and student achievement to map the performance-level descriptors (PLDs) adopted by New Hampshire onto an ordered-item booklet (OIB) constructed from items administered in the first operational test administration. The OIBs will be constructed to be proportional to the blueprint and to minimize information gaps between consecutive items.

We propose to provide panelists with contextual information to help inform their primarily content-driven performance standard recommendations. Because item parameters estimated from each of the ICCR participating state assessments are linked to the common ICCR scale, the performance standards and other benchmarks established in each of the ICCR state assessments can be represented on the ICCR scale. These linkages allow the Department to provide panelists with benchmark information about the location of performance standards across multiple assessments systems when considering the location of performance standards for the New Hampshire statewide assessments. Locations where performance standards from multiple important assessment systems converge represent neighborhoods in which performance standards likely reside.

For each assessment, panelists will be provided with the approximate location of performance standards for other important assessment systems. For ELA and math, in addition to the ICCR participating states (Arizona, Florida, Ohio, Tennessee, and Utah), the ICCR scales have also been linked to the Smarter Balanced scale, allowing those performance standards to be represented in the OIB as well. In addition, panelists recommending performance standards for the grades 4 and 8 summative assessments will be provided with the approximate location of relevant NAEP and TIMSS performance standards (with interpolated locations for grade 6).

As the Department is aware, AIR is working collaboratively with several states to develop and implement new science assessments designed to assess the three-dimensional science standards. Linkages between participating states and the underlying ICCR scale will also allow for identification of performance standards in those states to be located in the science OIBs.

Panelists will be asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists can use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

For ELA and math, we propose to convene grade-band standard-setting panels (e.g., grades 3–4, 5–6, and 7–8). Panels begin by recommending performance standards in anchor grades (e.g., grades 4, 6, and 8) using the standard Bookmark procedure. Following the recommendation of cut scores for the anchor grades, panelists convene to vertically articulate performance standards across the anchor grades. Once recommended cuts scores for the anchor grades have been established, we then interpolate the location of performance standards in the remaining grades. With interpolated performance standards in hand, the judgment task for panelists is modified so that panelists are now asked to judge whether, for each interpolated cut score, two-thirds of students who just barely qualify for entry into the performance level can respond successfully to the item on the interpolated page. If the judgment is yes, panelists are asked to endorse the interpolated ordered-item booklet page as the performance standard. If they cannot endorse the interpolated OIB page as the performance standard, panelists are asked whether they can locate an item very near the location of the interpolated OIB page where two-thirds of students just barely qualifying for entry into the performance level can respond successfully. This approach allows panelists to explicitly consider the important goal of recommending vertically coherent performance standards when recommending performance standards.

Because the new science assessments are administered once per grade band and can be expected to measure quite different content between the grade 5, grade 8, and grade 11 assessments, developing a vertical scale to measure growth over time is not feasible. Thus, while ELA and math panelists will be provided with feedback about how their performance standard recommendations relate to student growth over time, such feedback cannot be provided for the new science assessments. Nevertheless, the recommended performance standards for the science assessments must be viewed as part of a coherent system of standards, both between grade bands and across subject areas. Thus, panelists will also be provided with feedback about the vertical articulation of the impacts of their recommended performance standards across grade bands so that they can view the relationship between the locations of recommended cut scores for each grade-level assessment to the cut-score recommendations at the other grade levels. This approach allows panelists to view their cut-score recommendations as a coherent system of performance standards and further reinforces the interpretation of test scores as indicating not only achievement of current grade-band standards, but also preparedness to benefit from instruction in the subsequent grade band.

Online Bookmark Procedure. To ensure vertically articulated performance standards across grade-level assessments, AIR proposes to employ an augmented Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001) to establish grade-level cut scores for New Hampshire’s statewide assessments. In this approach, panelists first employ the Bookmark procedure to recommend performance standards for anchor grades. Following recommendation of anchor grade standards, the full panel convenes to vertically articulate the anchor grade performance standard recommendations. With vertically articulated performance standards for the anchor grades, initial performance standard recommendations the remaining grades are identified by interpolating the anchor grade performance standards. Panelists employ a modified judgment task to recommend performance standards for the interpolated grades. If panelists determine that two-thirds of just barely qualified for entry into the performance standard can respond successfully to the item on the interpolated OIB page, they are asked to adopt the interpolated location as the standard. If panelists cannot adopt the interpolated page as the performance standard, they are asked to identify an OIB page as near to the interpolated page where two-thirds of students just barely qualified for entry into the performance level can respond successfully.

We recommend that the Bookmark method be implemented in two rounds, providing panelists with contextual (benchmark and impact) data before round one and panel feedback before round two. We also propose to employ our web-based approach to the Bookmark method, in which panelists access OIBs, review benchmark and impact data, and set standards online. AIR's standard-setting tool allows panelists to view and interact with items in the way they are presented to students, which is crucial to their evaluation of the knowledge and skills required for students to respond correctly to the item types developed for online assessments. The standard-setting tool also enables immediate calculation and reporting of results, reducing the overall workshop time and burden on panelists.

Online Ordered Item Booklets. AIR proposes using an online OIB developed from items in the first operational test administration. Items will be selected so that they are proportional to the test blueprint and minimize information gaps between adjacent items. The OIB will include approximately 50% more items than an operational test administration. Increasing the number of items in the OIB modulates the impact of bookmark placements, especially in higher- and lower-ability regions, freeing panelists to focus on content considerations when placing bookmarks. Although ELA and math items will have been previously field tested, classical item analysis and differential item functioning (DIF) statistics will be evaluated for all items to ensure that the selected OIB test items are not flagged for item characteristics that could result in their rejection from the operational item pool. The online OIB will be presented to the panelists using AIR's online standard-setting tool and will enable the display of complex item types, as well as benchmark and impact data, and item attributes such as content alignment.

Workshop Panelists and Stakeholder Review Committees. AIR proposes to recruit classroom teachers and curriculum specialists with expertise in the grade-level content standards and instruction for the standard-setting workshops. For the science assessments, although tests are administered in grades 5, 8, and 11, panels will comprise educators with instructional expertise in each of the grades comprising the band. To ensure that the widest range of stakeholders has meaningful input into the standard-setting process, we propose to conduct a stakeholder review with a more general group after the standard-setting workshop. This group will consist of table leaders from the standard-setting workshops and representatives from other important stakeholder groups.

Performance-Level Descriptors. PLDs define the content area knowledge, skills, and processes that students at a specific performance level are expected to possess. The standard-setting panelists will base their judgments about the location of the achievement levels by using the PLDs as well as the subject area content standards. AIR will provide the Department with draft PLDs for review. The Department may wish to consider posting the PLDs for review by educators to incorporate important feedback from the field in the final set of PLDs. This procedure also ensures that panelists are already familiar with the PLDs before the standard-setting workshop, which greatly facilitates their performance.

Panelist Recruitment. To recruit panelists, AIR will assist the Department to actively seek individuals from the diversity of backgrounds found in New Hampshire. For ELA and math, AIR will convene three grade-band panels (3–4, 5–6, and 7–8). Each grade-band panel will include approximately 12 panelists, divided evenly between grades, and grouped into three tables. For science, AIR will also convene three separate panels for each of the grade-band assessments comprising 12 panelists arranged in three tables and representing educators from each of the grade levels assessed by the standards. One panelist at each table will be assigned as a table leader. Table leaders will have the additional responsibility of ensuring that table activities remain focused, ensuring that panelists understand their assignment, and alerting workshop leaders of any issues encountered by panelists.

Training. Training at the meetings will involve helping panelists become familiar with the assessment system and the standard-setting process. The training will include a review and discussion of the assessments, the content standards, and the PLDs for each performance standard. Central to panelists' training will be their use of the PLDs to develop a representation of students who just barely qualify for entry into each identified performance level. During this training task, panelists learn that although PLDs are written to characterize typical members of each achievement level, their bookmark placements will be directed toward characterizing and identifying the most minimally qualified members of each performance level. All panelists will receive training before the start of the meeting on how to access and use the standard-setting tool.

Further, panelists will receive training in the placement of bookmarks in the online OIBs. As part of this training, panelists will learn to use AIR's online standard-setting tool to place a bookmark that best delineates two achievement levels.

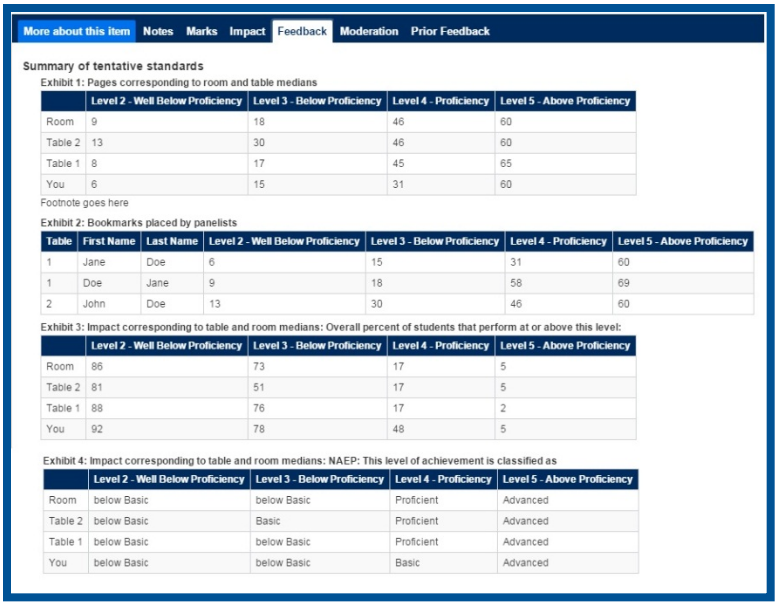
OIB Review. Following training on how to use the standard-setting tool, panelists will begin reviewing the online OIB for the grade-level assessment. For each item, panelists will be instructed to consider (1) what a student must know and be able to do to successfully answer each question and (2) what makes each item more difficult than the previous items. After reviewing the items individually, panelists engage in discussion with table members about the knowledge and skills required to respond to items in the OIB successfully. This process ensures that panelists have a solid understanding of the progression of item difficulty in the OIB and how the knowledge and skill requirements for successful responding relate to the PLDs. The process also ensures that panel members share a common understanding.

Bookmark Placement. Following review of the OIB, panelists are provided with additional training in the Bookmark procedure. This is done to ensure that they understand the procedure used to place bookmarks (e.g., place the bookmark on the last page in the OIB on which two-thirds of students who just barely meet the standard can respond successfully) and that they understand how to use the online tool to record their responses. Panelists complete a readiness assessment before recommending performance standards.

Panelists are instructed to review each page in the OIB and consider whether two-thirds of students who just barely qualify for entry into the performance level can respond successfully. They are further instructed to place their bookmark on the last page in the OIB where two-thirds of those just barely qualified students can respond successfully. While most standard-setting activities allow for and encourage group discussion, bookmark placement is an individual judgment task, and panelists are instructed not to consult with table members about their bookmark placements. Panelists begin their bookmark placements with the proficient-level performance standard. Once they feel comfortable about their placement of the proficient-level bookmark, they will move on to place bookmarks for the other performance standards.

Feedback and Impact Data. Immediately following the round one bookmark placements, AIR's online standard-setting tool will instantly evaluate and display the results from round one and create feedback reports for both the workshop table and the room as a whole. These reports include summary statistics showing the median, lowest, and highest bookmark placements for each performance standard. In addition, the tool contains impact data tables that contain scale scores and percentage of students at or above the level for each possible performance standard in the OIB for reference and discussion. Exhibit D1.7-1 illustrates the types of tables available to the panelists at feedback review.

Exhibit D1.7-1: Feedback from Standard-Setting Tool



More about this item | Notes | Marks | Impact | Feedback | Moderation | Prior Feedback

Summary of tentative standards

Exhibit 1: Pages corresponding to room and table medians

	Level 2 - Well Below Proficiency	Level 3 - Below Proficiency	Level 4 - Proficiency	Level 5 - Above Proficiency
Room	9	18	46	60
Table 2	13	30	46	60
Table 1	8	17	45	65
You	6	15	31	60

Footnote goes here

Exhibit 2: Bookmarks placed by panelists

Table	First Name	Last Name	Level 2 - Well Below Proficiency	Level 3 - Below Proficiency	Level 4 - Proficiency	Level 5 - Above Proficiency
1	Jane	Doe	6	15	31	60
1	Doe	Jane	9	18	58	69
2	John	Doe	13	30	46	60

Exhibit 3: Impact corresponding to table and room medians: Overall percent of students that perform at or above this level:

	Level 2 - Well Below Proficiency	Level 3 - Below Proficiency	Level 4 - Proficiency	Level 5 - Above Proficiency
Room	86	73	17	5
Table 2	81	51	17	5
Table 1	88	76	17	2
You	92	78	48	5

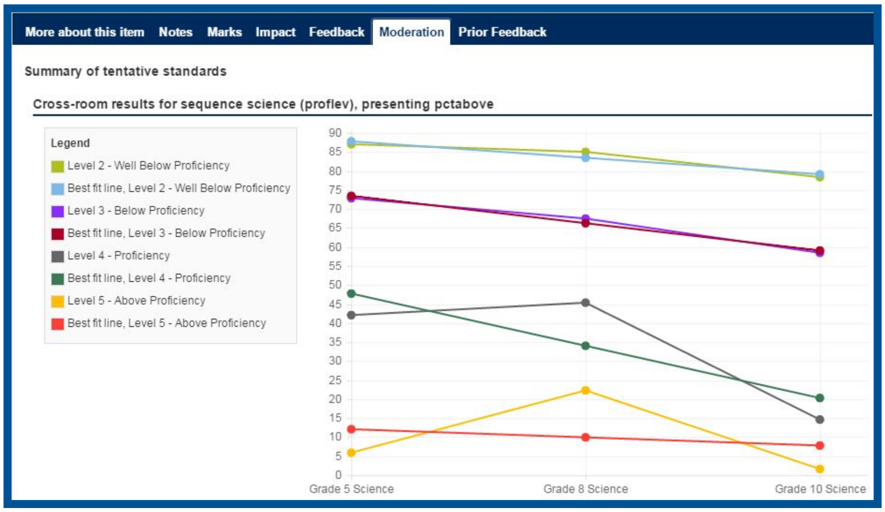
Exhibit 4: Impact corresponding to table and room medians: NAEP: This level of achievement is classified as

	Level 2 - Well Below Proficiency	Level 3 - Below Proficiency	Level 4 - Proficiency	Level 5 - Above Proficiency
Room	below Basic	below Basic	Proficient	Advanced
Table 2	below Basic	Basic	Proficient	Advanced
Table 1	below Basic	below Basic	Proficient	Advanced
You	below Basic	below Basic	Basic	Advanced

Round two begins with each table reviewing and discussing cut-score placements, focusing on the lowest and highest recommended cut scores and the table's median score. Discussion then expands to the room level, with each table reviewing for the group at large the basis for its recommendations. After completing these discussions, panelists again work through the OIB, placing their cut scores for all four performance levels.

Vertical Articulation and Moderation. Following round two recommendation of performance standards, AIR proposes to conduct a moderation session consisting of all table leaders from each panel. The table leaders will be tasked with evaluating whether the recommended performance standards are reasonable across all grade levels based on moderation tables generated in AIR's online standard-setting tool and to modify the location of the standards if necessary. This capabilities of the online standard tool is illustrated in Exhibit D1.7-2. If modifications to the performance standards are recommended at moderation, AIR's online standard-setting tool allows only a table leader to make the recommended performance standard change and captures the written reason for the amended performance standard recommendation.

Exhibit D1.7-2: Moderation in AIR's Standard-Setting Tool



Workshop Staffing. Each workshop panel will be staffed by an AIR senior workshop leader, a content specialist, and a research assistant. In addition, AIR psychometricians will be available to observe real-time data using AIR's online standard-setting tool. This allows the psychometricians to respond rapidly to recommendations being made in each workshop room.

Logistics and Expenses. AIR will conduct the standard setting workshop in a Department approved location. AIR will be responsible for all facets of meeting planning and execution, including meeting logistics and costs, participant notifications and registration, and materials production, distribution, and security, as well as stipends, and/or reimbursement costs for standard setting panelists.

Topic 26 Standard Setting Report

Technical Report. AIR will provide a comprehensive standard setting technical report that will be available for review by the Department and its Technical Advisory Committee (TAC). The technical report will detail the procedures, participants, and other relevant elements. The report will contain the following sections:

1. Executive summary, which is written for policy makers and can be used by the state for workshops
2. Background and history of the New Hampshire assessments
3. Preparation for the standard-setting workshop, which covers the development of the PLDs, psychometric details of the field test, item calibrations, estimation of impact data, explanation of the articulation process and benchmarking procedures, and the composition and demographics of the standard-setting panels
4. Conduct of the standard-setting workshop, which includes agendas, training procedures, details of the round one and round two estimates of cut scores and moderation sessions, security procedures, and panelist and stakeholder evaluation summaries
5. Results of the standard-setting workshop, which includes summaries of each round, reliability of the achievement-level standards, changes in cut scores resulting from the stakeholder review and moderation sessions, final cut scores, and the percentage at and above each achievement level

In addition, AIR will assist the state with any materials or any other assistance needed for the state's approval of the standards, including presentation of the recommended performance standards to the Board.

Topic 26.1 Standard Setting Validation

To provide consistency in test score interpretation and foster comparisons of student performance between cohorts as well as longitudinally, the Department will wish to maintain the performance standards adopted following the first operational test administration in spring 2018 despite anticipated revisions to the New Hampshire academic standards that are likely to impact test blueprints in subsequent test administration. In this section, we present our plan for evaluating the validity of adopted performance standards following modification of blueprints resulting from revision of the New Hampshire academic standards.

When test blueprints are modified, especially in substantial ways, the measurement construct being assessed may be significantly altered, casting doubt on the validity of some test score interpretations. In part, this concern arises because standard-setting panelists were guided by a set of PLDs and an OIB of test items constructed to meet the blueprints as originally enacted. The PLDs represent the valid inferences that can be made from student test scores and performance-level classifications. Changes to the test blueprints may affect the validity of some of the assertions in the PLDs.

To evaluate whether descriptions of performance levels continue to be valid descriptors of students achieving current performance standards, AIR will convene a panel of educators to review modified blueprints, PLDs, and OIBs to determine whether blueprint revisions warrant either modifications to the PLDs or reestablishment of performance standards.

Review of performance standards will be conducted by a panel of educators consistent with those recruited for the original standard-setting workshops. For ELA, where revisions to blueprints are systematic, we propose a single ELA panel with three representatives each from grade bands 3−5 and 6−8 and high school. For assessments with localized impacts, we propose assessment-specific panels comprising four participants.

Panelists will be provided with the current PLDs, the standard-setting OIBs, and an item map highlighting the location of items with respect to adopted performance standards. Panelists will also be provided with the original and revised blueprints, and the OIBs and item maps will highlight items that would no longer be administered given deletions in the revised blueprint or a range of additional items representing additions to the revised blueprint.

Panelists will review the PLDs and the notated OIBs and item maps in light of blueprint revisions to determine whether the PLDs as currently constructed continue to be valid descriptors of student achievement or whether the PLDs must be revised. Where revisions to PLDs are called for, panelists will work to revise the PLDs in ways that support valid interpretations of performance-level classifications.

AIR psychometricians will work with the Department and the TAC to determine whether results indicate that PLDs can stand with no or minor revisions, indicating that the assessed construct is reasonably similar following blueprint revision, or whether PLD revisions based on the revised blueprint indicate substantial alteration of the assessed construct and signify a need to adopt new performance standards.

D1.8 Reporting Portal

We propose to provide New Hampshire with a customized online administrative portal, a tailored one-stop shop for educators and other stakeholders to provide access to all components of the assessment system, including student enrollment, test administration, and online reporting systems, as well as practice sites, training sites, tutorials, and other resources. AIR will create an administrative portal as a central point of access for New Hampshire stakeholders, including educators, administrators, families, and students. The portal will house links to all online tests, including practice and training tests, and the supporting materials required for test administration. The structure of the portal may be organized according to role (test coordinator, teacher, technology coordinator, student/parent) and separate, subject-specific tabs such as Secure Browsers, Training, General Resources/Manuals, FAQs, and Contact Us. The portal tabs function like a table of contents, helping users organize and locate information as needed. The portal homepage also offers an easy-to-access location for important announcements and updates. Aside from functioning as a central information repository, the portal also provides each user role with password-protected access to only those components of the online testing system that are relevant to each specific role. Both the Online Reporting System (ORS) and AIRWays can be accessed through the portal as well. AIR staff will work with the Department to configure the portal in the most optimal way for educators and administrators. The administrative portal is discussed in further detail in Topic 4 Assessment Delivery Platform.

As described in Section D1.6 Reporting Portal, AIR proposes to use our ORS and AIRWays as the reporting portals that provide a platform to integrate data from state summative and interim assessments. Summative and interim comprehensive assessment results will be available through the ORS and the interim block assessment results will be available through AIRWays. Both systems provide educators with a user-friendly interface that increases access to assessment results. Educators can build customized online reports through a series of filters (e.g., demographic filters, accommodation filters, longitudinal filters) using the ORS.

In addition, AIR can offer different options for online reporting to parents. In our experience, the greatest barrier to statewide online reporting has to do with establishing and maintaining appropriate parent authorizations. States must rely on schools and districts to maintain appropriate parent/user lists. Should these be out of date or incorrect, the state may be liable for improper disclosure of student data. One solution to this problem is to make online reports available to districts in a format that can be easily uploaded into their existing parent portals. We are currently using this approach in multiple states including Utah and Florida.

In this way, the sensitive PDF reports are delivered to trusted district users who control access to them through their existing portals. AIR's ORS provides a second option. AIR maintains a near-real-time reporting system that parents can use, provided that the state or district authorizes parent users. This highly configurable system can make reports available to parents as the data become available. With the system deployed, it also offers substantial data exploration capabilities to teachers and schools.

D-2 CORPORATE OVERVIEW AND PROJECT MANAGEMENT

D2.1 Corporate Qualifications

Topic 27 Corporate Overview

Please see Section V: Corporate Qualifications for a complete narrative response.

Topic 28 Vendor Experience

Building on the Corporate Overview presented in Topic 27 Corporate Overview, this section details AIR's experience in each of the major areas of New Hampshire's RFP.

General Overview

Many clients are surprised to discover that AIR is now the undisputed leader in the United States for statewide summative assessment testing. In addition to delivering 43 different statewide assessment programs, including alternate assessments, End-of-Course (EOC) assessments, science assessments, and English language proficiency testing programs, AIR delivers the core grades 3–8 ELA and mathematics testing in 17 states, serving 34% of students nationwide. No other vendor serves more than 11 states or more than 21% of students nationwide, and no other vendor tests anywhere near as many students online as AIR.

Component A: Summative Assessments in ELA and Mathematics

AIR has successfully delivered and continues to deliver online statewide summative English language arts (ELA) and mathematics assessments for grades 3−8 in Oregon beginning in 2007; Delaware and Hawaii beginning in 2009; Ohio and Utah beginning in 2013; Arizona, Connecticut, California, Florida, Idaho, Maine, New Hampshire, South Dakota, Vermont, Washington, West Virginia, and the U.S. Virgin Islands beginning in 2015; and Montana and North Dakota beginning in 2016. These online programs are in addition to paper-based testing programs for many additional years.

Beyond administration, few companies can match AIR's test development achievements in the past three years. In ELA and mathematics, AIR has developed two complete banks of items aligned to Common Core State Standards (CCSS), including developing the test and item specifications and contributing significantly to the development of the Smarter Balanced item bank.

We have proposed to build New Hampshire's Summative ELA and mathematics assessments using ICCR. AIR has developed more than 5,800 ELA and mathematics items and more than 110 writing performance task families for AIR's ICCR Independent College and Career Ready (ICCR) bank, which continues to expand. AIR built ICCR to support the construction of state-specific college- and career-readiness assessments in ELA, mathematics, and science. The ELA and mathematics banks are well-populated, with hundreds of items available for each grade. The science bank, discussed in Component B, is in the early stages of development, but will be available for test administration in spring 2018. ICCR ELA and mathematics items have been field tested or used operationally in at least one of five participating states: Arizona, Florida, Ohio, Tennessee, and Utah. We developed ICCR to provide states with a way to achieve their individual goals. These goals include the following:

- Measuring the breadth and depth of college and career-readiness standards, with the ability to do so immediately.
- Bolstering their own state-owned test with an immediate startup by embedding custom-developed, state-owned items in field test positions.

- Comparing the rigor of performance standards across states and comparing performance against those standards.

ICCR offers a hybrid between an off-the-shelf product and a custom test: our calibrated item bank can be used alone or in conjunction with custom-developed items to build a statewide testing system. The system is designed to be primarily delivered online, but it includes paper forms to meet individual needs while maintaining comparability to the online versions. ICCR-based assessments can be either adaptive (as in Utah), fixed-form (as in Florida), or staged-adaptive. The ICCR item bank has been so well received that it has been adopted for testing programs in Arizona, Florida, Ohio, Tennessee (2016–2017), Utah, and recently Wyoming. In total, about 4 million students a year test with ICCR items, more than test with PARCC or ACT Aspire.

In addition to ICCR, AIR has developed the following item banks:

- *SAGE.* AIR developed more than 7,000 ELA and mathematics items and 160+ writing performance task families for Utah's Student Assessment of Growth and Excellence (SAGE). AIR also worked with Utah educators to develop and review 5,000 new items for embedded field testing in Utah in 2016 and 2017. The SAGE item bank has been so well received that it has been licensed to testing programs in Florida, Arizona, Tennessee (2015–2016), and Ohio.
- *Smarter Balanced.* In addition to SAGE and ICCR development, AIR made significant contribution to the development of the Smarter Balanced item banks, which included creating stimuli and items for all mathematics claims (more than 1,000 mathematics items and performance tasks) and more than 3,000 ELA items (including performance task families of 20 items across four or five tasks each). AIR did more than create items for Smarter Balanced. AIR and its partners conducted approximately 700 cognitive labs in the Smarter Balanced states of California, Connecticut, Hawaii, Idaho, Iowa, Michigan, Oregon, Pennsylvania, and South Carolina. AIR helped develop the 2014 field test ELA and mathematics blueprints. When Smarter Balanced's item banking system was delayed and eventually cancelled, AIR provided our item banking system (ITS) as the bank of record for the practice, small-scale trial, pilot, and field tests, and continues to do so today.

Custom Development

In addition to Smarter Balanced, SAGE, and ICCR, AIR has developed ELA and mathematics assessment items and managed large development efforts for customized assessments in Arizona, Delaware, Florida, Hawaii, Minnesota, and Ohio. In these states, AIR has delivered approximately 5,000 new state-owned ELA and mathematics items each year and supported major efforts to align existing item banks to new standards (more than 20,000 items).

We offer several quantitative measures to support our history of meeting both quality and schedule requirements for item development. AIR has consistently experienced item acceptance rates above 90% in our state assessment contracts, at volumes of up to 15,000 items per year. AIR consistently exceeds expectations. For example, in Utah last year we had 1,138 approved items after completing all content and bias reviews. This exceeded the contract requirement of 810. We had extra items in both ELA and mathematics. Utah accepted the extra items against a future year deliverable and was able to field test the items a year early. AIR can document similar performance examples in recent years in Delaware and Ohio. AIR has more than 125 full-time content, editorial, and production specialists dedicated to item and test development activities.

Component B: Summative Assessment in Science

AIR has successfully delivered and continues to deliver online, adaptive science assessment programs in Oregon beginning in 2007; Delaware and Hawaii beginning in 2009; Ohio and Utah beginning in 2013; Connecticut, Washington, and West Virginia beginning in 2015; and Idaho starting this year. In California, we are currently conducting a statewide science pilot test for 1.2 million students for the spring 2017 window in grades 5, 8, and 10–12. As we describe below, AIR is the lead contractor supporting the broadest currently active multi-state Next Generation Science Standards (NGSS) item and test development effort and we deliver approximately 1.5 million science tests each year in 10 states.

AIR is in the final stages of completing the initial ICCR science bank, which will serve as the basis for an operational field test in spring 2018 for New Hampshire. ICCR science follows directly from AIR's work successfully developing and delivering science assessments across the country, in paper-pencil, dual-mode, transition to online, and 100% online programs.

We have completed the item specifications for ICCR science, based on the National Research Council's A Framework for K–12 Science Education (NRC, 2012). Starting in mid-2015, AIR has been engaged in supporting multiple states in various ways with Next Generation Science Standards (NGSS) development efforts, including Connecticut, Hawaii, Idaho, Oregon, Washington, West Virginia, and the U.S. Virgin Islands. Connecticut is now shepherding an effort to formalize a Memorandum of Understanding for item sharing. Building on the same Council of Chief State School Officers (CCSSO)/WestEd Science Assessment Collaborative item cluster concepts, AIR has worked directly with science educators in most of these states to create item clusters that meet NGSS criteria and support three-dimensional learning. Each standard combines practices, cross-cutting concepts, and disciplinary core ideas. The item clusters span the performance expectations in physical science, life science, earth and space sciences, and engineering, technology, and applications of science. Several states have pilot tested the first rounds of AIR development, and AIR has more than 75 additional NGSS clusters in various stages of development and review. For this new NGSS-focused development, we have created stand-alone items, mini clusters that have one stimuli, and full multi-stimuli clusters. The items have addressed Matter and Its Interactions; Motion and Stability: Forces and Interactions; Energy, Waves, and Their Applications in Technologies for Information Transfer; From Molecules to Organisms: Structures and Processes; Ecosystems: Interactions, Energy, and Dynamics; Heredity: Inheritance and Variation of Traits; Biological Evolution: Unity and Diversity; Earth's Place in the Universe; Earth's Systems; Earth and Human Activity; and Engineering Design.

Beyond these NGSS endeavors, AIR has supported statewide science assessment programs in Delaware, Ohio, Oregon, and Utah.

Components C & D: Interim Assessments in ELA and Mathematics, and Science

AIR has delivered various sets of integrated interim assessments in ELA, mathematics, and science since 2007. AIR started delivering online, standards-based, adaptive interim or multiple-attempt summative assessments as part of the states' summative assessment systems in Oregon in 2007 and Delaware in 2009. This work continued with Utah starting in 2013, and in 14 Smarter Balanced States starting in 2014−2015.

As described in the Technical Proposal, AIR is proposing to build New Hampshire's interim assessments in ELA, mathematics, and science by licensing an existing item bank aligned to the Common Core State Standards (CCSS) and linked to the ICCR summative item bank. Options may include Utah's SAGE Interim Assessment Item Bank, which AIR helped Utah build into one of the largest and highest quality state-owned item banks in the country. As noted, Utah's SAGE summative item bank has been so well received that it has been licensed to testing programs in Arizona, Florida, Ohio, and Tennessee (2015–2016). The item bank was so robust that Utah was able to split the items between interim and summative assessment banks, leaving 3,800 items for interim assessments in ELA, mathematics, and science. These SAGE interim assessment item banks are already linked to the ICCR scale. Alternatively, AIR is working with other interim assessment item banks (our own and others we can license) in other states, and these banks will be linked to the ICCR scale in 2017–2018.

Component E: Reporting Portal

New Hampshire has used AIR's online reporting system (ORS) for the past three years, giving the state's educators the tools that go far beyond accessing standard, prepackaged reports and allow them to navigate student results in a manner that mirrors the instructional decision-making process. ORS is used in 24 states.

Beyond ORS, New Hampshire last year also began using AIRWays, AIR's interim/benchmark reporting system that provides teachers with more tools to track student progress throughout the year. Teachers and other users can see the items from the non-secure, fixed-form interim assessments, in addition to the actual student response, allowing them to pinpoint student strengths and weaknesses and gaps in student understanding.

General Description of Capabilities and Capacities for Key Project Activities

The RFP asks for general capabilities and capacities of AIR related to, among other areas, test development, administration of paper-based and online assessments, scoring, reporting, and psychometric activities. The following sections describe the structure and capabilities of AIR Assessment.

AIR Assessment includes more than 400 fulltime and 200 temporary staff members available to support New Hampshire and other state assessment clients. AIR's assessment work operates under the direction of Dr. Jon Cohen, AIR's Executive Vice President and President of Assessment; Steve Kromer, Senior Vice President and Chief Operating Officer of Assessment; and Heather Hayes, who directs all online testing projects at AIR. Dr. Cohen, Mr. Kromer, and Ms. Hayes have the authority to make corporate decisions related to this project. Dr. Cohen, Ms. Hayes, and other AIR senior managers are intimately involved in project operations and attend key planning meetings, progress meetings, and even technical meetings.

AIR Assessment Structure

We present the project team for New Hampshire. Here, we summarize the organizational structure and the depth of capability of our staff. AIR's corporate directors work directly on our state projects more than 90% of the time. AIR Assessment is organized functionally, and each functional area is led by a member of AIR Assessment's senior management team. Within each functional area and for each project, teams are formed and a team leader is designated. The team leader takes first-line responsibility for the team's products and services for the project, with ultimate responsibility resting with the director of the relevant functional area.

Dr. Cohen sets the vision and standards for the program; recruits, develops, and motivates staff; and ensures that needed infrastructure and procedures are in place. AIR's approach to managing personnel and staff ensures that there is a team to support each contract requirement and that required information never resides in a single person. Steve Kromer brings more than 25 years of assessment experience to his role as AIR Assessment's Senior Vice President and Chief Operating Officer. In this role, he is responsible for ensuring that repeatable processes are in place in each functional group and that we follow these processes in delivering each program. Given the specialized nature of our work and

AIR's unique approach to it, staff members are expected to master their capabilities through an apprenticeship system predicated on the gradual increase of individual responsibility. Group training sessions and collaborative interactions that encourage mentoring and supervision are at the core of this system. We achieve this goal by organizing work into structured teams, each with well-defined specifications and work processes: Project Management; Computer and Statistical Sciences Center (CSSC); Psychometrics and Statistics; Test Development; Operations and Scoring; and Communications and Reporting.

Project Management

Vice President:
Heather Hayes, J.D.

Project Manager:
Tom Glorfield, B.F.A.

Project Management

State assessment contracts are large and complex, involving many different skills, client management staff, school staff, vendors, systems, datasets, schedules, and other resources and stakeholders to operate efficiently.

Heather Hayes serves as AIR's Vice President of Assessment Programs and Client Services. She leads a team of 38 project management and operations staff members, 26 of whom hold master's or doctoral degrees. Ms. Hayes has more than 15 years of experience in online systems, including managing Oregon's online testing project and documenting and transferring best practices in online, multiple-administration testing in states including Delaware, Hawaii, Ohio, Oregon, and Utah. She received her J.D. from the University of Iowa.

To oversee and coordinate the multifaceted contract, AIR is proposing a seasoned and highly experienced program manager, Tom Glorfield. Mr. Glorfield has managed the deployment of online interim and summative assessments in five states. He is currently finishing the third year of Smarter Balanced testing in New Hampshire and Vermont, serving as AIR's main point of contact with the New Hampshire Department of Education and the Vermont Department of Education. In 2014–2015, he managed Maine's implementation of Smarter Balanced before that state left the consortium to develop its own assessment. In all three states, Mr. Glorfield had responsibility for the development and maintenance of interfaces between AIR and state data systems for information on test administrators, students, and test results. Mr. Glorfield has been a certified Project Management Professional (PMP) since 2005 and earned his B.F.A degree (with a minor in English) from Iowa State University.

Ms. Hayes and Mr. Glorfield will be completely familiar with the entire project and will participate in all meetings. Together, they will lead all internal management meetings with AIR staff, direct all higher-level communications with New Hampshire, and serve as a team for communication with other stakeholders. Mr. Glorfield will take the lead on coordinating schedules, producing minutes and other documentation, and implementing the project management and quality assurance procedures to ensure fidelity to the project's goals and policies.

Our Project Management team has the flexibility and the authority to authorize and execute change orders, make decisions, engage additional resources, and execute creative solutions to unusual or unforeseen problems.

CSSC/Online Testing

Vice President:
Selina Tolosa, M.B.A.

Technology Consultant:
Sonja Hubbard, M.S.

Computer and Statistical Sciences Center

The Computer and Statistical Sciences Center (CSSC)/Online Testing CSSC is responsible for both our software infrastructure and our software products, including the technology and support for AIR's student registration, online testing, reporting, learning management, and related systems.

CSSC takes as its mission the development of innovative products and systems that will keep AIR Assessment at the forefront of our field. CSSC combines a structured software development unit (developers and project directors) with a team of statisticians, psychometricians, and mathematicians. It is organized into groups supporting families of software systems:

- Pre-Production systems: Responsible for systems that support assessments before they are delivered, including test development, publication, and user management.

- Test delivery and reporting: Responsible for our high-demand, public-facing systems that deliver tests and reports, as well as other curricular and educational supports.
- Data analysis: Responsible for test data processing, analysis systems, and data deliveries.

Each family of systems has a technical lead, and together these technical leads form a committee responsible for overseeing peer review of key documents such as requirements, design documents, and test plans, as well as peer review of program code. CSSC's interdisciplinary team includes more than 160 staff members, 11 of whom have doctoral degrees in areas ranging from computer science to statistics and engineering. Eighty-nine of the software project directors and other engineers have MBAs and other advanced degrees.

Close interdisciplinary collaboration enables CSSC to develop remarkable technical products that effectively turn raw data into useful information and manage complex human processes to efficiently support test development, psychometrics, and score reporting. For more than nine years, Selina Tolosa, Vice President, Assessment Technology Services and Solutions, has served as team leader for the development and implementation of assessment software systems at AIR, including AIR's online test delivery system for operational tests in 20 states. Before joining AIR, Ms. Tolosa was the director of information services and the financial officer for New American Schools, a nonprofit organization funding the development of schoolwide designs aimed at transforming elementary and secondary schools. Ms. Tolosa also brings experience in financial forecasting and analysis, quality assurance of government enterprise systems, and the application of database skills, statistical sampling, and relational databases to data organization of assessment systems. Ms. Tolosa received her MBA from the Wharton School of Business, University of Pennsylvania.

AIR is proposing Sonja Hubbard, M.S., as the Technology Consultant. Ms. Hubbard has more than 12 years of experience as a software project manager and engineer, including more than three years at AIR working on data exchanges, single sign-on, and student registration systems in support of online testing and reporting. Her work with TIDE and OpenAM single-sign on make her the perfect candidate for New Hampshire's desired Technology Consultant role. She received her M.S. in information systems technology from George Washington University and her B.A. in media studies from Pomona College.

Psychometrics and Statistics

Vice President and Chief Psychometrician:
Gary Phillips, Ph.D.

Lead Psychometrician:
Ahmed Turhan, Ph.D.

Psychometrics and Statistics

AIR Assessment offers psychometric and statistical services that stand alone in terms of quality and innovation. The integration of psychometrics with statistics and sampling sets AIR apart from the competition.

Although testing firms often bring expertise in psychometrics, the quality of those services depends dramatically on the samples on which the data are based. Typical samples used in state testing programs can undercut the best psychometrics, leading to volatile test results from year to year and inaccurate classification of test-takers. AIR combines expertise in sampling and psychometrics; all of our samples are optimized, and our statistics accurately reflect the complexities of the sample designs.

Dr. Gary Phillips leads the Psychometrics and Statistics team, which includes 21 lead psychometricians with doctoral degrees and more than 20 statistical programmers, research associates, and other supporting staff. Dr. Phillips brings 26 years of experience, with distinctions such as having run the National Assessment of Educational Progress (NAEP) for the federal government and having served as acting commissioner of the National Center for Education Statistics (NCES). Prior to that, he served as the deputy commissioner of NCES. Previous responsibilities at NCES included overseeing the NAEP, the National Adult Literacy Survey (NALS), and the Third International Mathematics and Science Study (TIMSS). Dr. Phillips served as the architect and as the executive director of President Clinton's Voluntary National Test (VNT). He holds a Ph.D. from the University of Kentucky with an emphasis in statistics and psychometrics. He has published or presented more than 200 papers, taught dozens of advanced graduate-level statistics courses, and presented hundreds of workshops on advanced statistical and psychometric topics. His areas of interest are structural equation models, hierarchical linear models, item response theory (IRT) models, and Bayesian inference. He is nationally and internationally known for his expertise in large-scale assessments and complex surveys.

For Lead Psychometrician, we have proposed Dr. Ahmet Turhan. Dr. Turan has over 15 years of experience with large-scale assessment programs, including starting new national and statewide assessments. With AIR, he manages, plans, and supports form development, independent and operational field testing, data review, and various other psychometric tasks for state clients. Dr. Turhan served as a lead psychometrician for the Partnership for Assessment of Readiness for College and

Careers (PARCC) project, as well for the entire psychometric efforts for two major Florida Assessment programs. He received his Ph.D. in educational measurement statistics from Florida State University.

Across all projects, the Psychometrics and Statistics team is responsible for the following tasks:

- Sample design
- Field-test design
- Item analysis
- Analysis of differential item functioning
- Calibration, using both classical and IRT methods
- Parallel form equating and vertical linking of related tests
- Design and implementation of standard-setting sessions
- Design and implementation of special studies, including validity studies
- Randomized field trials
- Cross-form reliability studies
- Program and initiative evaluations

Test Development

Vice President:
June Zack, M.A.

Content Leads:
Jacob Wilkes, M.A. (ELA)
Rachel Azzerah, M.S. (Science)
Alyssa Kartee, (Mathematics)

Test Development

Developing more than 10,000 items annually, our Test Development team has been developing and aligning existing items to Common Core State Standards (CCSS) since its inception. This team is led by June Zack, M.A. Ms. Zack is AIR's Vice President of Assessment Content and Test Development and has led this team for the past 12 years, expanding the team to include more than 110 staff members (13 with doctoral degrees, 61 with master's degrees). Ms. Zack is primarily responsible for the development of items and the assembly of test forms for all of AIR's assessment projects. She is responsible for budget tracking, staffing, product development, and overall coordination of assessment development. Before joining AIR, Ms. Zack worked at ETS for 10 years, managing test development for numerous projects, including the GRE, GMAT, LSAT, SAT, NAEP, and Praxis assessments. Ms. Zack was also involved in project management for reading, writing, and mathematics development for the ISEE and the CTP IV. Ms. Zack holds an M.A. in educational psychology from New York University.

Our Test Development team is responsible for critical activities including the development of assessment frameworks, item and form development, and curricular content development. Across projects, the Test Development team is responsible for the following tasks:

- Framework development
- Test blueprint and item specification development
- Item development, review development, and review of scoring rubrics form construction
- Curricular content development
- Development of supporting materials for item release
- Client committee support and liaison with rangefinding services

Vice President:
Dee Wagner, M.B.A.

Operations and Scoring

Operations and Scoring

AIR's Operations and Scoring team includes 29 fulltime and more than 200 part-time professionals and is responsible for print test forms, production, and help desk operations, and for warehousing, distributing, collecting, security processing, scanning, editing, performance scoring, and preparing data files. The supporting software systems interface directly with operations relating to Pre-ID and ordering, student tracking, print tracking, and other AIR systems.

DeeAnn Wagner is Vice President of Assessment Programs and Operations. Ms. Wagner has more than 22 years of experience managing large, high-stakes assessment projects, including the past 13 years at AIR. Ms. Wagner has overseen test administration, security, and assessment operations, which include activities related to producing, delivering, receiving, scanning, and scoring test materials. She is also adept at managing AIR's subcontractors for scoring and processing online, paper-pencil, and dual mode testing projects. A Six Sigma Black Belt and ISO-Certified Lead Assessor, Ms. Wagner gets the most out of staff and consistently identifies and gives more responsibility to high performers. The results have been outstanding; AIR projects run on schedule, within budget, and to high client satisfaction. Ms. Wagner received her operations management (B.S.B.A.) degree from the University of Delaware and her MBA from Philadelphia University.

AIR will be responsible for all paper-pencil production, processing, and validation handscoring services for machine-scored writing performance tasks. Damon Hartzler will lead the performance scoring for the assessments. Mr. Hartzler has 18+ years of experience in the scoring field. He has several years of experience working for the Maryland State Department of Education on their statewide assessments for ELA, government, and mathematics. Currently, he leads the scoring of statewide assessments at AIR's Ohio Processing Center in Columbus. Mr. Hartzler collaborates with project managers and state department officials to ensure that the scoring of statewide assessments adheres to state standards. He holds B.A.s in both English and history from The Ohio State University.

Communications and Reporting

Vice President:
Bani Dheer, J.D.

Lead Reporting:
Margaret Won, M.A.

Communications and Reporting

AIR goes beyond simply reporting scores and other numbers to provide deep analyses of the data, reported in a clear, appealing, and actionable way. AIR is perhaps alone in the industry in delivering information, often in full color, with variable text and graphics. The reports begin with the same basic data that have been traditionally reported, but they extend much further. The basic data are reported using tested data displays that facilitate correct interpretation and impede misinterpretation. The implications (as well as limitations) of additional analyses, including the increasingly popular value-added analyses, are presented both graphically and in text that is customized for each individual report. Our professional development services focus on assessment administration and data use, helping our clients develop assessment literacy in their states.

Bani Dheer, J.D., is AIR's Vice President of Communications and Reporting, bringing 15 years of experience to this role. Ms. Dheer has led the development and on-time delivery of more than 20 million pages of score reports for AIR clients each year. Before joining AIR, Ms. Dheer designed and developed the messaging strategies for several national and state education groups, including the National Education Association (NEA), the Learning First Alliance, the Civil Society Institute, and the California Teachers Association. She also worked with lead Democratic strategist Mark Mellman to develop the communications strategy for a presidential campaign. Before entering the field of educational assessment and communications, Ms. Dheer worked for five years as managing editor for a Vancouver-based publishing group.

AIR is proposing Margaret Won, M.A., to serve as the lead for score reporting. With AIR, Ms. Won's work with Smarter Balanced clients has been in creating specifications documents, testing integration between systems, and performing quality control procedures. She manages a team of employees to ensure quality control throughout the development process for the online score reporting systems. Ms. Won received her M.A. from the Brandeis International Business School and her B.A. in economics from Wellesley College.

Vendors

AIR also engages many vendors for various services in support of projects. These include managed hosting services (Rackspace), printers for test materials and printed score reports (e.g., AGS), communications (e.g., multiple vendors for VOIP, WebEx, cell phones), shipping (e.g., USPS, FedEx), internal copying (e.g., Xerox, Océ), temporary employment agencies, supplies, and many other services required to support our staff and project work. Although not specifically procured for this project, AIR treats these agreements in the same manner described above at a corporate level. We procure these services after-careful analysis and/or competition, and develop formal contractual agreements.

Key AIR Facilities

These teams are supported by all of the required facilities. Beyond our typical office space and technology tools, AIR will support the paper-pencil processing and handscoring from our Columbus, Ohio facility. This 36,000-squarefoot facility has supported paper-pencil packaging, receipt and scanning, editing, and/or scoring for projects in Delaware, Hawaii, New Mexico, Ohio, Oregon, and South Carolina. The scanning system used by AIR employs a patent-pending technology to ensure accurate data capture, and our web-based handscoring system for constructed-response items integrates seamlessly with both our online test delivery and paper scanning systems. The handscoring software manages the scoring process with a broad range of configurable business rules that manage the workflow, assignment of scores, and quality procedures. As described in our technical approach, AIR supports application hosting with both external and internal resources. AIR hosts our ITS, TDS, reporting, and related critical assessment systems and websites in a managed hosting facility, because the scale and investment available from a dedicated hosting company far exceed what any one company can provide. We use several of Rackspace's managed hosting facilities to host our assessment systems. Rackspace is rated one of the "Magic Quadrant" leaders by Gartner Research.

Other Topic 28 Requirements

In addition to general description above, AIR presents the following responses to the other Topic 28 requirements

- AIR has been in business continuously as a not-for-profit corporation since 1946. AIR has no relevant buyouts, takeovers, IPO's, bankruptcies, litigations and claims, etc. within the last five (5) years.
- AIR has approximately 1,800 employees across 18 offices in the United States and 7 International offices. AIR does not have a physical presence in New Hampshire.
- Throughout the Technical Proposal and Appendices, AIR has provided examples of AIR's work products such as test and item specifications, items, forms, psychometric analysis plans, technical reports, and other deliverables. AIR commits that NH DOE will receive the same or better quality of work throughout the contract, including any extensions, as the examples we have provided in the proposal.

While AIR designs our systems and procedures to meet our clients' expectations of error-free performance, issues do arise, and AIR is proud of our record of performance in these situations. In all cases, AIR works closely with the client to remedy the situation, either with additional services or payment of liquidated damages. As our Hawaii client reported: "The key reasons that led to [Hawaii's] selection of AIR as the testing contractor … and why we have continued to collaborate with them over time include … honesty in addressing issues and providing solutions that support the integrity of the test administration process as well as the validity and reliability of student results."

Reportable Issues

Issue 1: Smarter Balanced Delayed Scoring. In spring 2015, AIR was committed to a two-week rolling scoring window, so students would receive their scores within two to three weeks of completing their test. This posed a unique challenge never tackled before in the industry. A typical, fixed-form test has only a small number of handscored items. Across grades, scorers may need to be trained for a total of about 100 items. The Smarter Balanced assessment is adaptive and has a very large item pool. Dozens or hundreds of handscored items might be delivered in each grade, so planning the scoring activity is enormously challenging. This challenge was compounded by the fact that this was the first operational administration of the Smarter Balanced assessment. Reporting was delayed in several states administering the Smarter Balanced adaptive assessments. Root cause analysis revealed that scorers were available at the wrong times, with too many early in the window and too few later; scorers were also allocated evenly across items, even though items were administered with quite divergent rates. In 2016, AIR met these scoring deadlines reliably, thanks to three improvements. First, having data from the previous year allowed us to know when students in each state were going to test. Second, a series of simulations indicated the exposure rate of each item. Together, these data enabled us to accurately predict workflows by item so that we could have the appropriate staff trained and in place to score the correct items. The third improvement is a new system that we built. Our Ledger system tracks tests and responses in real time by communicating with our online testing system and analysis systems. All items that will be handscored pass through Ledger, which tracks the progress of each individual response. Aggregate aging reports tell us which items are running behind so that resources can be reallocated to those items before deadlines are missed. Ledger also lets us retrieve the status of any test, or tests from any school, to better respond to inquiries from the field.

Issue 2: Login system unavailable for three hours in Florida in 2015. In April 2015, our login system experienced an outage that lasted approximately three hours, affecting testing in the state of Florida. Investigation revealed that the problem arose when a staff member, working on non-production systems, accidentally disturbed the database functioning of a production system. We replicate databases across systems for a variety of reasons. In one instance, the database on the replicated system required an additional index that was not required on the source system. When installing additional servers to support back-end activities, the staff member accidentally reinitialized the replication to our Florida login server. With new replication, the additional index was missing, slowing the server down. It took three hours to identify and fix the problem. Our root cause analysis revealed that the staff member had not abided by our production control process, and that our policies were ambiguous in this regard. Because the new servers being installed were not production servers, he did not believe that he needed to abide by our production control processes. In response, we clarified our policy—any work that touches the production environment must go through the formal production control process. That process would have had his plans reviewed, and critically, a second verifier would work alongside him to ensure that only the correct servers were touched. Formal post-implementation tests would have been implemented. Our policies were clarified, and several meetings with all staff who have access to the production environment were held to retrain them on the clarified policies. We repeat these trainings now at the beginning of each testing season.

Issue 3: In spring 2015, on March 3 and 5, AIR experienced Distributed Denial of Service (DDoS) attacks in Florida, lasting approximately 20 minutes. Testing continued for some users. Delays and failures occurred during the login process for some users. Schools and districts made local choices to change testing calendars. With the information from the attacks, AIR has been able to configure our Intrusion Detection Systems and has not experienced interruptions from any other DDoS attacks. Finally, AIR users occasionally experience local student and school interruptions during testing windows due to issues with individual computers and/or local school issues. AIR has designed our systems to include many safeguards to ensure that data is not lost and to protect the validity of the test even if computers fail, sessions are interrupted, or Internet service goes down.

For Issues 2 and 3, the Florida legislature engaged Alpine Testing Solutions to conduct a validity study on the administration of the Florida Standards Assessments (FSA). Alpine confirmed that the FSA is an accurate way to measure Florida’s students’ mastery of the standards. The report goes on to say that the results can be used for group-level decisions, so the department is moving forward with those activities, which include calculating school grades, calculating scores for inclusion in teacher evaluations and setting achievement-level cut scores. AIR did not think it was productive to debate whether the state was entitled to Liquidated Damages under the terms of the contract. AIR did agree, however, that it fell short of its own high standards of performance and agreed to pay for the independent Alpine report ($594,310) and reduce our invoices to Florida by $4,202,675, or approximately 8% of the annual value.

Issue 4: In 2016, AIR experienced two issues in Utah subject to Liquidated Damages. On Tuesday and Wednesday, March 25 and 26, 2014, AIR received reports of Utah students’ SAGE tests timing out. During both days, the majority of students were able to test with no issues. The initial problem on March 25 caused students to get an error when they opened the secure browser. This problem was quickly identified and resolved by failing over to redundant servers. The login errors were caused by a problem in how indexes were rebuilding on the server. The first help desk calls were received at 11:55 a.m. EST, and the problem was identified and resolved by 12:54 p.m. EST. At that time, all users who had contacted the help desk were called and informed that the problem was resolved and that the system was functioning normally. The next day, we encountered a second issue. We had reports of some students having trouble starting a new test, resuming an existing test, or receiving error messages when trying to save their work. The problem was sporadic; some students were able to test successfully. This is evidenced by the 8,572 tests completed and 14,592 tests successfully started that day. The root cause of both issues was an update made the previous weekend that led to missing indexes on one server. Standard AIR procedure requires deployments to be performed by two staff members, with one implementing a set of documented changes and the other independently checking those changes. A second quality assurance procedure uses an automated check of database configuration. Procedures were followed, but this index was omitted from the scripts and documents. AIR undertook a review of our deployment documents to identify and fix any similar omissions. AIR did not think it was productive to get into a debate as to whether Utah was entitled to liquidated damages ($20,000 under the terms of its contract). AIR did agree, however, that it fell short of its own high standards for performance and wanted to offer an accommodation that would provide additional value to Utah and allow the parties to put the matter behind them. AIR proposed, and the Utah State Board of Education accepted, a set of services to advance the SAGE program. This included, at no cost to USBE, the summer 2014 administration, reductions in test administration manual and training costs over three years, and support for a content review of writing prompts.

To support USBE's accountability goals, USBE and AIR agreed that AIR would post final data files to the SFTP site by September 9, 2014. Due to various reasons, science files were posted on September 11, mathematics files were posted on September 12, and ELA files were posted on September 15. AIR acknowledges these delays, paid $200,000 for this 2014 Late Return of Data.

Issue 5: In Arizona, AIR has provided $132,500 as compensation for the following issues:

- In 2015, AIR shipped Parent Guides at the same time as print reports as required, but the separate Parent Guide shipments were delayed, arriving two days after the reports.
- In spring 2016, AIR's TIDE Manual included incorrect information about the size of on-demand labels (due to a change in label stock) and the Test Coordinators Manual included incorrect information regarding FedEx return labels.
- In June 2016, AIR used the wrong file layout for test results. The correct file was posted three days later.
- In June 2016, AIR determined 15 ELA items had not been scored correctly for some of the possible response patterns for these items. Corrected scores were available in the ORS seven days later.

Topic 28.1 Relevant Experience

Relevant Experience Table

Exhibit D2.1-1 shows a list of similar AIR projects and includes the following columns required by the RFP:

- Client/program name
- Length of contract
- Content area, grades
- Number of students
- Administration mode (paper-pencil or computer-based), with total number of tests administered and the highest number of concurrent testers
- Use of scoring (human and artificial intelligence)
- Client contact information

Current Use of Vendor Proposed Software – Current Implemented Sites of Vendor Proposed Software

As described in Section E-1.1.4.1, AIR is proposing to use the same student registration, test delivery, online reporting, and related software systems that we deployed successfully in New Hampshire for the past three years, and in all of the states and jurisdictions listed in Exhibit D2.1-1.

Exhibit D2.1-1: Similar AIR Projects

Program	Contract Value	Length	Content Areas and Grades	Number of Students	Number of Online Tests	Peak Number of Concurrent Testers	Constructed Response Scoring	Summary	Contact information
New Hampshire Smarter Balanced Assessments #2014-018	$3.9M	2014–2016	ELA 3–8 Math 3–8	14,000 per grade	150,874	6,835	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced assessments for grades 3–8, 11 for SY15; removed grade 11 for SY16–17	New Hampshire Department of Education 101 Pleasant Street Concord, NH 033010 Scott J. Mantie, Ph.D., Administrator Bureau of Accountability and Assessment (603) 271-3844 scott.mantie@doe.nh.gov
Connecticut Smarter Balanced Assessments #155DE0018AA	$13.6M	2014–2017	ELA 3–8 Math 3–8	41,000 per grade	487,727	16,935	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced assessments for grades 3–8, 11 for SY15; removed grade 11 for SY16–17	Connecticut Department of Education 165 Capital Ave Hartford, CT 06145 Abe Krisst (860) 713-6860 abe.krisst@ct.gov
Connecticut Science Test Administration and NGSS Development #155DE0170AA	$15.3M	2015–2019	Science in grades 5, 8, and 10	41,000 per grade	116,708	Included above	Human Machine-scored Constructed Response	Administer, score, and report on existing CT science tests and develop and pilot NGSS items for fully operational NGSS assessments beginning in 2018–2019	Connecticut Department of Education 165 Capital Ave Hartford, CT 06145 Jeff Greig (860) 713-6860 jeff.greig@ct.gov
Arizona State Assessment #ADED-083299	$55.8M	2014–2017	ELA 3–11 Math 3–11	89,000 per grade	1,031,600	70,305	Human Machine-scored Constructed Response Machine-scored Essay with Human Validation	Test development and full-year online adaptive summative ELA and math assessments for grades 3–8 and 11 and EOC	Arizona Department of Education 1535 West Jefferson Street Phoenix, AZ 85007 Audra Ahumada Deputy Associate Superintendent (602) 542-5423 Audra.Ahumada@azed.gov
Oregon Assessment of Knowledge and Skills 1#ODE1007-07 2#ODE9573	1$31M 2$24M	2007–2014 2014–2024	ELA 3–11 Math 3–11 Science in grades 5, 8, and 11	45,000 per grade	787,989	16,325	Human Machine-scored Constructed Response	Full-year, multiple-opportunity, online adaptive summative and interim assessments for all students, including English Language Proficiency Assessments (ELPA through 2015; ELPA21 2015–2024)	Oregon Department of Education 255 Capitol Street NE Salem, OR 97338 Kathleen Vanderwall, Manager, Test Design & Implementation Oregon Department of Education (503) 947-5600 kathleen.vanderwall@state.or.us
Hawai'i State Assessment System #CO-60111	$75M	2006–2014	ELA 3–11 Math 3–11 Science in grades 4, 8, and 11	15,000 per grade	288,630	15,593	Human Machine-scored Constructed Response	Full-year, multiple-opportunity, online adaptive summative and interim assessments for all students (2010–2014); paper-pencil testing (2006–2010)	Hawai'i Department of Education 1390 Miller Street Honolulu, HI 96813 Tom Saka, Director (808) 586-3288 tom_saka@notes.k12.hi.us

Exhibit D2.1-1: Similar AIR Projects (continued)

Program	Contract Value	Length	Content Areas and Grades	Number of Students	Number of Online Tests	Peak Number of Concurrent Testers	Constructed Response Scoring	Summary	Contact information
Hawai'i End-of-Course Assessments #CO-10361	$5M	2012–2015	Algebra I Algebra II Writing U.S. History	15,000 per grade	22,624	Included above	Human Machine-scored Constructed Response	Online adaptive summative assessments for all students	Hawai'i Department of Education 1390 Miller Street Honolulu, HI 96813 Tom Saka, Director (808) 586-3288 tom_saka@notes.k12.hi.us
Hawai'i State Alternate Assessments #CO-10328	$4.9M	2012–2016	ELA 3–8, 11 Math 3–8, 11 Science in grades 4, 8, 11	300 per grade	4,123	N/A		Task-based adaptive assessment for students with disabilities	Hawai'i Department of Education 1390 Miller Street Honolulu, HI 96813 Tom Saka, Director (808) 586-3288 tom_saka@notes.k12.hi.us
Hawai'i Smarter Balanced, State Science and EOC Assessments #CO10551	$14.6M	2014–2017	ELA 3–11 Math 3–11	15,000 per grade	189,666	15,593	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced assessments for grades 3–8, 11 Develop, administer, score, and report science assessments and EOC assessments	Hawai'i Department of Education 1390 Miller Street Honolulu, HI 96813 Tom Saka, Director (808) 586-3288 tom_saka@notes.k12.hi.us
Utah Student Assessment #136199	$39M	2012–2017	ELA 3–11 Math 3–11 Science in grades 4–8 and EOC exams in Earth	50,000 per grade	1,152,109	30,612	Human Machine-scored Constructed Response Machine-scored Essay with Human Validation	Online adaptive interim and summative assessments for all students	Utah Department of Education 250 East 500 South P.O. Box 144200 Salt Lake City, UT 84114-4200 Jo Ellen Shaeffer, Director (801) 538-7811, JoEllen.Shaeffer@schools.utah.gov
California CAASP (sub to ETS)	[1]$15.2M [2]$20.5M [3]$77M	2013–2014 2014–2015 2015–2018	ELA 3–11 Math 3–11 Science in grades 4–8, and EOC exams in Earth Science, Biology, Physics, and Chemistry	480,000 per grade	6,563,452	334,862	Not scored by AIR	[1-2]Administer online adaptive summative and interim Smarter Balanced Assessments for grades 3–8, 11 [3]Administer online adaptive summative and interim Smarter Balanced Assessments for grades 3–8, 11; online CA Alternate, NGSS, and Primary Language assessments	Educational Testing Services 600 Rosedale Road Princeton, NJ 08541 USA Tom Foster, CAASP Program Manager (916) 403-2402 tfoster@ets.org

Exhibit D2.1-1: Similar AIR Projects (continued)

Program	Contract Value	Length	Content Areas and Grades	Number of Students	Number of Online Tests	Peak Number of Concurrent Testers	Constructed Response Scoring	Summary	Contact information
'lorida Standards Assessment '14-652	$133M	2014–2017	ELA 3–10 Math 3–11	210,000 per grade	2,973,680	218,596	Human Machine-scored Constructed Response Machine-scored Essay with Human Validation	Test development and full-year online adaptive summative ELA and math assessments for grades 3–8 and 11 and EOC	Florida Department of Education Office of Assessments 325 W. Gaines St., Turlington Bldg, #404 Tallahassee, FL 32399 Vince Verges, Asst Deputy Commissioner (850) 245-0513 vince.verges@fldoe.org
'ermont Smarter Balanced ssessments '28388	$4M	2014–2017	ELA 3–11 Math 3–11	7,000 per grade	84,092	3,877	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced assessments for grades 3–8, 11	Vermont Agency of Education 219 North Main Street, Suite 402 Barre, VT 05641 Michael Hock, Director of Educational Assessment (802) 479-1288 michael.hock@state.vt.us
'aine Smarter Balanced ssessments '20140905*856	$2.4M	2014–2015	ELA 3–11 Math 3–11	14,000 per grade	180,000	10,000	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced assessments for grades 3–8, 11 Administer Interim Smarter Balanced assessments for grades 9 and 10	Maine Department of Education 23 State House Station Augusta, ME 04333 Charlene Tucker (270) 624-6827 charlene.tucker@maine.gov
'ontana Smarter Balanced ssessments (sub to Measured 'rogress)	$396k	2015–2016	ELA 3–11 Math 3–11	12,000 per grade	132,926	5,020	Not scored by AIR	Administer, machine score, and report online adaptive summative and interim Smarter Balanced assessments for grades 3–8, 11	Measured Progress 100 Education Way Dover, NH 03821
daho Smarter Balanced ssessments '125023	$10.3M	2014–2017	ELA 3–11 Math 3–11 Grades 5, 7 and EOC Biology and Chemistry	23,000 per grade	324,239	12,868	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced Assessments for grades 3–8, 11 Develop, administer, score, and report Smarter Balanced EOC assessments for grades 9 and 10	Idaho State Department of Education 650 West State Street Boise, ID 83720-0027 Cheryl Finley, Director of Assessment and Accountability (208) 332-6976 cfinley@sde.idaho.gov
outh Dakota Smarter alanced Assessments '2015C-229	$4.1M	2014–2017	ELA 3–11 Math 3–11	11,000 per grade	145,330	11,159	Human Machine-scored Constructed Response Machine-scored Essay with Human Validation	Administer, score, and report online adaptive summative and interim Smarter Balanced Assessments for grades 3–8, 11	South Dakota Department of Education 800 Governors Drive Pierre, SD 57501-2294 Jan Martin, Administrator (605) 773-3246 jan.martin@state.sd.us

Exhibit D2.1-1: Similar AIR Projects (continued)

Program	Contract Value	Length	Content Areas and Grades	Number of Students	Number of Online Tests	Peak Number of Concurrent Testers	Constructed Response Scoring	Summary	Contact information
U.S. Virgin Islands Smarter Balanced Assessments #PC063DOE15	$1M	2014–2017	ELA 3–11 Math 3–11	1,000 per grade	15,487	831	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced Assessments for grades 3–8, 11 Develop, administer, score, and report Smarter Balanced EOC assessments for grades 9 and 10	U.S. Virgin Islands Department of Education Alexandria Baltimore-Hookfin, State Assessment Director (340) 773-1095 abaltimore-hoofkin@stx.k12.vi
North Dakota Smarter Balanced Assessments (sub to Measured Progress)	$304k	2015–2016	ELA 3–11 Math 3–11	9,000 per grade	107,317	4,244	Not scored by AIR	Administer, score, and report online adaptive summative and interim Smarter Balanced assessments for grades 3–8, 11	Measured Progress 100 Education Way Dover, NH 03821
Washington Statewide Assessment Program #20150204	$72.9M	2014–2017	ELA 3–11 Math 3–11 Science in grades 5 and 8 and EOC Biology	82,000 per grade	1,419,791	68,347	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced Assessments for grades 3–8, 11 Develop, administer, score, and report EOC alternate assessments, and state-specific science assessments; added ELPA21 administration for SY16	State of Washington Office of the Superintendent 600 Washington Street South Olympia, WA 98504 Tony Alpert, Chief Operating Officer (503) 622-9270 tony.alpert@smarterbalanced.org
West Virginia Assessment Testing #20150204-EDD1500000004	$14.8M	2014–2017	ELA 3–11 Math 3–11 Science in grades 4, 6, and 10	22,000 per grade	500,285	21,488	Human Machine-scored Constructed Response	Administer, score, and report online adaptive summative and interim Smarter Balanced Assessments for grades 3–8, 11 Develop, administer, score, and report EOC and state-specific science and social studies assessments	West Virginia Department of Education Office of Assessment and Accountability 1900 Kanawha Blvd E, Bldg 6, Rm 825 Charleston, WV 25305 Vaughn Rudy (304) 558-2546 vrhudy@k12.wv.us
Smarter Balanced Assessment Consortium 1Pilot Items #5043.0 2Test Delivery System #20130180 3Test Admin #20130205	1$5.5M 2$28M 3$2.7M	2012–2014	ELA 3–11 Math 3–11		6 million in 20 states in 2014			1Providing communication and recruiting support, test administration manuals, training modules, webinars, and Tier 1 help desk services for about 1 million students 2Administering online adaptive pilot and field tests for about 6 million students in 20+ states, developing and delivering opensource test delivery system, Tier 2 and Tier 3 help desk services 3Developing pilot and field-test items for Smarter Balanced, opensource machine scoring engines, small-scale tryouts	State of Washington Office of the Superintendent 600 Washington Street South Olympia, WA 98504 Tony Alpert, Chief Operating Officer (503) 622-9270 tony.alpert@smarterbalanced.org

Exhibit D2.1-1: Similar AIR Projects (continued)

Program	Contract Value	Length	Content Areas and Grades	Number of Students	Number of Online Tests	Peak Number of Concurrent Testers	Constructed Response Scoring	Summary	Contact information
Delaware Comprehensive Assessment System #DOE2010-02	$29M	2009–2019	ELA 3–8 Math 3–8 Science in grades 5, 8, and 10 EOCs in Algebra II and Integrated Math III	10,000 per grade	178,946	5,647	Human Machine-scored Constructed Response	Full-year, multiple-opportunity, online adaptive summative, interim, and formative assessments for all students	Delaware Department of Education 401 Federal Street, Suite #2 Dover, DE 19901-3639 Brian Touchette, Acting Director, Accountability Resources (302) 735-4090 brian.touchette@doe.k12.de.us
Delaware Comprehensive Assessment System Alternate Assessment #DOE-C11-42	$7.9M	2011–2016	ELA 3–11 Math 3–11 Science in grades 3–8, 10 Social Studies in grades 4, 7, and 9	400 per grade	3,681	Included above		Task-based adaptive assessment for students with disabilities	Delaware Department of Education 401 Federal Street, Suite #2 Dover, DE 19901-3639 Brian Touchette, Acting Director, Accountability Resources (302) 735-4090 brian.touchette@doe.k12.de.us
Ohio Achievement Assessments [1]K–5: #CSP0A02021 [2]6–8: #CSP0A03031 [3]OTELA: #CSP906007	[1]$114M [2]$76M [3]$312M	2002–2005	ELA 3–8 Math 3–8 Science 5, 8	140,000 per grade	Paper-based	N/A	Human	Full paper-pencil testing program with online adaptive formative assessments, online pilot and field tests of science and social studies assessments, K–2 assessments	Ohio Department of Education 25 South Front Street Columbus, OH 43215-4183 Jim Wright, Director (614) 466-0233 james.wright@education.ohio.gov
Ohio Graduation Tests #CSP901307	$258M	2006–2015	ELA, Writing, Math, Science	140,000 per grade	Paper-based	N/A	Human	Full paper-pencil testing program with online adaptive formative assessments	Ohio Department of Education 25 South Front Street Columbus, OH 43215-4183 Jim Wright, Director (614) 466-0233 james.wright@education.ohio.gov
Ohio Computer-Based Assessments #CSP903714	$172.9M	2013–2018	ELA 3–11 Math 3–11 Science 5, 8 Physical Science Biology Social Studies 4, 6 American History American Government	140,000 per grade	2,517,220	90,941	Human Machine-scored Constructed Response Machine-scored Essay with Human Validation	Full-year, online fixed form summative assessments for science and social studies assessments; beginning in SY15-16 ELA and math	Ohio Department of Education 25 South Front Street Columbus, OH 43215-4183 Jim Wright, Director (614) 466-0233 james.wright@education.ohio.gov

Exhibit D2.1-1: Similar AIR Projects (continued)

Program	Contract Value	Length	Content Areas and Grades	Number of Students	Number of Online Tests	Peak Number of Concurrent Testers	Constructed Response Scoring	Summary	Contact information
Ohio Alternate Assessments #CSP906614	$8.8M	2014–2018	ELA 3–8, 11 Math 3–8, 10 Science 5, 8, 10 Social Studies 4, 6, 10	3,000 per grade	59,192	N/A		Portfolio-based alternate assessments for students with disabilities (2004–2012); transition to task-based adaptive assessment (2013–2016)	Ohio Department of Education 25 South Front Street Columbus, OH 43215-4183 Jim Wright, Director (614) 466-0233 james.wright@education.ohio.gov
Ohio English Language Proficiency Assessment #CSP903416	$2.3M	2015–2018	K–11	5,000 per grade	53,475	Included above		Administer, score, and report online and paper assessments for English Language Proficiency Assessment	Ohio Department of Education 25 South Front Street Columbus, OH 43215-4183 Jim Wright, Director (614) 466-0233 james.wright@education.ohio.gov
South Carolina Alternate Assessment 1#04-S6592-A10743 2#4400002282	Program 1 - $9M Program 2 - $6M	2004–2015	ELA 3–8, 10 (through 2014) Math 3–8, 10 (through 2014) Science in grades 3–8, 10 (through 2015), grade 11 (through spring 2016) Social Studies in grades 3–8	500 per grade	Paper-based			Task-based adaptive assessments for students with disabilities	South Carolina Department of Education 1429 Senate Street, Suite 200 Columbia, SC 29201 Leslie Dawes, Program Coordinator (803) 734-4944 ldawes@ed.sc.gov
New Mexico Alternate Performance Assessment 1#06924P5270157 2#11924P5270016 3#92400-0000051817 4#92400-0000056962	Program 1 - $8.3M Program 2 - $5.5M Program 3 - $1.2M Program 4 - $5.6M	2006–2020	ELA grades 3–8, 10–11 (through 2014) NCSC (starting 2016)	400 per grade	Paper-based			Task-based adaptive assessment for students with disabilities	New Mexico Public Education Department 300 Don Gaspar Ave, Room 109 Santa Fe, NM 87501 Fran Lucero, Staff Manager (505) 827-5861 frances.Lucero@state.nm.us

Exhibit D2.1-1: Similar AIR Projects (continued)

Program	Contract Value	Length	Content Areas and Grades	Number of Students	Number of Online Tests	Peak Number of Concurrent Testers	Constructed Response Scoring	Summary	Contact information
Wyoming Alternate Assessments	$2.1M	2014–2017	ELA 3–11 Math 3–11 Science in grades 4, 8–11	150 per grade	Paper-based			Task-based adaptive assessment for students with disabilities	Wyoming Department of Education 2300 Capitol Ave Hathaway Bldg, 2nd Floor Cheyenne, WY 82002 Deb Lindsey (307) 777-8753 deb.lindsey@wyo.gov
Iowa English Language Proficiency Assessment #008316	$12M	2015–2017	K–11	3,000 per grade	28,783	468		Administer, score, and report online and paper assessments for English Language Proficiency Assessment	Iowa Department of Education 400 East 14th Street Des Moines, IA 50319 Colleen Anderson (515) 281-3249 colleen.anderson@iowa.gov

Reference

National Research Council, Division of Behavioral and Social Sciences and Education, Board on Science Education, Committee on a Conceptual Framework for New K-12 Science Education Standards. (2012). A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: National Academies Press.

D2.2 Project Management

A thorough description of AIR's approach to project management, including procedures and processes for scheduling, communication (both internal and with the Department), and coordination across tasks and parties that AIR uses to ensure the effective and efficient management of all assessment projects is provided in D2.3 Project Plan. Topic 39 Support Center outlines AIR's approach to interactions with districts and schools.

AIR is not proposing any subcontractors for this effort.

Topic 29 Management Team

In Topic 28 Vendor Experience, we introduced the structure of AIR Assessment and the capabilities of each of our teams. In this section, we identify all named staff involved in the New Hampshire assessments. The project organization chart with the named AIR personnel is presented in Exhibit D2.2-1. Resumes can be found in Appendix A.

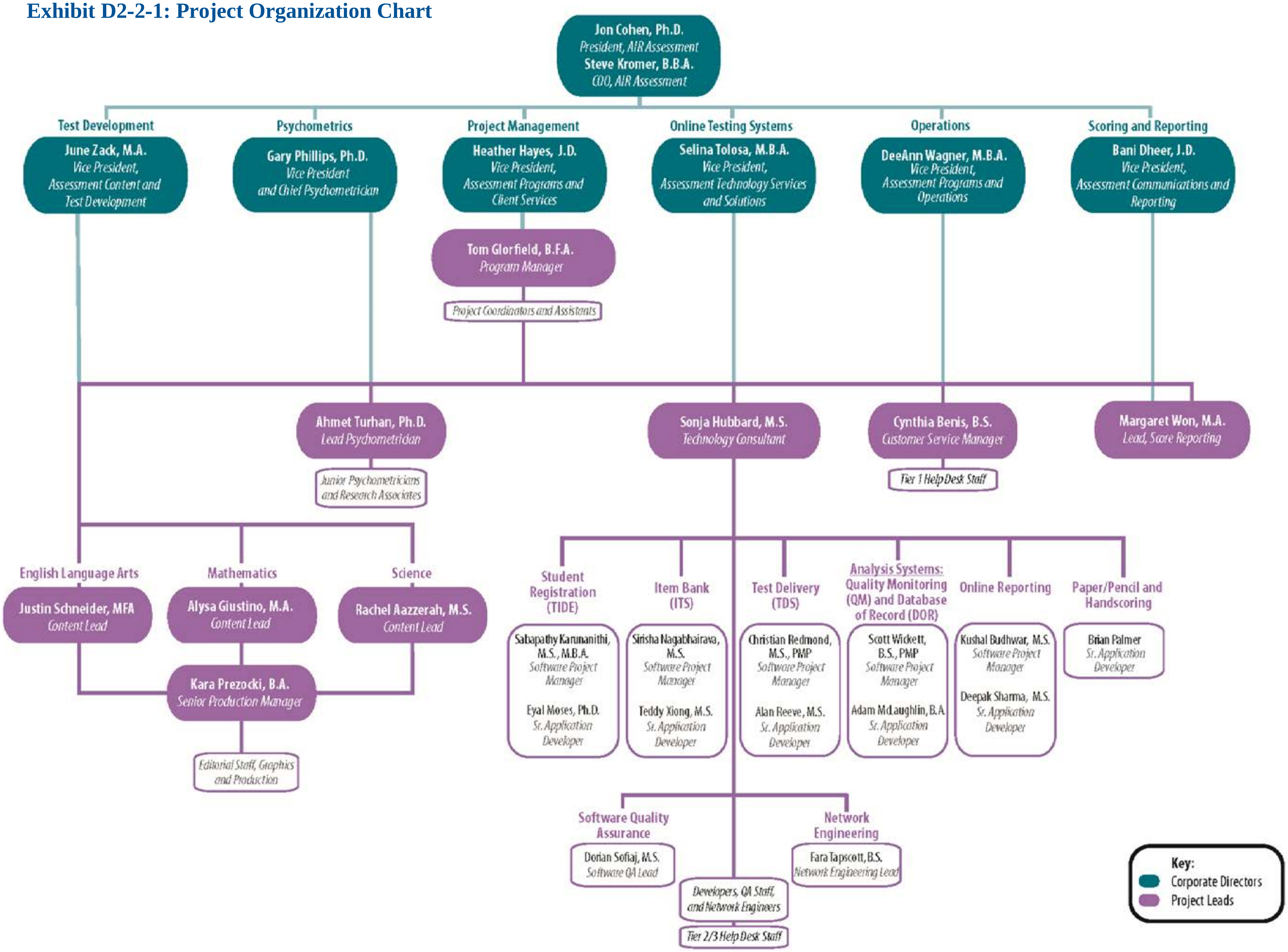
- Tom Glorfield, B.F.A., PMP, Program Manager, has managed the successful delivery of online testing in New Hampshire and other states for the past three years.
- Sonja Hubbard, M.S., Technology Consultant, has the technology experience with AIR's data exchanges, single sign-on, and student registration systems in support of online testing to work with NH DOE and district technology officials, and the communications background and experience to communicate with New Hampshire users.
- Ahmet Turhan, Ph.D., Lead Psychometrician, has over 15 years of experience with large-scale assessment programs, including working with ICCR in Florida, serving as the lead psychometrician for the Partnership for Assessment of Readiness for College and Careers (PARCC) project, and supporting older assessment projects in Florida and Mississippi.
- Content Leads, including Justin Schneider, MFA (ELA), Alysa Giustino, M.A. (mathematics), and Rachel Aazzarah, M.S. (science), who have experience with developing new statewide assessments using ICCR and other item banks proposed, as well as importing existing item banks from states.
- Software Project Managers and Senior Application Developers who have proven their ability to deploy AIR's industry-leading item banking, student registration, test delivery, scoring, and online reporting systems for more than 25 states, including the New Hampshire Smarter Balanced ELA and mathematics assessments for the past three years.

AIR's New Hampshire-specific staff will be supported by a large core team at AIR. New Hampshire-specific requirements may include interfaces with existing data systems, a dedicated portal, designated testing windows, user acceptance testing, and training plans that fit within New Hampshire's existing training infrastructure. At the same time, New Hampshire will benefit from the teams at AIR that are already dedicated to statewide assessments. For example, dedicated AIR technical project managers and software engineers configure and operate each of our major systems (student registration, test delivery, reporting, etc.) for all of our online testing clients.

In addition to highly qualified staff, our management approach relies on an extensive system of planning and specifications documents. It is not enough for all participants in the process to hear the same decisions in a meeting; decisions must be documented in a place where they can be easily found, referenced, reviewed, and considered. For this reason, each aspect of the testing program is governed by a specification or planning document. This approach is discussed in Topic 31 Work Plan.

In Appendix S, we have included each staff member's assigned responsibilities and time allocated to the project. All of our staff are available for the hours proposed. We have not assigned any individual beyond full time.

Exhibit D2-2-1: Project Organization Chart



New Hampshire Named Staff

Project Management

Tom Glorfield, B.F.A., PMP, Program Manager, has managed the deployment of online interim and summative assessments in five states. He is currently finishing the third year of Smarter Balanced testing in New Hampshire and Vermont, serving as AIR's main point of contact with the New Hampshire and Vermont Departments of Education. In 2014–2015, he managed Maine's implementation of Smarter Balanced before that state left the consortium to develop its own assessment. In all three states, Mr. Glorfield had responsibility for the development and maintenance of interfaces between AIR and state data systems for information on test administrators, students, and test results. In addition to this Smarter Balanced implementation work, Mr. Glorfield served as project director for the startup year of AIR's subcontracts with Measured Progress (MP) to deliver Smarter Balanced interim and summative assessments with Montana and North Dakota. For this work, Mr. Glorfield worked through MP and with state officials to establish the interfaces for single sign-on, student registration, accommodations, test delivery and reporting, as well as new interfaces between AIR's testing systems and MP's handscoring systems. Mr. Glorfield has been a certified Project Management Professional (PMP) since 2005 and earned his B.F.A. degree (with a minor in English) from Iowa State University.

In addition to the named staff below, Mr. Glorfield will be supported by project managers/coordinators and assistants for production and control of the schedule, specifications, and tracking documents, as well as regular project conference calls and meetings.

Ahmet Turhan, Ph.D., Lead Psychometrician, has over 15 years of experience with large-scale assessment programs. With AIR, he manages, plans, and supports form development, independent and operational field testing, data review, and various other psychometric tasks for state clients. Before joining AIR, Dr. Turhan was a senior research scientist with Pearson, serving as a lead psychometrician for the Partnership for Assessment of Readiness for College and Careers (PARCC) project, as well for the entire psychometric efforts for two major Florida assessment programs: The Florida Comprehensive Assessment Test (FCAT) 2.0 and Florida End-of-Course (EOC) Assessments. Dr. Turhan also served as a research scientist for the Mississippi Subject Area Testing Program, 2nd Edition (SATP2) large-scale high-stakes assessment program, conducting psychometric work, including calibration, equating, and scaling of both field and operational test forms using three-parameter logistic (3-PL) model for multiple-choice items. Dr. Turhan received his Ph.D. in educational measurement statistics from Florida State University.

Dr. Turhan will be supported by operational psychometricians, statisticians, data analysts, and research associates/assistants for executing the technical plans, conducting data review and standard setting meetings, producing technical report sections, and leading related activities. AIR's psychometric and statistics team is experienced with providing essential support throughout the project life cycle, from forms development and item analyses to standard setting and reporting. For example, they review forms and perform actual testing using strict protocols before deployment or distribution of assessments. They support all aspects of standard setting and production of score reports and technical reports.

Test Development

Our named test development content leads will be responsible for recording item and test specifications, reviewing item banks, selecting items, building forms, conducting committee meetings, and assisting our psychometricians in data review and standards setting. They lead a large cadre of test development specialists and item editors.

Justin Schneider, M.F.A., is an English language arts test developer in AIR's Assessment Program. Over the past two years, Mr. Schneider has assisted in the development of assessment items and performance tasks for the Smarter Balanced Assessment Consortium, as well as items and passages for clients including the states of Arizona, Florida, Ohio, and Utah. Before working at AIR, Mr. Schneider spent two years as a teaching fellow at the Catholic University of America, providing instruction in writing and rhetoric as well as serving as an undergraduate academic advisor. He holds an M.F.A. and MLitt in Shakespeare and Renaissance literature performance from Mary Baldwin University, and B.A.s in philosophy and English literature from American University.

Alysa Giustino, M.A., Mathematics Content Lead. Ms. Giustino has over six years of experience in mathematics education. During her time with AIR, she has served as an item developer for various AIR projects including Arizona, Florida, Ohio, Utah, and the Internal ICCR project. She has worked as a grade-band lead for the Florida Standards Assessments, and grade-level lead for the Internal ICCR project. As a grade-band lead, Ms. Giustino led an internal team of content specialists to develop over 200 online test items each cycle, aligned to the Florida blueprint and elementary content standards. Ms. Giustino led meetings including item review, rubric validation, test construction, standard setting, and achievement level descriptors review in Florida attended by Department of Education representatives and Florida educators. Ms. Giustino also assisted in the configuration of the Smarter Balanced Assessment Consortium Field Test and Practice Test for Mathematics and ELA. Prior to joining AIR, Ms. Giustino taught two years of summer school mathematics and worked as a substitute teacher at various grade levels for three years. Ms. Giustino received her B.A. in mathematics with a concentration in secondary education from St. Joseph’s College, and her M.A. in mathematics from Stony Brook University.

Rachel Aazzerah, M.S., Science Content Lead. Ms. Aazzerah brings over 15 years of science item and test development experience, most recently as the Science and Social Sciences Assessment Specialist at the Oregon Department of Education (ODE). At ODE, she was in charge of item development for the Oregon Assessment of Knowledge and Skills (OAKS) for science and social sciences. She was in charge of item writer training sessions for teachers, content review committee meetings, rangefinding and standard setting meetings, bias and sensitivity review meetings, accessibility review meetings, and rubric validation meetings. She also provided professional development for Oregon educators on both the Next Generation Science Standards (NGSS) and Smarter Balanced. At ODE, she also co-led the adoption of the NGSS as the 2014 Oregon Science Standards and participated in both public and private reviews of the Framework for K–12 Science Education and NGSS before they were finalized and released to the public in 2013. She has been involved with the Council of Chief State School Officers State Collaborative on Assessment and Student Standards (SCASS) for Science and has developed formative, interim, and summative assessment item clusters and tasks that are aligned to NGSS. Before her five-year tenure at ODE, she was a middle and high school educator for 13 years and holds an Oregon Professional Teaching License with endorsements to teach Biology, Chemistry, Integrated Science, Mathematics, Middle School Science, and Physics at the K–12 level. She is now responsible for managing a seven-member team of Next Generation Science Standards (NGSS) item and test developers at AIR. This involves overseeing the development of item clusters, task types, item specifications, and test blueprints that measure the breadth and depth of the three dimensions of the NGSS (Science and Engineering Practices, Disciplinary Core Ideas and Crosscutting Concepts). Ms. Aazzerah received her B.S. in chemistry and M.S. in science and mathematics education from Oregon State University.

Our test development leads are supported by 100+ content-specific test development specialists, along with the following dedicated staff:

- Operations Forms Managers. While AIR’s test development staff members are excellent at building forms, we have found that the physical form-construction efforts require dedicated forms managers. This role includes responsibility for the quality control of all paper test materials and online forms, as well as managing the schedule of internal reviews and client reviews in order to meet all print and online review deadlines.
- Editorial, Graphics, and Production Specialists. There are two specific roles for editorial, graphics, and production staff. The first is in the development of items, and for this work, AIR’s ELA, mathematics, and science teams include dedicated editorial staff experts in their respective content areas, along with free-hand computer graphics specialists for item graphics. The second role involves producing error-free paper and online forms and some other related materials (manuals, etc.). As described in the technical proposal, AIR’s systems are designed to

produce forms directly from our ITS, but a human eye is still required to finalize these forms and materials. AIR has a pooled team of editorial, graphics, and production staff for these tasks.

The RFP asks for a Special Populations Consultant. AIR has available expert staff in both accessibility of testing and in developing alternate assessments for students with disabilities. For example, Matthew Greathouse, M.Ed., has developed and delivered assessments and training on AIR's alternate projects in Delaware, Hawaii, New Mexico, Ohio, South Carolina, and Wyoming. Mr. Greathouse also reviews and assists in the development of alternate assessment items at multiple levels in the development process. Additionally, he trains teachers and administrators in the use and scoring of state assessments. Mr. Greathouse is AIR's representative at the ASES/SCASS meetings in order to stay up-to-date on any information or trends that are in the field. Mr. Greathouse received his B.A. from Virginia Wesleyan College, his M.Ed. in special education from The College of William and Mary, and his M.Ed. in educational leadership from George Mason University.

Information Technology

Selina Tolosa, MBA, Vice President, Assessment Technology Services and Solutions. Ms. Tolosa leads AIR's Computer and Statistical Sciences Center (CSSC) and has served as team leader for the development and implementation of assessment software systems at AIR for eight years, including AIR's online test delivery engine (TDS) and Pre-ID and student tracking for all of AIR's clients. Before joining AIR, Ms. Tolosa was the director of information services and the financial officer for New American Schools, a nonprofit funding the development of schoolwide designs aimed at transforming elementary and secondary schools. Ms. Tolosa brings experience in financial forecasting and analysis; quality assurance of government enterprise systems; and the application of database skills, statistical sampling, and relational databases to data organization of assessment systems. Ms. Tolosa received her MBA from the Wharton School of Business, University of Pennsylvania.

AIR is proposing to implement our proven test delivery, item banking, form construction, student registration, and related systems to support New Hampshire. These systems will be configured and deployed for New Hampshire by the following team. Each software system at AIR has a lead software project manager and a lead or senior application developer.

Sonja Hubbard, M.S., Technology Consultant. Ms. Hubbard has more than 12 years of experience as a software project manager and engineer, including more than three years at AIR working on data exchanges, single sign-on, and student registration systems in support of online testing and reporting. Her work with TIDE and OpenAM single-sign on (SSO) make her the perfect candidate for New Hampshire's desired Technology Consultant role. With her experience with data exchanges, SSO, and TIDE, she will be assisting NH DOE and the districts with the transition to the assessment system; working with NH DOE information staff to ensure the accurate and efficient transfer of data to and from NH DOE; creating, defining and reviewing file layouts; and providing assistance in the verification of demographic data. Her experience in media studies will help her communicate with NH DOE and district assessment coordinators about the use of AIR's websites and functionality. Before AIR, Ms. Hubbard had a decade of experience delivering web solutions for the U.S. government; non-profit, international development, education, and news media organizations; and other private sector organizations. She brings experience in technical project management, requirements analysis, and usability and user experience design. She develops requirements, budgets, and schedules for software development projects. She assumes day-to-day responsibility for ensuring that projects are completed on time and within budget and that all deliverables are of the highest quality. Ms. Hubbard facilitates requirements meetings and status meetings with clients and the project team, and she maintains client relations and communications. She received her M.S. in information systems technology from George Washington University and her B.A. in media studies from Pomona College.

Software Project Manager. In addition to the technical lead, each system or family of systems has a software project manager (SPM). The SPM is responsible for ensuring that appropriate protocols are followed, schedules are met, and the software developers have the resources they need to meet their schedules. The SPM serves as the primary liaison to clients and the rest of the project team. The SPM owns the requirements document and the final test plan; and works with the developers, quality assurance team, clients, and senior project director to ensure that the requirements are clear, accurate, and sufficient and that the test plan adequately addresses all requirements. He or she also manages the development schedule, ensuring that resources are allocated appropriately to meet the schedule, notifying the senior project director of variances from the schedule, and working with the senior project director and development team to remediate any deadlines that are likely to be missed. Finally, the SPM documents requested changes to the production systems and the testing of those changes, receives technical lead approval for the requested changes, and submits changes to the Production Control Board.

Technical Lead/Senior Application Developer. This high-level software engineer takes responsibility for the design and implementation of the systems in his or her area. For each new system or modification to a system, the technical lead works with the team to develop the overall architecture; assigns components or modules to senior developers to design, review, and approve detailed requirements and design documents; reviews and approves unit test and integration test plans; and approves all requests for deployment changes going to the Production Control Board. The software project manager and senior application developer for each system are as follows:

Test Delivery (TDS)

- S. Christian Redmond, M.S., PMP, Software Project Manager
- Alan Reeve, M.S., Senior Application Developer

Student Registration (TIDE)

- Sabapathy Karunanithi, M.S., M.B.A., Software Project Manager
- Eyal Moses, Ph.D., Senior Application Developer

Item Bank (ITS)

- Sirisha Nagabhairava, M.S., Software Project Manager
- Teddy Xiong, M.S., Senior Application Developer

Analysis Systems (Quality Monitor and Database of Record)

- Scott Wickett, B.S., PMP, Software Project Manager
- Adam McLaughlin, B.A., Senior Application Developer

Online Reporting

- Kushal Budhwar, M.S., Software Project Manager
- Deepak Sharma, M.S., Senior Application Developer

Under these two main roles, lead developers are responsible for creating design documents and software and for addressing any issues that arise within the production system. The lead developer in the team distributes work assignments, oversees the work of other developers, and implements software. The lead developer is responsible for technical direction, the production of design documents from requirements documents, distribution and oversight of work assignments to the team, review of software modules developed by the team, direct development of some modules, and contribution to and technical review of test plans.

Finally, each system has data analysts responsible for preparing data in accordance with analysis specifications. In a primarily online program, much of the data analysis is automated, and the data analyst's two main roles involve preparation of data for the field-test analyses, special studies, and quality assurance investigations. For this program, the data analyst's main responsibilities include extracting item tryout data, ensuring their quality, and preparing them for analysis by the Psychometrics and Statistics team; providing periodic quality assurance checks to validate automated quality assurance systems; and preparing analysis files for special analyses, validity studies, and other data sets that may be required by the client or the Psychometrics and Statistics team.

Software Quality Assurance for Systems

Software quality assurance (QA) is a critical function. AIR maintains a substantial QA team, and our QA process involves everyone from the senior project director to the developers, QA engineers, and leads. The QA process begins with requirements definition. As requirements for a system or modification are developed, the software project manager (SPM) works with the QA engineer and lead to begin development of the QA design. Once requirements are complete, the SPM, QA lead, QA engineer, and lead developers work together to produce a multifaceted QA plan, which is then implemented during development (unit testing) and after development (integration testing).

Dorian Sofiaj, M.S., Lead, Software Quality Assurance, is responsible for ensuring that every AIR system deployment meets specifications, functions as intended, provides a satisfying user experience, and accurately captures or generates data. He has test responsibility for every system, including AIR's system for item banking, test delivery, electronic reporting, and websites. He develops plans for load testing, interface/functional testing, and data and calculation model testing. He ensures that every system in development has a thorough testing plan that includes automated and thorough regression testing at scale (rather than on small data sets) and that every system documents procedures to ensure security and confidentiality. Mr. Sofiaj joined AIR after more than eight years in software QA at Fannie Mae, where he progressed to senior technical risk specialist. At Fannie Mae, he had managerial and hands-on experience with all aspects of testing, including front and back end, white box and black box, system, regression, failover, shakeout, and performance testing. He has an M.S. in computer science, an M.S. in physics, and a B.S. in mechanical engineering. He has extensive additional training in Oracle/PLSQL, Java, UNIX, security, test methodologies, software development lifecycle (SDLC), management, culture, diversity, and other subjects. Mr. Sofiaj leads AIR's QA team and is responsible for the technical quality of QA plans and their successful implementation, as well as the management and scheduling of the QA testers. He designs the overall architecture of the QA plan, which is embodied in the QA design document, reviews and approves detailed QA plans, and allocates human and hardware resources to implement the QA plan. Under his direction, individual QA leads work with the SPM to draft the QA plan and oversee its implementation. The QA lead participates in the development of the QA design, works on the team led by the SPM to develop the QA plan that implements the QA design, and implements the QA plan, writing automation scripts and overseeing the work of a team of QA testers. Our QA leads on this team have bachelor's degrees (e.g., B.S., commerce with statistics and computers; B.E., electronics and communications), have more than eight years of experience managing software defect-tracking processes, and often hold relevant Microsoft and Cisco certifications.

Network Engineering

As described in our technical proposal, AIR supports application hosting with both external and internal resources. AIR hosts our Item Tracking System (ITS), test delivery system (TDS), reporting, and related critical assessment systems and websites in a managed hosting facility because the scale and investment available from a dedicated hosting company far exceed what any one company can provide. We use several of Rackspace's managed hosting facilities to host the online testing for Delaware, Hawaii, Ohio, Oregon, Utah, and other smaller efforts, as well as other core AIR assessment systems. Rackspace is rated one of the "Magic Quadrant" leaders by Gartner Research. Internally, AIR has our own staff of network engineers dedicated to our assessment infrastructure. This team is led by **Fara Tapscott, B.S., Network Engineering Lead,** who has 16+ years of experience in system engineering, network engineering, DevOps, and full life cycle IT management. She has served as the lead system engineer for AIR's online testing system for the past nine years, helping to design and implement the system and network platforms for the online testing system. Additionally, Ms. Tapscott has served as the primary architect for the

DevOps platform used to automate provisioning of new systems and deploying system configuration changes to all of AIR's online systems. Prior to joining AIR, Ms. Tapscott served as an enterprise systems engineer for Legg Mason Wood Walker, where she was responsible for engineering and support of critical high-volume transactional financial systems. Ms. Tapscott is a part of a system engineering team that includes nine additional systems engineers, with a combined 75+ years of experience in systems and network engineering, administration, and management of 24/7 mission-critical operational environments, as well as multiple full-time and part-time Tier 2 network help desk support staff.

Score Reporting

Margaret Won, M.A., Lead, Score Reporting. Ms. Won will serve as the lead for score reporting. With AIR, Ms. Won's work with Smarter Balanced clients has included creating specifications documents, testing integration between systems, and performing quality control procedures. She manages a team of employees to ensure quality control throughout the development process for the online score reporting systems. Ms. Won received her M.A. from the Brandeis International Business School and her B.A. in economics from Wellesley College.

Validation Handscoring

Damon Hartzler, B.A., Lead, Performance Scoring. Mr. Hartzler will lead the performance scoring for the social studies assessments. Mr. Hartzler has 18+ years of experience in the scoring field. He has several years of experience supporting the Maryland State Department of Education on their statewide assessments involving ELA, government, and mathematics. Currently, he leads the scoring of statewide assessments at AIR's Ohio Processing Center in Columbus. Mr. Hartzler collaborates with project managers and state department officials to ensure state standards are adhered to when scoring statewide assessments. He holds B.A. degrees in both English and history from The Ohio State University.

Customer Service

Cynthia Benis, B.S., Customer Service Manager, manages AIR's Help Desk across all assessment programs, with a focus on support for online testing. Ms. Benis has 25 years of customer service experience, including the past six years at AIR, leading our Help Desk across multiple programs. She manages AIR's Help Desk knowledge bases, inquiry tracking systems, and telephone and electronic communications systems. She oversees full-day support across five time zones for 10,000 monthly inquiries. She manages a team that expands to approximately 50 staff members during peak periods, providing services from 5:00 a.m. to 10:00 p.m. EST, Mondays through Fridays, with occasional Saturday support periods. As the Help Desk manager, she establishes and enforces processes and best practices, adheres to proper training methodologies, provides staff training, produces management reports, ensures that the call center adheres to standards, appropriately communicates program statuses and issues with program management, oversees our clients' service level agreements, and utilizes standard help desk metrics to track and manage workflow. Ms. Benis graduated from The Ohio State University with a B.S. in marketing.

Other Section 29 Requirements

- AIR affirms all key personnel proposed are available to assume the work as proposed.
- AIR has assigned one person (Tom Glorfield) to function as the Program Manager.
- AIR has identified a highly qualified Technology Consultant.
- AIR agrees NH DOE reserves the right to interview and approve assignment of, and changes to, the assigned program manager, program coordinator, lead psychometrician, content development lead, content specific area lead, special populations consultant, and technology consultant, except for those resulting from separation of services. In the event that NH DOE requests removal of specific vendor personnel, AIR agrees to provide acceptable

replacement(s) with no impact to the project. Replacement(s) shall have qualifications which meet or exceed the original staff member proposed or the staff member holding the position previously and shall be approved by NH DOE.

- AIR agrees all personnel who will work on-site at NH DOE or school sites may be required to be pre-approved for site access via a criminal background check paid for by the vendor.

Topic 30 Staff Qualifications and Experiences

Qualifications of all named personnel are presented in Topic 29 Management Team. Resumes are included in Appendix A.

Other Section 30 Requirements

- We believe New Hampshire will find Dr. Gary Phillips, Dr. Ahmet Turhan, and their team to be able to execute routine functions and provide a sophisticated level of expertise to guide the psychometric decisions to be made and re-evaluated as the program evolves.
- Our named content leads will facilitate the blueprint and item review meetings with educators. Our content leads and psychometric staff will facilitate standard setting, if selected. These staff have performed these roles successfully for other clients, and they bring the expertise in assessment and instruction required to ensure success.
- The project organization chart in Topic 28 identifies the Program Manager and named personnel, along with lines of authority, for both AIR Assessment and for the NH DOE project team.
- As described and as New Hampshire has experienced directly, AIR's Senior Management Team, including Dr. Jon Cohen, President of AIR Assessment, work full-time on projects and are readily available to clients to resolve any vendor/client disagreements.
- The Program Manager (Tom Glorfield) and all of AIR's Senior Management Team are available during and outside of normal business hours to assist with any urgent situations.

D2.3 Project Plan

In order to ensure both AIR and the Department stay informed about project activities and timelines, including events that may cause schedules to shift, the AIR project manager will utilize the following management documents to orchestrate the work of the project:

1. Schedules
2. Planning documents
3. Specification documents
4. Tracking documents

These project documents facilitate communication among the Department, AIR project management, and the functional teams doing the work. The owner-to-stakeholder relationships ensure that the information stays up-to-date and that everyone who needs to know about a change is promptly informed.

Planning

Schedules

We produce schedules at different levels, based on the amount of detail required. At the project level, we produce the overall project calendar, which documents work start dates and completion or delivery dates for key activities, milestones, and deliverables. AIR and the Department will discuss and set these key dates at appropriate points throughout the project. After approval of the milestones, we will produce team schedules. Schedules that include tasks requiring interaction with the management partner will be developed in collaboration with appropriate staff. We typically maintain the project schedule in Microsoft Project.

AIR has drafted a preliminary project schedule for this program that includes all deliverables, milestones, review dates, and responsible parties. This project schedule is included in Appendix D. We will work with the Department to update the schedule as needed to accommodate revisions and modifications during the life of the contract. AIR understands that the Department has a very small number of staff available to work on this project and that the Department requires schedules that allow sufficient time to thoroughly review all products and deliverables, as well as reminders in advance of deadlines. Exhibit D2.3-1 summarizes our standard schedules.

Exhibit D2.3-1: Schedules

Team	Schedule	Description	Owner
Project Management	Project Schedule	Defines project milestones and start dates for activities	Project Manager
Psychometrics and Statistics	Analysis Schedule	Lists schedule of data flows and analysis activities	Lead Psychometrician
Operations	Delivery Schedule	Details milestones for packaging and shipping materials	Operational Lead
Operations	Receipt and Scoring Schedule	Defines milestones for document receipt	Operational Lead
Test Development	Batch Delivery Schedule	Lists dates for delivery of item batches and item reviews	Item Development Manager
Test Development	Committee Meeting Schedule	Lists dates and milestones for the preparation and covering of annual committee meetings	Item Development Manager
Computer and Statistical Sciences	Development and Deployment	Outlines schedules for development and deployment	Software Project Manager

Center (CSSC)	Schedules		

Planning Documents

Planning documents kick off complex activities. They outline the general approach to completing the activity, assigning responsibilities, and establishing milestones. Some of these milestones may be additional specifications documents or elements needed to complete a part of the overall plan. Exhibit D2.3-2 summarizes our standard planning documents.

Exhibit D2.3-2: Planning Documents

Team	Document	Description	Owner
Project Management	Project implementation Plan	Summarizes the overall project, including all deliverables, goals, milestones, grades, subjects tested, and modes of testing	Project Manager
Project Management	Administration Summary	Lists characteristics of the administration, such as grades and subjects available for testing, the number of opportunities offered for each, the number of field-test items to be included, and the start and end of the administration period	Project Manager
Psychometrics and Statistics	Field-Test Plan	Describes the sampling methods used to field test items, the statistics used to analyze them, the ways in which risks of item exposure will be mitigated, and other aspects of the field-test plan	Lead Psychometrician
Psychometrics and Statistics	Analysis Plan	Lists objectives of the analysis and the specific analysis methods for calibration and linking for operational psychometrics; special analyses such as generalizability analyses, comparability studies, etc., are described in each study's individual analysis plan	Lead Psychometrician or Study Lead
Psychometrics and Statistics	Standard-Setting Plan	Details the standard-setting workshops, including text of instructions to panelists, statistical analyses, and construction of item booklets	Standard-Setting Lead
Operations	Special Version Summary	Summarizes the special versions to be offered, including translations, alternate language audio, Braille; details the requirements for each test	Project Manager
Operations	Ancillary Abstract	Summarizes ancillary documents to be developed, including administration guides, user manuals, and	Project Manager

		other materials; includes special instructions or requirements	
Test Development	Annual Pool Analysis and Development Plan	Describes annual charts summarizing the state of the item pool relative to the requirements of the item pool or forms to be deployed the subsequent year	Item Development Manager
Test Development	Committee Meeting Summary Sheet	Summarizes committee meetings, including the number of items to be reviewed, number of committee members, and dates	Item Development Manager
CSSC	Project Charter	Describes the goals, approach, and nature of new systems or system modifications	Software Project Manager

Exhibit D2.3-2: Planning Documents (continued)

Team	Document	Description	Owner
CSSC	Testing Design	Details the design document describing the testing strategies to be employed for a new system or the modifications to an existing system; lays the groundwork for the test plan	Software Project Manager
CSSC	Test Plans	Details plans for load testing, individual test cases, and regression tests to be implemented	Software Project Manager
CSSC	Software Release Plans	Lists step-by-step activities for the software development and network engineering teams to deploy software systems for scheduled releases	Software Project Manager
CSSC	Production Control Forms	Lists steps required by AIR's Production Control Board for evaluating and approving the test plans for a configuration change in production systems in the middle of the year	Software Project Manager
Reporting	Reporting Goals	Lists objectives of the elements to be included in the report; describes the goals and the types of actions that the report might motivate, and suggests data elements that might support those actions	Reporting Lead

Implementation

Specifications Documents

Our team leaders work with our clients to develop detailed specifications documents. These function as the detailed requirements documents for each program phase. Each document has an owner and a list of stakeholders. The owner is responsible for keeping the document up-to-date and for ensuring that all stakeholders sign off on any changes to it. These stakeholders include project team members whose work depends directly or indirectly on the work described in the document. In almost every case, client staff members are stakeholders in these documents as well. Exhibit D2.3-3 summarizes our standard specifications documents.

Exhibit D2.3-3: Specifications Documents

Team	Document	Description	Owner
Psychometrics and Statistics	Analysis Specifications	Describes the details of analyses, including software settings, locations of data files	Lead Psychometrician
Psychometrics and Statistics	QC Specifications	Identifies quality control-specific analyses and comparisons to be undertaken	Lead Psychometrician
Psychometrics and Statistics	Scoring-Engine Specifications	Summarizes scores to be reported at various levels, attemptedness rules, and	Lead Psychometrician

		scoring computations that are performed by the test-scoring engine for all tests	
Psychometrics and Statistics and Test Development	Test Specifications	Includes test blueprints, item specifications, and style guides	Lead Psychometrician, Item Development Manager
Test Development	Field-Test Summary	Shows numbers and types of items to be field tested each year and form construction details	Item Development Manager

Exhibit D2.3-3: Specifications Documents (continued)

Team	Document	Description	Owner
Test Development	Item Tracking System (ITS) Specifications	Defines the attributes to be collected for each item, the content review sequences, and the item layout templates	Item Development Manager
Operations	Packaging Specifications	Describes materials to be shipped, shipping carrier, expected packing list format, materials to be packed, and details of packaging	Packaging and Shipping Manager
Operations	Receipt Specifications	Details secure and non-secure documents to be received, expected quantity, receipt start and end dates, and special instructions	Packaging and Shipping Manager
CSSC	Requirement Documents	Defines the detailed requirements for the software or software modification to be implemented	Software Project Manager
CSSC	Test Information Distribution Engine (TIDE) Specifications	Defines detailed requirements for materials ordering, student management, and user management, including user roles within the suite of online testing applications	Software Project Manager
CSSC	Data File Generator Specifications Specifications	Defines detailed requirements for student and item-score files delivered after testing	Software Project Manager
CSSC	Online Administrative Portal	Defines the layout, skin/theme, user cards, and content sections	Software Project Manager
CSSC	Test Delivery System Specifications	Defines the test settings for the Proctor and Student Interfaces and the business rules that govern test opportunities	Software Project Manager
CSSC	Online Reporting System Specifications	Lists the reporting settings for student and aggregate reports	Software Project Manager
Reporting	Reporting Specifications	Details the reports, data elements to be included, calculation rules for each data element, special rules for merging and cleaning data, and other similar items	Reporting Lead
Reporting	Report Mockups	Presents an annotated graphic mockup of reports, providing the	Reporting Lead

		look and feel and identifying the critical elements of their appearance	

Operations

Tracking Documents

Our Program Management team relies on several tracking documents to monitor project status, coordinate work among internal teams, and communicate progress toward milestones to our clients. The priority list identifies the document name, the responsible party, the date the document is due to be returned, and the date the document is due to be finalized. This document also assigns a priority level to help clients manage and prioritize work during busier times in the schedule.

Each document has an owner and a list of stakeholders. The owner is the person responsible for keeping the document up-to-date and ensuring that all stakeholders sign off on any changes to the document. Stakeholders include project team members whose work depends directly or indirectly on the work described in the document. In almost every case, client staff members are stakeholders in these documents as well. Exhibit D2.3-4 summarizes our standard tracking documents.

Exhibit D2.3-4: Tracking Documents

Team	Document	Description	Owner
Project Management	Issues Log	Updated weekly; tracks issues and near- term deliverables and notes progress, problems, risks, assignment, and agreed-upon changes to scope or progress	Project Manager
Project Management	Priority List	Lists the testing and administration documents currently being exchanged between AIR and the Department; assigns a priority value to each document; identifies the document's review level due dates	Project Manager
Project Management	Risk Register	Identifies and classifies project risks organized by function and area; summarizes proposed mitigation strategies	Project Manager
Test Development	Item Release List	Lists specific items to be released each year for each subject	Lead Psychometrician, Item Development Manager
Test Development	Batch Delivery Tracking Sheet	Tracks the item batches by grade and subject, including the number of items in each batch, number of items accepted, and number of items accepted with revisions; tracks due dates	Item Development Manager

Repository for Management Documents

Given the complex nature of assessment projects, the sheer number of documents alone, and the complexity of the review and approval process to track iterations of these documents, maintaining a repository of these documents can be a challenge. Therefore, AIR proposes that all major project documents be stored in KnowledgeTree, including those specified in Exhibits D2.3-1 through D2.3-4.

KnowledgeTree is an online document management system on a secure cloud computing platform. KnowledgeTree manages document workflow and file sharing, alerts document followers to changes, and maintains version control, including authors and document history. Documents are checked in and checked out for editing. Access to documents can be controlled through a user roles and permissions system.

This project will start with a clean slate of project documents. Managing documents through KnowledgeTree from the project kickoff will increase productivity, increase efficiency, reduce error and miscommunication, provide a single reference point for all documents that house key project decisions, and increase collaboration across internal teams and across our organizations. KnowledgeTree has three distinguishing features:

1. **Ease of use:** A document management system can be effective only if it can be sustainably used by users of varying technical ability. KnowledgeTree is user friendly and requires little to no training, even for nontechnical users. For example, creating files by authorized users is as intuitive as doing these same tasks in Windows Explorer.
2. **Workflow management:** The process of fully vetting specification and schedule changes is both complicated and critical. Our process allows participants to identify a stake in a document, thereby becoming stakeholders. Proposed changes are automatically circulated among stakeholders for review and approval. This provides an opportunity for those who must accomplish the work to identify dependencies or implications that may have been initially overlooked.
3. **Built-in version control system:** Documents can be checked in and checked out for editing. Review workflows can also be built into each document type. AIR proposes to establish a secure file transfer protocol (SFTP) site for the transfer of large, secure databases. SFTP offers an efficient mechanism for transferring large-scale data. Such transfers do not require the workflow or overhead associated with the KnowledgeTree site. AIR will maintain the same users across systems.

AIR will maintain the KnowledgeTree and SFTP sites. Before deploying documents to the site, AIR will propose a list of users and their access rights. The Department will have an opportunity to modify this list before providing final approval. AIR will work with the Department to configure the system with user roles and respective file permissions. KnowledgeTree can be programmed to have a hierarchy that is arbitrarily deep.

Change Management

The purpose of change management is to ensure that changes are communicated, managed, and controlled. This is accomplished by working with the Department to clearly define the change control processes during the planning stage so that when the inevitable change requests surface, there will be no confusion about roles, responsibilities, and process requirements. Otherwise, uncontrolled or mismanaged change can negatively impact budget, scope, schedules, and possibly quality.

Weekly conference call meetings between AIR and the Department cover near-term deliverables and address issues as they arise. Weekly meetings are discussed in Topic 32 Management Meetings. We note that though in Topic 32 the Department states a preference for detailed monthly reports to be produced, the AIR project manager will produce additional weekly reports for the Department that will include status of items such as deliverables, key accomplishments, upcoming activities, any changes to the schedule, or any issues. Decisions are documented in the meeting minutes and in the issues log, which is updated weekly and tracks issues and near-term deliverables. The issues log notes progress, problems, risks, and assignments. A sample issues log is included in Appendix H and a sample status report is included in Appendix M.

A potential change will be identified and documented on a change request form and sent to the project manager. The project manager will log all change requests into the change control log and then assign a subject matter expert (SME) to analyze the impact of the change request on the scope/deliverables, schedule, budget, and overall project quality. The SME analyzes the request and documents the impact, including the feasibility of making the change and any impacts and risks associated with the change. In some instances, it may be necessary to define the impact if the change is not made. The project manager will review the analysis and provide it to the key stakeholders, including the Department, to either authorize or reject the change. Whether the change is approved or rejected, the project manager will update the change control log to reflect the updated status. A sample change control log can be found in Appendix N.

Changes are reflected in the appropriate communication documents through a carefully coordinated process. The project manager is responsible for communicating the changes to the document owner, who reviews the changes and amends the document. The document is then routed to each stakeholder who has subscribed to the document. Changes to the document are not final until they receive sign-off from each stakeholder. The document owner is responsible for ensuring that the changes are accurately drafted, reviewed, and accepted by each stakeholder, including the Department.

Issue Management

AIR understands that these are high-stakes administrations and any issue that arises must be dealt with quickly and effectively. The project manager will maintain an issues log that will be updated regularly and reviewed at each weekly meeting.

Once an issue is identified, it will be reported to the project manager, who will evaluate the issue with the project team. If the issue is considered valid, it will be added to the issues log with all pertinent information, and a priority will be assigned. This log will be reviewed during the weekly conference calls. However, if the issue is assigned a priority of medium or high, the issue will be immediately shared with AIR management and Department staff as needed, according to proper escalation protocol. Program staff will be assigned to determine the resolution. This information will be shared with Department staff, but no action will be taken until approval is received from the Department to proceed.

The issues log will be managed by the project manager, and it will be available to all stakeholders through the document management system. Each issue will receive a unique number when entered into the log. The entry in the log will include a description of the issue, the date identified, the person who identified it, and the method used to report it. Each issue will be evaluated to measure the affected program team members, define the impacted areas of the project, and be assigned a priority. Each issue entry will indicate the owner of the issue, the action steps to be taken, and the expected resolution date. Each issue must be approved by the Department before any action can be undertaken. Finally, the actual resolution date and status will be updated. Like most other project documentation, the issues log will be reviewed by the project team regularly to ensure that issues are being resolved. The document will be updated and communicated to all project stakeholders as updates are made.

A sample issue log can be found in Appendix H.

Risk Management

AIR understands the need to analyze and monitor risks in all our programs. It is critical that the risk management plan be vetted and completed as early in the project as possible. AIR looks forward to working with Department staff to complete and approve this process. In conjunction with the Department, an initial risk management plan defining all processes used in risk management will be developed and delivered to the Department within a mutually agreeable timeframe. Because of the high stakes associated with this program, it is vital that schedules be monitored closely and that sufficient time be provided for reviews and production of materials. The risk log will be reviewed at each cycle kickoff meeting and during weekly progress meetings throughout the life of the project.

We look forward to working with the Department to adapt and merge our program management documentation tools and processes to fit the needs of the Department going forward.

Topic 31 Work Plan

The RFP calls for a thorough project Work Plan to address all work offered in the proposal. In addition to our online testing platform (described in D1.2 Solution Technology) that is already in place to deliver and report New Hampshire's Smarter Balanced Assessments during the 2016–2017 administration, we propose to offer AIRCore test items in order to develop test items and forms for the 2017–2018 operational assessments.

AIR would be pleased to offer our full suite of services described throughout the proposal in the development, administration, and reporting activities associated with the New Hampshire Statewide Assessments in ELA, mathematics, and science to the Department upon award of the contract and would be pleased to tailor our Work Plan to the Department's needs throughout the term of the project.

Our standard approach to managing the planning, implementation, operational phases of statewide assessment projects is described in D2.3 Project Plan.

Topic 31.1 Preliminary Work Plan

We refer readers to D2.3 for a description of our comprehensive approach to the planning and implementation phases of statewide assessment projects.

We note that major milestones are included in the project schedule, Appendix D, and the payment schedule and staffing information is included in our separate cost proposal that addresses all of the requirements of Section VII Pricing Model.

Critical success factors of the project include the development of clear modes of communication between AIR and Department project staff and the successful management of documents in the planning, implementation, and operational phases of the project.

Topic 31.2 Project Plan and Schedule

AIR has included a detailed project schedule that is reflective of the work proposed throughout the proposal that includes the following:

- Tasks
- Subtasks
- Beginning date
- End date
- Responsible party/functional group for each deliverable and process specified in the schedule

The schedule clearly identifies and includes both key activities related to the field in addition to key transfer dates between AIR and the Department related to all phases of the project's processes that includes but is not limited to the following:

- Production
- Shipping and receipt
- Administration
- Scoring
- Data processing
- Reporting and psychometric activities

We note that, as we are proposing the entry of student responses on paper-based assessments into the data entry interface (DEI) for immediate processing, machine/AI scoring, and reporting, we have not included any scanning processes in our proposed Work Plan and project schedule (Appendix D).

AIR understands the requirement to develop a detailed project plan and schedule in preparation for each fiscal year that distinguishes the development for the next year's assessment and reporting for the prior year's assessment from activities related to the current year's assessment.

AIR understands and will comply with the requirement to conduct a review of the project schedule included in Appendix D, followed by the Department's approval of the schedule. We understand that this review for the first contract period should occur within two weeks of the contract award and the necessity to mutually agree upon final dates.

AIR understands the importance of continual joint monitoring of the project schedule and ensures that all schedule adjustments agreed upon with the Department will allow for final deliverable dates to be met. AIR understands that, if it becomes necessary, timelines and schedules may be revised with prior approval of the Department and an executed contract amendment for all deliverables subject to liquidated damages.

AIR understands and will comply with all terms set forth in Topic 31.2 Project Plan and Schedule.

Topic 32 Management Meetings

The Project Management team is responsible for orchestrating delivery and reporting for New Hampshire Statewide Assessments. The project manager will lead all internal management meetings and all communications with the Department. In addition, the team will support and engage with the Department to effectively analyze and monitor project activities, communicate regularly about progress toward project goals, and document outcomes of project activities. These strategies improve efficiency in completing tasks. AIR will assume the responsibility for the costs associated with all project management meetings.

We note that in all meetings, AIR will designate an AIR New Hampshire project team staff member to prepare and circulate meeting agendas, background (including any updates to the Work Plan discussed in Topic 31 Work Plan), and minutes for each meeting specified below. Additionally, AIR staff assume the responsibility to draft formal presentations as necessary to include the kickoff meeting.

Kickoff Meeting

AIR will work with the Department to schedule an initial two-day, in-person management meeting to be held within two weeks of the contract award. This meeting will include key vendor staff and the NH DOE Project Management team to enable project leaders from AIR and the Department to become acquainted and establish any preliminary project procedures. AIR anticipates that the kickoff meeting will focus on finalizing the project work plan and schedule as indicated in Topic 31 Work Plan.

Having a comprehensive agenda that keeps the meeting focused on important issues is one of the keys to a successful meeting. Senior AIR project and management staff will meet internally immediately upon award of the contract to discuss what topics need to be covered at the kickoff meeting and what decisions need to be made during that meeting for the project to have a smooth and successful start-up. Simultaneously, the AIR project manager will reach out to the NH DOE project officer to get his or her input on topics that need to be on the agenda for the kickoff meeting. Based on the input from both AIR staff and the NH DOE project officer, the AIR project manager will draft an agenda for the kickoff meeting and share it with the NH DOE project officer. We anticipate that the two project leads—one from AIR and one from the Department—will circulate the draft agenda to their project staff for input as necessary and work together to finalize the agenda. The final agenda should include a time estimate for each item on the agenda and designate the person responsible for coordinating and leading the discussion on each agenda topic.

AIR also anticipates that we will work with the Department at the kickoff meeting to develop a Communications Plan that specifies the following:

- When the information needs to be communicated (on a specific date or at regular intervals)
- The person responsible for communicating the information
- The method used to communicate the information
- The documentation processes for information that has been communicated and signoffs that have been received (if applicable)

The Communications Plan is outlined in greater detail in Topic 33 Project Communication.

We have included a sample agenda for the kickoff meeting included in Appendix J.

Annual Planning Meeting

During the annual planning meeting, the Department and AIR will review schedules for the upcoming year, establish the baseline plan, and work through scope of work modifications or enhancements. Attendees will include the project manager and leads in each program area. AIR will draft an agenda for review, revision, and approval by the Department. We will discuss any challenges encountered while performing the work under this contract and changes to processes and procedures as needed to allow the project to run more smoothly.

Weekly Status Meeting

Although the RFP requires weekly phone calls in only year one of the project, AIR considers weekly conference calls/web-based meetings to be an important and routine element of effective communication throughout the entirety of the project. The assembly of key stakeholders provides a forum for solving complicated issues and ensuring that AIR’s approach coincides with the Department’s vision for its Statewide Assessment program. AIR will host the weekly program update meetings, which will be led by the project manager, to discuss the week’s progress and to identify upcoming deadlines and possible challenges. The weekly project status report, distributed in advance of the meeting, will provide the basis for much of the discussion, and it will be updated after the meeting as needed. During peak intervals in the project schedule, AIR proposes to host the meetings more frequently if necessary.

AIR maintains a suite of communication tools and services to support project communication. These tools are listed in Exhibit D2.3-5 in Topic 33.1 Ongoing Communication. AIR will provide a toll-free conference line for all project-related conference calls. AIR can also offer web-conferencing solutions that enhance a virtual meeting. AIR has made significant investments in the hardware, software, and connectivity required to support virtual meetings. We use well-known industry-standard solutions that are easily accessible.

A sample weekly status meeting agenda is included in Appendix K.

Status Meetings

AIR recognizes and supports the need to conduct a minimum of two status meetings per month to include AIR New Hampshire project team staff and Department staff in order to discuss overall project status, the reports produced, and additional topics as necessary. We propose to host a weekly status meeting to facilitate regular discussion and status communications between the Department and AIR New Hampshire project team staff.

Reports

AIR agrees to submit reports in accordance with the requirements highlighted in the RFP. Reports will delivered in a method, format, and timeline as agreed upon by AIR and the Department during the kickoff meeting. We agree to include the following items, at a minimum, in each report:

- Project status as it relates to the Work Plan
- Deliverables status
- Accomplishments during the weeks being reported
- Planned activities for the upcoming two-week period
- Future activities
- Issues and concerns requiring resolution
- Reports and remedies in case of falling behind schedule

Special Meetings

AIR recognizes and supports the need to include the additional special meetings at the discretion of the Department and AIR New Hampshire project team staff to address specific issues as they arise.

Exit Meeting

AIR recognizes and supports the need to include an exit meeting in order to transition the testing program to the vendor at the conclusion of the contract. Our approach to ensuring the successful transition of materials and knowledge is outlined in Topic 45 Transition.

A sample exit meeting agenda is included in Appendix L.

Topic 33 Project Communication

We are committed to working closely with Department staff to coordinate project activities, and we will be available and on call to ensure that staff are kept informed of project status. At AIR, management is about analysis and communication. Analysis is the process of breaking down tasks, risks, or issues into their component parts. Once the component risk points are isolated, we can address each and identify mitigation strategies. Communication encompasses many stages:

- Gathering information about goals, objectives, and the current context
- Communicating work breakdown structures to those who must complete the tasks
- Tracking progress

Clear, timely information paves the way for success.

AIR has developed a series of tools that enable our project directors to analyze and implement the project requirements more efficiently. These tools form the basis for clear, open, and accessible communication about the project schedule, deliverables, and specific requirements. Our project directors act as coordinators among the technical experts in our other functional areas, much as a general contractor orchestrates the work of specialized subcontractors on a construction job.

As noted in Topic 31 Work Plan, schedules, tracking documents, change management processes, and other tools facilitate communication among Department staff, AIR project management personnel, and the functional teams doing the work.

Communications Plan

AIR understands that for this project to be successful, we will need to collaborate with many stakeholders. AIR will make every effort to be flexible, provide information as required, and keep this important project on track.

Upon award of the contract, AIR will work with the Department to develop a communications plan that specifies

- who needs to receive specific information;
- when that information needs to be communicated (on a specific date or at regular intervals);
- the person responsible for communicating the information;
- the method used to communicate the information; and
- the processes for documentation that the information has been communicated and sign-offs have been received (if applicable).

AIR will create a project communication organizational chart that will include all key team members, their contact information for immediate access, and chain of command contact information. AIR will assist the Department in designing a comparable project-based organizational chart, if desired, so that contact information is at everyone's fingertips.

Moreover, we recognize the importance of an established and predictable escalation protocol. AIR will coordinate with the Department during the resolution of all escalated issues to ensure timely resolution and communication with affected users of the assessment system. We will also work closely with the Department to develop and document an effective strategy for ensuring that the Department is given advance notice of changes that may disrupt service to users of the online system.

Please see Appendix I for a sample communications plan.

Topic 33.1 Ongoing Communication

Real-Time Communication with the Department

In addition to the formal communications plans identified in this section, the project manager will communicate proactively with Department staff, informing them of any project change that might affect the schedule or deliverables. AIR's proposed communications strategy will ensure that the Department is informed in advance of emergent issues.

Weekly Meetings

AIR considers weekly conference calls and web-based meetings to be an important and routine element of effective communication. The assembly of key stakeholders provides a forum for solving complicated issues and ensuring that AIR's approach coincides with the Department's vision for its assessment program.

AIR will host the weekly program update meetings, which will be led by the program manager, to discuss the week's progress and identify upcoming deadlines and possible challenges. The weekly project status report, distributed in advance of the meeting, will provide the basis for much of the discussion at the meeting, and it will be updated after the meeting as needed. During peak intervals in the project schedule, AIR proposes to host the meetings more frequently if necessary.

The project plan and schedule, policy decisions that need to be made, and upcoming tasks and timelines that need to be met will be reviewed. At each meeting, the Department will receive general progress updates from each team leader. The Department will have the opportunity to speak directly with AIR's technical, systems engineering, content development, operations, and project management team leaders.

Decisions made during the meeting will be documented in the appropriate specifications document and in the weekly meeting minutes. Changed specifications will be circulated among internal and Department stakeholders before the decision becomes final. In this way we ensure that (1) the entire team is aware of the change and (2) each functional group has an opportunity to identify risks or issues that may arise as a result of a change in another functional area.

Communication Media

AIR maintains a suite of communications tools and services to support project communication. These tools are listed in Exhibit D2.3-5.

AIR will provide a toll-free conference line for all project-related conference calls. AIR can also offer web-conferencing solutions that enhance a virtual meeting. AIR has made significant investments in the hardware, software, and connectivity required to support virtual meetings. We use well-known industry-standard solutions that are easily accessible.

Exhibit D2.3-5: Communication Tools

Tool	Description
GoToMeeting	GoToMeeting may provide the best option and will facilitate meetings involving attendees in multiple locations. Again, AIR will adapt to existing conference services solutions as needed. GoToMeeting connects us with real-time, online webinars and telephone conferencing. Whether we are meeting with Department staff or assisting in training state personnel or local staff, GoToMeeting will assist us in connecting. In addition to sharing presentation slides, participating in live video, and recording events, attendees can interact regardless of the size of the meeting using a live Q&A.
Global Crossing Conference Phone Line	The Global Crossing Conference Phone Line is a traditional audio conference line and may be suited for smaller Department/AIR meetings.
Skype	Increasing in popularity, Skype offers video and audio conferencing, instant messaging, and desktop sharing. Skype is easily downloaded and requires little technical support. A user guide and a tutorial are available to assist users accessing Skype.
Tandberg Video Conferencing Services	For complete video conferencing across multiple locations, AIR offers Tandberg Videoconferencing Services. Although this solution requires local support (e.g., video cameras, connectivity), Cisco's Tandberg offers screen-based video communication systems that enhance our ability to communicate objectives for a particular meeting.
KnowledgeTree Site	AIR will maintain a KnowledgeTree site for the Department. Before deploying documents to the site, AIR will propose a list of users and their access rights. The Department will have an opportunity to modify this list before providing final approval.
Secure File Transfer Protocol Site	AIR proposes to establish a secure file transfer protocol (SFTP) site for the transfer of large databases and secure materials. The SFTP site offers an efficient mechanism for transferring large-scale data. An SFTP site will also be available for the transfer of documents and data that do not require the version control of KnowledgeTree. This site will be available to both the Department and the field as needed. Such transfers do not require the workflow or overhead associated with the KnowledgeTree site. AIR will maintain the same users across systems.

Topic 33.2 Timeliness of Communication

AIR understands the importance of timely communication with Department staff and commits to response times of 24 hours or less with the Department for routine e-mail and telephone communications. AIR will provide the Department with advance notification of dates on which the program manager will be unavailable and provide alternate contacts and contact information to maintain a direct line between the Department and the AIR New Hampshire project team.

Topic 33.3 Monthly Reports

AIR understands the requirement to produce a monthly report to summarize all project actions, issues tracked in the issues log maintained by the project team and discussed in Topic 35, Risk Management and Quality Assurance, upcoming milestones and important dates, and additional information to support the project needs. AIR understands and will adhere to the requirement that the monthly reports be sent monthly to the Department, producing retrospective reports on the third business day of the following month via e-mail and to the personnel identified during the initial project kickoff meeting.

Topic 34 Program Improvement Plan

AIR understands the need to maintain a program improvement plan for each phase of the assessment program. We believe that program success may be achieved through collecting continuous feedback both from the Department and from the field about potential improvements. We agree that there is a need to document issues and successes as they arise. At the conclusion of each phase listed below, AIR will produce for the Department's review a copy of the Program Improvement Plan we have maintained:

- Development
- Production
- Shipping and receipt
  - We note that we propose the data entry interface (DEI) for the collection and entry of student responses at the local level as opposed to scanning and handscoring responses on paper-pencil materials. However, we commit to the secure distribution and collection of these materials as required by New Hampshire policies and procedures.
- Administration
- Scoring
  - We note that we propose 100% machine and artificial intelligence (AI) scoring.
- Data processing
- Reporting
- Psychometric services

The report will include a detailed explanation of the activities performed for each phase of work and will provide recommendations for improvement in the next administration cycle. Reports will be produced for Department review within one month of completing each phase of work. We understand that the report should detail any errors, problems, or discrepancies by district and school to allow for the analysis of patterns within the data that can be used to clarify instructions provided in Test Administration Manuals and the Test Coordinator Manual.

AIR proposes to review each individual report produced in the Program Improvement Plan during the annual planning meeting, to discuss lessons learned from each aspect of the administration period in order to determine areas for improvement or opportunities for clarifications to the field. The New Hampshire Help Desk also serves as an important indicator for areas of program improvement, based on suggestions from the field, frequently asked questions, and recurring issues. AIR commits to collecting suggestions received from the field through the Help Desk, in addition to tracking patterns in reported issues (e.g., system issues, lack of clarity from user guides or manuals) to maximize the Program Improvement Plan. In addition, we note that during each year of the Smarter Balanced contract, AIR facilitated a meeting with New Hampshire's Technical Advisory Committee (TAC) to discuss the most recent administration and gather feedback and improvement requests for the upcoming year, and we propose to continue facilitating these meetings for New Hampshire's new statewide assessments.

During the course of each administration period, AIR collects feedback from all clients in order to produce a consolidated log of requested system enhancements. AIR conducts a careful review of this list to determine which enhancements can be built and deployed to best serve the full range of states and their assessment programs. AIR commits to presenting the Department with a visual walk-through of each of the system enhancements to be deployed during each upcoming administration period. We note that the Test Administrator Interface (TA Interface) Test Selection Tree was modified based on feedback from New Hampshire and other AIR clients for improved system functionality. These changes are illustrated in Exhibit D2.3-6 and Exhibit D2.3-7. Exhibit D2.3-6 is the TA Interface Test Selection Tree during the New Hampshire Spring 2015 assessment administration and Exhibit D2.3-7 is the TA Interface Test

Selection Tree during the New Hampshire Spring 2016 and 2017 assessment administration. As is evident, in addition to other TA Interface enhancements, we added expandable and collapsible tree branches for each assessment offered and each grade level. Color coding was added to allow TAs to easily differentiate each assessment (i.e., summative vs. interim). We are pleased to implement innovative system enhancements to improve the quality of the assessment programs we support. Currently, we are developing system enhancements to be implemented during the 2017–2018 administration and would embrace the opportunity to present these system enhancements to New Hampshire upon the award of the contract and during the project kickoff meeting.

Exhibit D2.3-6: New Hampshire TA Interface Test Selection Tree Spring 2015

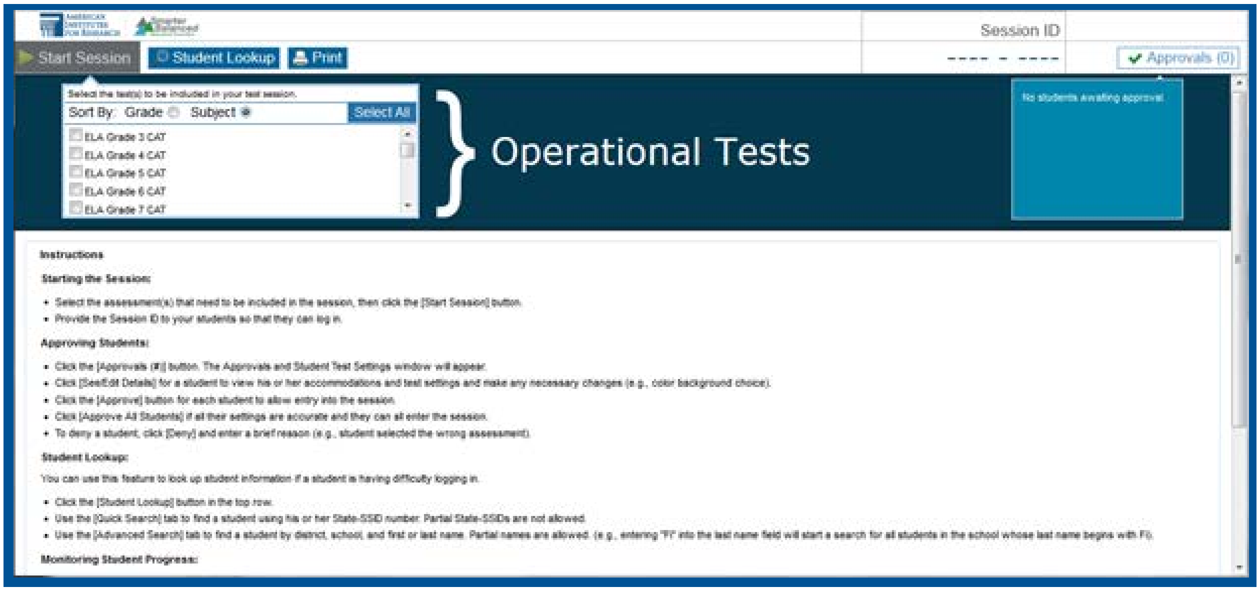
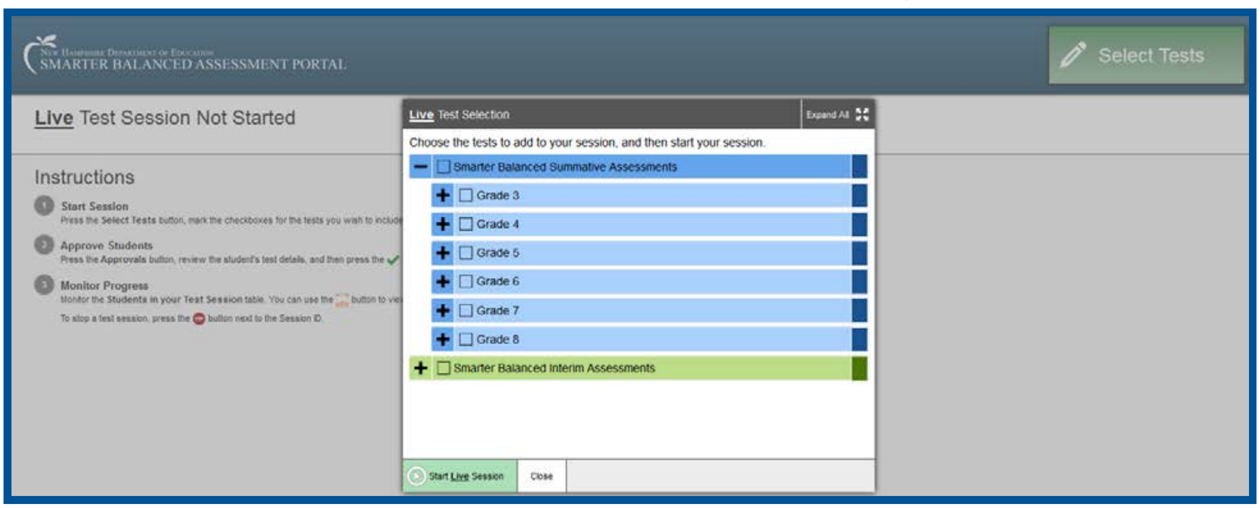


Exhibit D2.3-7: New Hampshire TA Interface Test Selection Tree Spring 2016 and 2017



In addition to feedback from the Department and from the field through the Help Desk, another valuable source of program feedback is the Technical Advisory Committee (TAC). If it is the Department's preference, we will collect systems and overall program feedback from the TAC during seasonal TAC meetings that include AIR New Hampshire project team staff, Department project leads, and TAC members. Discussions and decisions will be documented by a designated minutes-taker and AIR project team staff will discuss the feedback collected with Department staff in order to make decisions about program improvements and items that may be added to the enhancements log.

Topic 35 Risk Management and Quality Assurance

AIR understands the need to analyze and monitor risks in all our programs in order to identify risks and take the appropriate action on issues and risks as they arise throughout the project's lifecycle. It is critical that the risk management plan be vetted and completed as early in the project as possible. AIR looks forward to working with Department staff to complete and approve this process. In conjunction with the Department, an initial risk management plan defining all processes used in risk management will be developed and delivered to the Department within a mutually agreeable timeframe. Because of the high stakes associated with this program, it is vital that schedules be monitored closely and that sufficient time be provided for reviews and production of materials. The risk log will be reviewed at each cycle kickoff meeting and during weekly progress meetings throughout the life of the project. We look forward to working with the Department to adapt and merge our program management documentation tools and processes to fit the needs of the Department going forward. A sample risk register is included in Appendix C.

AIR understands that these are high-stakes test administrations and any issue that arises must be dealt with quickly and effectively. The project manager will maintain an issues log that will be updated regularly and reviewed at each weekly meeting.

Once an issue is identified, it will be reported to the project manager, who will evaluate the issue with the project team. If the issue is considered valid, it will be added to the issues log with all pertinent information, and a priority will be assigned. This log will be reviewed during the weekly conference calls. However, if the issue is assigned a priority of medium or high, the issue will be immediately shared with AIR management and Department staff as needed, according to proper escalation protocol. Program staff will be assigned to determine the resolution. This information will be shared with Department staff, but no action will be taken until approval to proceed is received from the Department.

The issues log will be managed by the project manager, and it will be available to all stakeholders through the document management system. Each issue will receive a unique number when entered into the log. The entry in the log will include a description of the issue, the date identified, the person who identified it, and the method used to report it. Each issue will be evaluated to measure the affected program team members, define the impacted areas of the project, and be assigned a priority. Each issue entry will indicate the owner of the issue, the action steps to be taken, and the expected resolution date. Each issue must be approved by the Department before any action can be undertaken. Finally, the actual resolution date and status will be updated. Like most other project documentation, the issues log will be reviewed by the project team regularly to ensure that issues are being resolved. The document will be updated and communicated to all project stakeholders as updates are made. A sample issues log can be found in Appendix H.

D-3 PROJECT EXECUTION

D3.1 Implementation and Operation

Topic 36 Implementation Approach

AIR is proposing to utilize the same online testing systems used for the current Smarter Balanced assessments for all components of assessment administration and reporting, with the exception of the data entry interface (DEI) to record student responses on paper-pencil assessments. As a result, the transition to the New Hampshire Statewide Assessments in ELA, mathematics, and science will be smooth for Department, district, and school personnel who are already familiar with our systems and processes for student registration, materials ordering, test administration, and reporting.

We are also pleased to say that the College Board has selected AIR to provide the test delivery system for their suite of assessments, including the SAT. Thus, should the Department award AIR the contract for New Hampshire's new statewide assessments, not only will the grades 3 through 8 ELA and mathematics assessments be administered on a platform familiar to New Hampshire students and educators, the grade 11 administrations of the SAT will also be administered by the same familiar and reliable test delivery system.

For paper testers, instead of asking districts to return the answer document for scanning, AIR is proposing that teachers enter the student responses from the paper books into AIR's DEI system for faster turnaround of results. AIR would be delighted to produce additional ancillary materials (e.g., quick guides) to support those who will use the DEI to input student responses. AIR also proposes to include use of the DEI as a topic in the four half-day trainings for the field. Though this system is intuitive and may not warrant an online training (webinar) for system users, we would be happy to support this if the Department deems it is necessary.

Because AIR's online testing systems are already in place in New Hampshire and users are familiar with the system functionalities, we consider it an extremely low risk in transitioning to delivering the Statewide Assessments. In fact, little system maintenance is required to get these systems ready for the 2017–2018 administration, and we would be happy to implement any offered system enhancements for New Hampshire users.

The time frames for major milestones are provided in the project schedule, included in Appendix D.

Topic 37 User Acceptance Testing

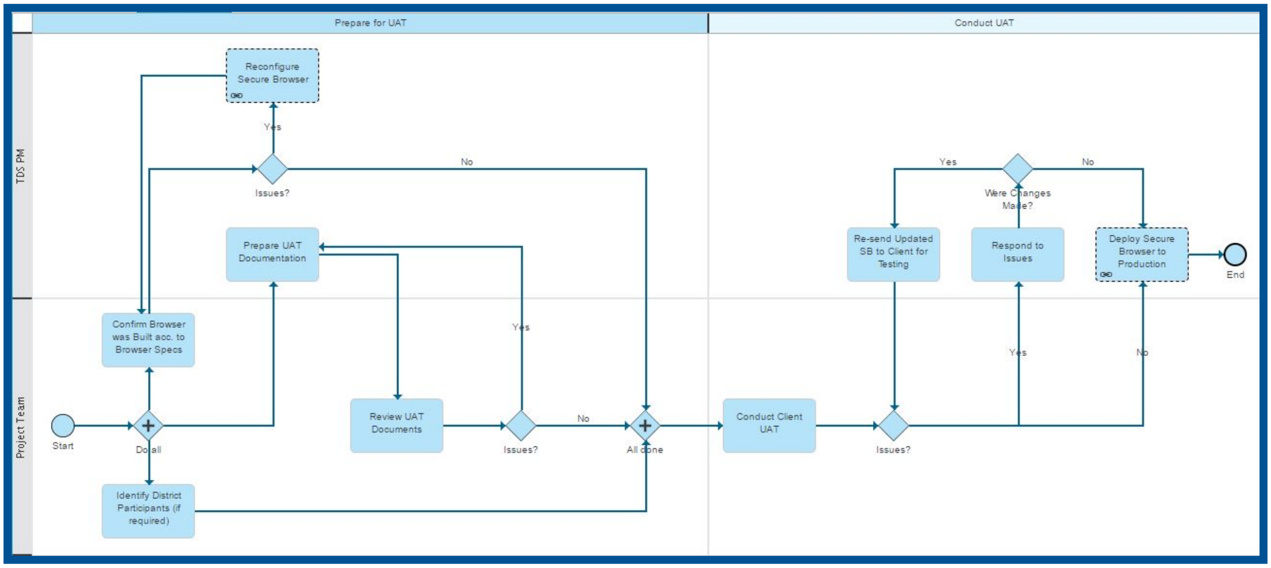
Testing Methodology and Proposed Test Plan

Prior to deployment, the testing system, as well as all supporting systems, and content are deployed to a staging server where they are subject to user acceptance testing (UAT). The UAT environment reflects the production environment and serves as a software and content evaluation period, giving the Department an opportunity to interact with all AIR systems before they are deployed to production and are made available to users.

AIR will first conduct internal UAT to detect any functionality or appearance defects in the system and will initiate the process for correcting system defects before giving the Department access to the UAT site for review. If defects are not corrected before the Department has access to the site, AIR will provide the Department with a list of known issues and a timeframe for their resolution to eliminate redundancy in the review process. AIR will also provide the Department with a system- and New Hampshire-specific UAT document that provides important login information and outlines test cases and expected system functionality. An example of a Test Information Distribution Engine (TIDE) UAT document is included in Appendix E. AIR and the Department will collaborate regularly during the UAT period to ensure that any system defects are identified and communicated. The AIR New Hampshire project team will be responsible for working with other AIR staff and the systems team to resolve issues detected during the UAT period before each system's deployment.

In order to give the Department visibility into AIR's internal and external UAT processes, we illustrate the UAT process for the secure browser in Exhibit D3.1-1.

Exhibit D3.1-1: User Acceptance Testing Internal and External Processes for the Secure Browser



Include the time the State will need to complete User Acceptance Testing of a component.

Please see the list of systems and associated tentative UAT review periods for each system outlined in Exhibit D3.1-2. The dates included in this tentative schedule reflect the dates used for the UAT cycle for the relevant AIR systems used to administer New Hampshire's Smarter Balanced Assessments in the 2016−2017 test administration that are proposed for the 2017−2018 administration period. The exact dates are included purely for explanatory purposes. Each UAT cycle will have its own schedule to ensure the Department is comfortable with each system and that all identified issues have been adequately addressed. Some systems (e.g., the TDS) require two rounds of UAT to ensure that all content and functions perform as expected and some systems require downtime before deployment. AIR will also conduct system integration testing internally (e.g., ensure that a form assignment and student settings indicated in TIDE are communicated to the TDS and scores generated from answers given in the TDS are reflected in the ORS) to ensure that all systems are communicating with each other as expected. We note that TDS UAT includes testing of the secure browser and the TA Interface simultaneously.

Not all systems require downtime. However, where required, downtimes will be scheduled in logical, agreed-on timeframes with the Department to minimize the impact to users. AIR will also draft approved verbiage to communicate scheduled downtime to system users to be posted as an announcement on the online administrative portal well in advance of the downtime start date. AIR will work with the Department to determine the most adequate amount of time for each system's UAT during the initial kickoff meeting. AIR is prepared to provide additional support to the Department in the UAT processes as necessary.

Exhibit D3.1-2: UAT Schedule for 2016–2017 New Hampshire Smarter Balanced Assessments Systems

System	UAT Period Start (AIR)	UAT Period Start (Department)	UAT Period End (Department)	System Downtime Start*	System Downtime End*	System Deployment
Secure Browser (Student Interface)	8/7/16	8/9/16	8/15/16	N/A	N/A	8/16/16
TIDE	8/29/16	9/6/16	9/21/16	9/24/16	10/6/16	10/6/16
TA Certification	9/19/16	9/26/16	9/30/16	N/A	N/A	10/6/16
TDS (Interim Assessments) UAT 1	9/6/16	9/19/16	9/23/16	9/26/16	10/3/16	10/18/16 (TDS Testing Window Opens)
TDS (Interim Assessments) UAT 2	10/3/16	10/3/16	10/17/16	10/10/16	10/17/16	10/18/16 (TDS Testing Window Opens for Interim Assessments)
TDS (Summative Assessments) UAT 1	1/23/17	2/6/17	2/10/17	2/14/17	2/21/17	3/14/17 (TDS Testing Window Opens for Summative Assessments)
TDS (Summative Assessments) UAT 2	2/21/17	2/21/17	2/27/17	3/9/16	3/14/17	3/14/17
ORS (Interim Assessments)	9/6/16	9/19/16	10/7/16	10/6/16	10/18/16	10/18/16
ORS (Summative Assessments)	1/23/17	2/6/17	2/27/17	N/A	3/14/17	3/14/17

*Not all systems require downtime before deployment. All scheduled system downtimes will be scheduled in timeframes agreed on with the Department.

Support Provided to Prepare the State Staff during UAT

Because the Department staff are familiar with AIR's proposed systems, there will be no learning curve to adapt to a new system. If new staff become involved with UAT, staff would like a refresher on UAT processes or system use and functionality, AIR introduces enhancements to the testing systems, or AIR introduces a new system altogether, AIR would be pleased to offer a broad range of support to the Department during the UAT processes. Support may include, but is not limited to

- telephone, e-mail, or web conferencing technical support to walk Department staff through the UAT environment or examine and discuss a system issue in real time;
- drafting and posting approved verbiage to the online administrative portal to communicate systems downtime or go-live (deployment) dates;
- maintaining, with AIR and Department staff, a collaborative, version controlled, up-to-date list of identified issues and issue resolution timeframes to be discussed in agreed-on intervals (at a minimum during weekly status meetings).

AIR would be eager and available to provide additional support to prepare Department staff during UAT as it becomes necessary.

Preparation Required for Testing the Configured Software

Before any AIR system is subject to Department review in UAT, AIR software project managers and the dedicated AIR New Hampshire project team ensure preparation for UAT of each system by

- confirming system alignment with the system specifications and New Hampshire-specific configurations requested by the Department (if applicable);
- preparing and reviewing UAT documentation to include applicable test cases, expected outcomes, and appropriate login information to be distributed to the client; and
- conducting internal UAT and, if any issues arise, documenting and tracking the issues to resolution. (Outstanding issues at the time of Department engagement with UAT will be communicated, and a resolution timeframe will be provided).

AIR would be pleased to offer necessary systems training to Department staff to prepare staff for engagement with UAT.

Please refer to Exhibit D3.1-21 for a visual illustration of AIR's internal UAT preparation processes.

Documentation Available to the Testing Team for Configuration

As previously discussed, AIR will provide UAT documentation to Department staff to test cases and expected outcomes to determine alignment with system configurations requested by the Department. An example of a TIDE UAT document is included in Appendix E. AIR will also provide a template to be shared by the Department and the AIR New Hampshire project team to document and track issues discovered in UAT and their resolution. We propose to house this document on KnowledgeTree, a document management repository, in order to ensure version control.

Defects Likely to be Encountered Based on Previous Experience

At the beginning of the 2016–2017 New Hampshire Smarter Balanced test administration, our TIDE user interface underwent a facelift to improve user functionality. This afforded the Department and the New Hampshire project team the opportunity to familiarize themselves with the system functionality and to discover minor system defects.

One such defect we encountered during this time was the inability for a user to download materials utilized by our Teacher Handscoring System (THSS) during this administration. User acceptance testing (UAT) revealed that the user received an error message when trying to access these materials. Our dedicated New Hampshire project team communicated with our TIDE project team in order to quickly investigate and resolve the issue to return the system to its intended functionality.

Another example of a non-functional defect we encountered during this time was that the TIDE User Guide embedded in the system did not match the TIDE User Guide that had been approved by the Department but which was not yet posted to the online administrative portal. This issue was resolved quickly and simply.

Time Frames for Investigation of Defects

AIR will immediately escalate the identification of any system defects to the appropriate systems project management team for investigation and resolution. AIR will track the issue resolution status internally and ensure that the issue resolution status is promptly communicated to the Department.

Time Frame for Defect Correction

While it is impossible to provide a definitive timeframe for defect correction in the UAT environment, the AIR New Hampshire project team and all systems project management teams will work diligently to immediately investigate the report of a system defect, whether the defect is discovered by AIR or by the Department. To ensure prompt resolution, AIR and the Department will determine appropriate methods to communicate the detection of a systems defect. Issue resolution is contingent on the detection of and immediate reporting of issues to the AIR New Hampshire project team when Department staff discover the defect. AIR will diligently adhere to all defect identification and reporting processes as decided on during the initial kickoff meeting with the Department.

Provide a Sample User Acceptance Test Plan from a Completed Project as an Appendix

An example of a TIDE UAT document is included in Appendix E.

D3.2 Ongoing Operations

Topic 38 Help Desk Support

To ensure that the NH DOE and all New Hampshire district test coordinators, school test coordinators, schools, users, educators, and test administrators have prompt and accurate assistance for all questions and requests, users may contact the AIR Help Desk by calling a dedicated toll-free customer support line, by sending an e-mail, or by chatting real-time with a dedicated agent who is ready to assist. Currently, this comprehensive support is available Monday through Friday from 7:00 a.m. to 4:00 p.m. EST outside of the summative testing window and 7:00 a.m. to 7:00 p.m. EST during the summative testing window. We are prepared to provide comprehensive support to New Hampshire users from the hours of 7:00 a.m. to 5:00 p.m. EST Monday through Friday as stated in Topic 39 Support Center of the RFP if that is the Department's preference. Users may leave a voicemail or send an e-mail outside of these hours and will receive a response within one hour of when the Help Desk opens the next business day.

AIR understands the importance of responding to and resolving user requests as quickly as possible. For this reason, our Help Desk is organized in a tiered manner that provides an efficient support structure to assist with the full range of queries received, from routine queries to more in-depth requests. This tiered structure is comprised of Tier 1, 2, and 3 teams staffed with team members who are well trained and knowledgeable about policies and procedures specific to the New Hampshire Statewide Assessment program. The AIR project team will collaborate with the Department to determine the degree of direct contact with districts and schools and which inquiries will be directed to the NH DOE to address. At a minimum, AIR will address inquiries specific to administration processes as included in the associated manuals and ancillary materials.

Tier 1

Tier 1 agents are dedicated by project and are trained to answer routine questions and to provide explanations and clarifications using scripted and semi-scripted responses. Agents will address inquiries specific to user roles, training, registration, reporting, and test administration processes as included in the associated manuals and ancillary materials. Examples of types of support include user role issues, application login, location of user guides and manuals, access to score reports, Secure Browser installation, test proctor and student system interfaces, resolution of system access issues that might occur in a school, and special case processing such as test invalidations or test resets. Help Desk staff will be able to identify questions that are matters of policy and belong to the NH DOE and those that are to be answered directly by AIR staff. AIR Help Desk agents will make initial contact regarding any inquiries within 24 hours of receipt; during testing windows, the response time will be within two hours.

The majority of user contacts are resolved at the Tier 1 level with a first call resolution rate of approximately 80%, and an average call duration of under eight minutes. Ninety percent of all cases are resolved in less than 24 hours.

Tier 1 team activities are supervised by a dedicated Help Desk project coordinator who will be a subject matter expert on all New Hampshire assessment policies and procedures. Project coordinators report directly to the assistant site manager who oversees the day-to-day activities and performance of the overall Tier 1 Help Desk.

AIR will establish a Tier 1 New Hampshire team with a dedicated Help Desk Project Coordinator who will be a subject matter expert on all assessment program policies and procedures. The Project Coordinator reports directly to the Assistant Site Manager, who oversees the day-to-day activities and performance of the overall Tier 1 Help Desk. The New Hampshire Help Desk agents will be trained and knowledgeable about each assessment program, including materials, processes, and use of the secure site for functions related to administration, scoring, and reporting. Agents will be knowledgeable of state-specific policies and concerns and be able to determine which inquiries to answer and which to escalate to the AIR New Hampshire project team.

All telephone calls, e-mails, and chat interactions between Help Desk agents and users are recorded within the 8x8 Virtual Call Center system. The 8x8 system provides supervisors the ability to listen to recorded calls, monitor real-time call queue statistics, and retrieve case histories sorted by caller.

When the Help Desk is contacted, a case is automatically generated with a unique case number, and the following information is captured:

- Caller's contact information
- Help Desk agent name
- Summary of the request
- Status of the case
- AIR system (e.g., TDS, TIDE)

AIR personnel who are not directly associated with the Help Desk, such as project team members and department heads, are made aware of new cases and case updates via e-mail correspondence. When replies are made within e-mail, these replies are automatically updated back into the 8x8 case details.

All Help Desk personnel have access to Superhelp, an online knowledge management tool designed to provide consistent responses to calls about the same issue by providing an electronic repository of FAQ documents, guides, manuals, and other resources. The use of this tool is a cornerstone in ensuring that agents provide up-to-date, accurate information to every user during every contact. The content contained within Superhelp is posted by the project team and provides a common baseline for all agents.

Data is captured in the 8x8 system in real time and is readily available for reporting purposes on a weekly basis. The project team will e-mail an aggregated transaction report that reflects the following metrics:

- Total number of incoming calls offered
- Total number of incoming calls answered
- Total number of calls abandoned by the caller
- Percent of calls abandoned by the caller
- Average time to abandonment
- Average speed to answer
- Maximum wait time to answer
- Average talk time
- Total number of e-mails and chat sessions answered

The AIR project team will work with the Department to identify additional reporting metrics that the Department wants to see and will then include those reporting metrics on the standard reports.

AIR will provide regular reporting access to the New Hampshire's call log, issue log, and information and performance metrics. Information from Help Desk interactions will be reviewed for program improvements.

Tier 2

When a user presents a question or query that requires more in-depth analysis, the case details are promptly escalated from Tier 1 to Tier 2. This escalation process is continually monitored and, upon receipt of an escalation, a Tier 2 agent is immediately assigned as the owner of the case. The Tier 2 agent reviews the details provided within the 8x8 case notes and initiates contact with the user either by telephone or by e-mail, based on user preference. While the case is open, the Tier 2 agent will reach out to the user as needed to obtain additional information, and will keep the user updated as the case progresses. This contact strategy continues through to case resolution.

Each Tier 2 team is staffed with technical support agents who possess specific areas of expertise that allow them to quickly resolve cases (e.g., network support, testing support, reporting). All Tier 2 agents are required to have a bachelor's degree in an IT-related field.

Tier 2 team activities are supervised by team leads who provide agents additional guidance related to troubleshooting and replication activities, when needed. In addition, they monitor the status of each case to ensure that cases are progressing toward resolution satisfactorily. The team leads report to the Tier 2 senior manager who oversees the day-to-day activities and performance of the overall Tier 2 Help Desk.

Tier 3

When an issue cannot be resolved within Tier 2, the case is escalated to dedicated resources within the project team who manage and track the progression of each case with the appropriate AIR technical team (e.g., network engineer, software engineer). This team is comprised of experienced program management professionals who act as conduits between Tier 2 and the technical teams.

Preventative Maintenance and Upgrade Installations

AIR provides one major version upgrade prior to the beginning of each school year, with minor releases throughout the year as needed to repair problems, add features, or enhance reliability and security. Importantly, these upgrades do not require any changes to the browser at the school. Each year we release a Secure Browser prior to the school year, and this usually does not require any upgrades throughout the year. One exception is when manufacturers release a new operating system, which may require its own version of the Secure Browser. Upgrades to our system are typically invisible to end users (except for the enhanced features to which they may have access). When regular maintenance is scheduled during a mutually agreed-upon time with the Department, the project team can post an announcement to the New Hampshire administrative portal to inform system users. Help Desk agents will also be made aware of any scheduled system downtime for preventative maintenance and upgrade installations. Routine and emergency maintenance procedures are discussed in greater detail in Topic 10 Backup and Recovery.

Topic 39 Support Center

Ensuring Timely and Accurate Assistance

Timely Assistance

To ensure that New Hampshire users have prompt and accurate assistance for all questions and requests, users may contact the AIR Help Desk by calling a dedicated toll-free number, by sending an e-mail, or by using real-time chat sessions with a dedicated agent who is ready to assist by answering questions and providing clarifying information. More information about Help Desk operating hours and methods that may be used to contact the Help Desk is included in Topic 38 Help Desk Support.

As discussed in Topic 38 Help Desk Support, AIR understands the importance of responding to and resolving user requests as quickly as possible. For this reason, our Help Desk is organized in a tiered manner that provides an efficient support structure to assist with the full range of queries received, from routine queries to more in-depth requests. This tiered structure is comprised of Tier 1, 2, and 3 teams staffed with team members who are well trained and knowledgeable about policies and procedures that are specific to the New Hampshire Statewide Assessment program.

Staffing levels within Tier 1 and Tier 2 are continuously monitored to ensure that agent resources are appropriately aligned to call volume and case load so that all cases can be assigned and worked as they are received.

Determining agent schedules and the appropriate staffing levels for Tier 1 and Tier 2 contributes significantly to user satisfaction and the Help Desk's ability to meet service level agreements. To ensure that sufficient staffing capacity is readily available at all times, many of our Tier 1 agents are cross-trained on several projects so that they can provide backup assistance during peak load periods.

Staffing levels are calculated by using historical volumes and call arrival patterns experienced by similar projects. Staffing levels will vary and will be agreed upon with the Department well in advance of the testing window. Staffing levels will be adequate so that individuals are not:

- put on hold for long periods of time;
- unable to reach someone due to busy signals; or
- otherwise unable to receive assistance in a timely manner.

AIR has experience managing assessment programs with large case volumes; our two largest projects have daily averages during the testing window in excess of 600 cases with spikes of over 1,000 per day during a peak month.

The majority of user contacts are resolved at the Tier 1 level with a first call resolution rate of approximately 80%, and an average call duration of under eight minutes. Ninety percent of all cases are resolved in less than 24 hours.

Accurate Assistance

All Help Desk personnel have access to Superhelp, an online knowledge management tool designed to provide consistent responses to calls about the same issue by providing an electronic repository of FAQ documents, guides, manuals, and other resources. The use of this tool is a cornerstone in ensuring that agents provide up-to-date, accurate information to every user during every contact. The content contained within Superhelp is posted by the project team and provides a common baseline for all agents.

Monitor and Document the Efficiency and Accuracy of the Service Provided

As discussed in Topic 38 Help Desk Support, AIR uses 8x8 Virtual Contact Center (Version Package 9.4.1) to log, track, and escalate all telephone calls, e-mails, and chat interactions between Help Desk agents and users. The 8x8 system provides supervisors the ability to listen to recorded calls, monitor real-time call queue statistics, and retrieve case histories sorted by caller. AIR personnel who are not directly associated with the Help Desk, such as project team members and department heads, are made aware of new cases and case updates via e-mail correspondence. When replies are sent using e-mail, they are automatically recorded in the 8x8 case details. All case information and follow-up communication is systematically recorded in 8x8, and the data is captured in real time and readily available for reporting to the Department.

Data is captured in the 8x8 system in real time and is readily available for reporting purposes on a weekly basis. The project team will e-mail an aggregated transaction report that reflects the following metrics:

- Total number of incoming calls offered
- Total number of incoming calls answered
- Total number of calls abandoned by the caller
- Percent of calls abandoned by the caller
- Average time to abandonment
- Average speed to answer
- Maximum wait time to answer

- Average talk time
- Total number of e-mails and chat sessions answered

The AIR project team will work with the Department to identify additional reporting metrics that the Department wants to see and will then include those reporting metrics on the standard reports.

AIR will provide regular reporting access to the New Hampshire call log, issue log, and information and performance metrics. Information from Help Desk interactions will be reviewed for program improvements.

Standards for Performance and Customer Service

AIR is committed to providing the highest quality customer service possible through our Help Desk. Because of this, we have a rigorous recruiting and hiring process to ensure that we have qualified agents and staff personnel assisting the New Hampshire users. All candidates must successfully pass a writing proficiency test, and Tier 1 candidates must pass a computer literacy test and a typing test. If successful, an in-person interview is then conducted by a member of the Help Desk management team. Tier 1 agents must have a minimum two years of previous customer service experience, and Tier 2 agents must have a bachelor’s degree in a field related to information technology or computer science.

AIR invests heavily in Help Desk training for newly hired agents and in existing agents through refresher and project launch classes. This investment ensures that information provided to users is accurate and that agents are able to respond in an efficient, time-sensitive manner.

New-hire training consists of one week of instructor-led classroom training, followed by one week of on-the-job training in which students sit side-by-side with a veteran agent to observe and reinforce the skills learned in the classroom. Courses are delivered by a variety of subject matter experts, including Help Desk managers, project coordinators, team leads, and veteran agents with demonstrated proficiency in a particular area. Topics covered during the classroom portion of new-hire training include

- AIR policies (attendance, schedules, human resources manual);
- 8x8 call handling, case creation, and resolution procedures;
- systems (e.g., 8x8, TIDE, TDS, ORS, AIRWays, Superhelp);
- the Family Educational Rights and Privacy Act (FERPA);
- project administrative portals, policies, and procedures;
- proper escalation and complaint handling procedures;
- quality expectations;
- security awareness (PII handling, badges, entering and exiting facility); and
- hands-on device testing within the training lab environment.

Throughout the new-hire training experience, quizzes and student presentations play an integral part in confirming that agents are building a strong knowledge foundation prior to graduation and progression onto their teams. Refresher training is conducted as needed based upon management observations, system enhancements, procedural changes, quality expectations, or other considerations. Daily communications are broadcast to all project agents, highlighting new or updated information or reminding agents of important project facts. Metrics are closely monitored to ensure that all Tier 1 agents adhere to our internal service level agreements for each transaction. Individual metrics are reviewed with each agent on a weekly basis and in real time when the metric report identifies an issue with an agent. Project launch training is delivered by an AIR New Hampshire project team member prior to testing to ensure that agents are current on all policies and procedures specific to each assessment.

The majority of user contacts are resolved at the Tier 1 level with a first call resolution rate of approximately 80%, and an average call duration of under eight minutes. Ninety percent of all cases are resolved in less than 24 hours.

AIR will provide regular reporting access to the New Hampshire call log, issue log, and information and performance metrics. Information from Help Desk interactions will be reviewed for program improvements.

Processes, Procedures, or Systems Used to Ensure that All Interactions with Districts and Schools are Documented and Maintained in a System that Allows for Efficient Access and Review

As previously described above and in Topic 38 Help Desk Support, data from interactions with districts and schools is documented and maintained in the 8x8 system in real time and is readily available for reporting purposes on a weekly basis. The project team will e-mail an aggregated transaction report that reflects the following metrics:

- Total number of incoming calls offered
- Total number of incoming calls answered
- Total number of calls abandoned by the caller
- Percent of calls abandoned by the caller
- Average time to abandonment
- Average speed to answer
- Maximum wait time to answer
- Average talk time
- Total number of e-mails and chat sessions answered

The AIR project team will work with the Department to identify additional reporting metrics that the Department wants to see and will then include those reporting metrics on the standard reports.

AIR will provide regular reporting access to the New Hampshire call log, issue log, and information and performance metrics. Information from Help Desk interactions will be reviewed for program improvements.

The vendor will provide for provide customer support to districts and schools throughout the registration, testing, and reporting cycles, with an emphasis on service provided at key periods such as registration of students and test administration.

AIR understands the expectation to provide customer support to districts and schools throughout the registration, testing, and reporting cycles, with an emphasis on service provided at key periods such as registration of students and test administration.

Currently, comprehensive support is available to New Hampshire Smarter Balanced Assessment users Monday through Friday from 7:00 a.m. to 4:00 p.m. EST outside of the summative testing window and 7:00 a.m. to 7:00 p.m. EST during the summative testing window. We are prepared to provide comprehensive support to New Hampshire users from the hours of 7:00 a.m. to 5:00 p.m. EST Monday through Friday as stated in the requirements of the RFP if it is the Department's preference. If the Department feels that additional Help Desk support is needed during the student registration period, AIR can support this.

The vendor will guarantee that help desk staffing will increase and/or decrease based on call volume and wait time/caller. When staffing increases/decreases will be determined in consultation with and with approval from the NH DOE management team.

AIR understands and the expectation that Help Desk staffing will increase and/or decrease based on call volume and wait time/caller. Staffing decisions will be made in consultation with and approval from the NH DOE management team.

The vendor will provide help desk and technical support via toll-free phone, e-mail, and/or other online methods Monday through Friday from 7:00 a.m. EST/EDT through 5:00 p.m. EST/EDT. This includes a dedicated technical support line for NH districts, schools, and state representatives.

Currently, comprehensive support is available to New Hampshire Smarter Balanced Assessment users Monday through Friday from 7:00 a.m. to 4:00 p.m. EST outside of the summative testing window and 7:00 a.m. to 7:00 p.m. EST during the summative testing window. We are prepared to provide comprehensive support to New Hampshire users from the hours of 7:00 a.m. to 5:00 p.m. EST Monday through Friday as stated in the requirements of the RFP if it is the Department's preference. If the Department feels that additional Help Desk support is needed during the student registration period, AIR can support this.

To ensure that the Department and all New Hampshire district test coordinators, school test coordinators, schools, users, educators, and test administrators have prompt and accurate assistance for all questions and requests, users may contact the AIR Help Desk by calling a dedicated toll-free customer support line, by sending an e-mail, or by chatting real time with a dedicated agent who is ready to assist. Users may leave a voicemail or send an e-mail outside of the designated Help Desk hours and will receive a prompt response within an hour of when the Help Desk opens the next business day. More information about Help Desk operating hours and methods that may be used to contact the Help Desk is included in Topic 38 Help Desk Support.

Tiered Levels of Customer Support to District and School Administrators and Educators

As described in Topic 38 Help Desk Support, AIR understands the importance of responding to and resolving user requests as quickly as possible. For this reason, our Help Desk is organized in a tiered manner that provides an efficient support structure to assist with the full range of queries received, from routine queries to more in-depth requests. This tiered structure is comprised of Tier 1, 2, and 3 teams staffed with team members who are well trained and knowledgeable about policies and procedures specific to the New Hampshire Statewide Assessment program.

Tier 1 agents are dedicated by project and are trained to answer routine questions and to provide explanations and clarifications using scripted and semi-scripted responses. The agents will respond to routine questions with answers that have been approved by the Department. Agents will address inquiries specific to user roles, training, registration, reporting, and test administration processes as included in the associated manuals and ancillary materials. Tier 1 team activities are supervised by a dedicated Help Desk project coordinator who will be a subject matter expert on all New Hampshire assessment policies and procedures. Project coordinators report directly to the assistant site manager who oversees the day-to-day activities and performance of the overall Tier 1 Help Desk.

When a user presents a question or query that requires more in-depth analysis, the case details are promptly escalated from Tier 1 to Tier 2. This escalation process is continually monitored and, upon receipt of an escalation, a Tier 2 agent is immediately assigned as the owner of the case. Each Tier 2 team is staffed with technical support agents who possess specific areas of expertise that allow them to quickly resolve cases (e.g., network support, testing support, reporting). All Tier 2 agents are required to have a bachelor's degree in an IT-related field. This contact strategy continues through to case resolution.

When an issue cannot be resolved within Tier 2, the case is escalated to dedicated Tier 3 resources within the project team who manage and track the progression of each case with the appropriate AIR technical team (e.g., network engineer, software engineer). This team is comprised of experienced program management professionals who act as conduits between Tier 2 and the technical teams.

Please see Topic 38 Help Desk Support for additional information on tiered structuring of Help Desk agents.

The vendor and state will agree upon the type of questions and issues that will be addressed by the vendor, what actions the support center and other vendor staff will take to resolve and/or answer those questions and issues, and the type of questions and issues that will be forwarded to the NH DOE for resolution.

AIR acknowledges the need to agree upon the type of questions and issues that will be addressed by the vendor, what actions the support center and other vendor staff will take to resolve and/or answer those questions and issues, and the type of questions and issues that will be forwarded to the Department for resolution. At a minimum, AIR will address inquiries specific to administration processes as included in the associated manuals and ancillary materials.

The AIR New Hampshire project team will work closely with the Department to develop and obtain approval on all resource materials—such as forms, scripts, user guides, etc.—that will be used by the Help Desk. Upon approval, the AIR New Hampshire project team will work in concert with the Help Desk management team to determine the optimal dates for training Help Desk staff on New Hampshire Statewide Assessment specifics and to post all resource materials to Superhelp, our online knowledge management tool.

All Help Desk personnel have access to Superhelp, an online AIR-developed repository that utilizes a Google search appliance and allows Help Desk agents to quickly locate applicable approved responses and material. This tool was designed to ensure that all agent responses to the same issues are answered consistently every time. AIR developed Superhelp during the spring of 2015, and it is used in support of all AIR Assessment projects, including New Hampshire's Smarter Balanced Assessments.

This electronic repository includes all FAQ documents, guides, manuals, and information found on the New Hampshire Statewide Assessment Portal (administrative portal). The New Hampshire Superhelp knowledge base will be updated and maintained throughout the entirety of the testing window just as has been done in support of the New Hampshire Smarter Balanced Assessments. Our agents will provide consistent responses to the same issues coming in from the field.

New inquires and approved responses are uploaded into Superhelp, and the tool is a cornerstone in ensuring that agents provide up-to-date, accurate information to every user during every contact. The content contained within Superhelp provides a common baseline for all agents. Approved information loaded into the knowledge database can be easily located by our agents using the search functionality.

Support Center Staff Ability to Reopen Accidentally Closed Tests

Help Desk staff may be granted access to TIDE and will be able to submit an appeal to reopen closed tests through TIDE on the user's behalf for NH DOE and project team review/approval. The Help Desk agent user system permissions and access to TIDE is configurable based upon the preference of the Department.

Qualifications and Training of Support Center Staff and Other Vendor Staff

AIR is committed to providing the highest quality customer service possible through our Help Desk. Because of this, we have a rigorous recruiting and hiring process to ensure that we have qualified agents and staff personnel assisting the New Hampshire users. All candidates must successfully pass a writing proficiency test, and Tier 1 candidates must pass a computer literacy test and a typing test. If successful, an in-person interview is then conducted by a member of the Help Desk management team. Tier 1 agents must have a minimum two years of previous customer service experience, and Tier 2 agents must have a bachelor's degree in a field related to information technology or computer science.

AIR invests heavily in Help Desk training for newly hired agents and in existing agents through refresher and project launch classes. This investment ensures that information provided to users is accurate and that agents are able to respond in an efficient, time-sensitive manner.

New-hire training consists of one week of instructor-led classroom training, followed by one week of on-the-job training in which students sit side-by-side with a veteran agent to observe and reinforce the skills learned in the classroom. Courses are delivered by a variety of subject matter experts, including Help Desk managers, project coordinators, team leads, and veteran agents with demonstrated proficiency in a particular area. Topics covered during the classroom portion of new-hire training include

- AIR policies (attendance, schedules, human resources manual);
- 8x8 call handling, case creation, and resolution procedures;
- systems (e.g., 8x8, TIDE, TDS, ORS, AIRWays, Superhelp);
- the Family Educational Rights and Privacy Act (FERPA);
- project administrative portals, policies, and procedures;
- proper escalation and complaint handling procedures;
- quality expectations;
- security awareness (PII handling, badges, entering and exiting facility); and
- hands-on device testing within the training lab environment.

Throughout the new-hire training experience, quizzes and student presentations play an integral part in confirming that agents are building a strong knowledge foundation prior to graduation and progression onto their teams. Refresher training is conducted as needed based upon management observations, system enhancements, procedural changes, quality expectations, or other considerations. Daily communications are broadcast to all project agents, highlighting new or updated information or reminding agents of important project facts. Metrics are closely monitored to ensure that all Tier 1 agents adhere to our internal service level agreements for each transaction. Individual metrics are reviewed with each agent on a weekly basis and in real time when the metric report identifies an issue with an agent. Project launch training is delivered by an AIR New Hampshire project team member prior to testing to ensure that agents are current on all policies and procedures specific to each assessment.

Service Level Agreement (SLA)

Availability

Help Desk Availability

AIR Help Desk staff will be available through the methods described and during the times indicated above and in Topic 38 Help Desk Support.

System Availability

Our test delivery system (TDS) consists of hundreds of servers organized into three constellations. Each constellation is shared by a set of clients and can be scaled virtually infinitely. Currently, each of our constellations is designed to support 400,000–500,000 students signing on within 20 minutes of each other and testing simultaneously. Across constellations, our capacity approaches 1.5 million simultaneous users. Each year, we project peak loads and have the capability to scale any constellation as much as necessary. We discuss disaster recovery plans in Topic 10 Backup and Recovery.

Planned maintenance downtime to AIR systems will be coordinated with the Department, and mutually acceptable windows for system updates and improvements will be negotiated.

Reliability

AIR is proud to note that our systems operate reliably, without interruption. Our systems incorporate redundancies and fail-safes at every potential point of failure, and we maintain redundant data stores in the case of a catastrophic event. We are able to deliver tests with virtually no systemic interruptions.

Below, we propose several service-level agreements, which we will be pleased to tailor to the Department's needs. AIR makes the following commitments for level of service:

- Our TDS, our Test Information Distribution Engine (TIDE, our administrative system), and our Online Reporting System (ORS) will be available 99.99% of the scheduled uptime. This excludes agreed-upon maintenance periods. A system will be considered unavailable if more than 10% of users cannot access the system.
- The design of our TDS limits the impact of adverse events. We can make the following commitments:
  - Interruptions in the login system will be resolved within 30 minutes. We anticipate no interruptions in the login system, and availability of the TDS login system is included in the uptime commitment.
  - Interruptions in service delivery affecting fewer than 10% of users will be resolved within 30 minutes.
- Our system will support up to 100,000 simultaneous New Hampshire users as required. Typically, we assume that the number of concurrent users log in over a 20-minute period.

We offer two limitations on our service level commitments, both of which limit our commitment to factors within our control. First, we will integrate with the NH DOE single sign-on (SSO) system. If the NH DOE SSO system is down, users will be unable to access the system. Second, we cannot assume responsibility for outages caused by third party internet service providers, district users, etc. Finally, we have state-of-the-art defenses against distributed denial-of-service (DDoS) attacks, and these have successfully fended off malicious attacks without interruption numerous times; however, some very large DDoS attacks can overwhelm even the best defenses, and one recently shut down large swaths of the Internet. If the Department chooses to negotiate with us, we will share confidential information about our defenses and the limits of the attack size that we can successfully mitigate.

Latency

Once the tests are deployed, AIR's servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first sign that trouble may be ahead. Applications log not only errors and exceptions, but latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed or if it is starting to slow down or experience a problem.

In addition, latency data is captured for each assessed student, including data about how long it takes to load, view, or respond to an item. All of this information is logged as well, enabling us to automatically identify schools or districts that may be experiencing unusual slowdowns, often before they even notice.

AIR's Help Desk also continuously monitors any reports from the field for any issues with items or software. Any reports of issues are escalated using defined procedures as discussed in Topic 38 Help

Desk Support and Topic 39 Support Center, and the AIR Project Manager is kept informed of the status. Issues that interfere with the delivery of tests are given top priority by the Help Desk.

Disaster Recovery Plan

As described in Topic 10 Backup and Recovery, AIR is continuously improving our ability to protect our systems from interruptions. Data integrity is paramount. Our system is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. Each system is redundant, and critical student response data is transferred to a different data center each night.

We have developed a unique monitoring system that is very sensitive to changes in server performance. Most monitoring systems provide warnings when something goes wrong. Ours does, too, but it also provides warnings when any given server is performing differently from its performance over the prior few hours, or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them *before* a failure.

For example, in 2016, the system alerted us to a drive that had some sectors going bad. On multiple occasions, our monitoring system has enabled us to make adjustments and replace equipment before any problems occurred.

We have an escalation procedure in place that enables us to alert clients within minutes of any disruption. Our emergency alert system notifies our executive and technical staff by text message, who immediately join a call to understand the problem. Our VP of Assessment Programs and Client Services is among the participants, and she is charged with ensuring that clients are promptly notified if there is (or is likely to be) impact on the field. Please see AIR's Business Continuity Plan Addressed in Topic 11 Assurance of Business Continuity.

Server Backup Plan

As described in Topic 11 Assurance of Business Continuity, AIR has a comprehensive server backup plan. Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure the safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and proven real-time backup processes prevent loss of student data, even in the unlikely event of system failure. We refer readers to Topic 11 Assurance of Business Continuity for more information on our server backup plan.

Recovery Point Objective

Our disaster recovery procedures are described above, in Topic 10 Backup and Recovery, and in Topic 11 Assurance of Business Continuity.

Issue Resolution Times

The Help Desk staff will be able to identify which questions that are matters of policy and belong to the Department and which are to be answered directly by AIR staff. AIR Help Desk agents will make initial contact regarding any inquiries within 24 hours of receipt; during testing windows, the response time will be within two hours.

The majority of user contacts are resolved at the Tier 1 level with a first call resolution rate of approximately 80%, and an average call duration of under eight minutes. Ninety percent of all cases are resolved in less than 24 hours.

Maintenance Windows

Planned maintenance downtime to AIR systems will be coordinated with the Department. Mutually acceptable windows for system updates and improvements will be negotiated.

Service Reporting

As described above, the AIR New Hampshire project team will work with the Department to establish a reporting schedule as requested. Typically, AIR supplies daily reports throughout the registration window and testing window. The AIR New Hampshire project team will e-mail an aggregated transaction report that reflects the following metrics:

- total number of incoming calls received;
- total number of incoming calls answered;
- total number of calls abandoned by the caller;
- percentage of calls abandoned by the caller;
- average time to caller abandonment;
- average speed to answer;
- maximum wait time to answer;
- average talk time; and
- total number of e-mails and chat sessions answered.

AIR will work with the Department to provide other Help Desk metrics and information as required for the monthly Help Desk report upon request based on experience with the performance of the Help Desk and/or unanticipated types or volumes of support calls. AIR typically provides the following:

- total number of cases for a specific time period;
- percentage of cases resolved on the first call;
- percentage of cases resolved within 24 hours and 72 hours;
- percentage of cases escalated to Tier 2; and
- percentage of cases assigned by category (we classify roughly 100 unique category types).

Support Hours

As described in Topic 38 Help Desk Support and Topic 39 Support Center above, currently, comprehensive support is available to New Hampshire Smarter Balanced Assessment users Monday through Friday from 7:00 a.m. to 4:00 p.m. EST outside of the summative testing window and 7:00 a.m. to 7:00 p.m. EST during the summative testing window. We are prepared to provide comprehensive support to New Hampshire users from the hours of 7:00 a.m. to 5:00 p.m. EST Monday through Friday as stated in the requirements of the RFP if it is the Department’s preference. If the Department feels that additional Help Desk support is needed during the student registration period, AIR can support this. Users may leave a voicemail or send an e-mail outside of these hours, and AIR staff will promptly respond when the Help Desk opens the next business day.

Support Contact Information

As discussed in Topic 38 Help Desk Support and Topic 39 Support Center above, users may contact the AIR Help Desk by calling a dedicated toll-free customer support line, by sending an e-mail, or by chatting real-time with a dedicated agent who is ready to assist. Users may leave a voicemail or send an e-mail outside of these hours, and AIR staff will promptly respond when the Help Desk opens the next business day.

Escalation

AIR understands the importance of responding to and resolving user requests as quickly as possible. For this reason, our Help Desk is organized in a tiered manner that provides an efficient support structure to assist with the full range of queries received, from routine queries to more in-depth requests. This tiered structure is comprised of Tier 1, 2, and 3 teams staffed with team members who are well trained and knowledgeable about policies and procedures that are specific to the New Hampshire Statewide Assessment program. The AIR project team will collaborate with the Department to determine the degree of direct contact with districts and schools and which inquiries will be directed to the Department to address.

It is important to note that all telephone conversations between New Hampshire users and Tier 1 and Tier 2 agents are automatically recorded and captured in 8x8. If requested, AIR is able to forward any recorded conversation to the Department, and we are able to use the Tier 2 responses to help with subsequent training of Tier 1 staff on new issues.

Tier 1

Tier 1 agents are dedicated by project and are trained to answer routine questions and to provide explanations and clarifications using scripted and semi-scripted responses. Agents will address inquiries specific to user roles, training, registration, reporting, and test administration processes as included in the associated manuals and ancillary materials. Examples of types of support include user role issues, application login, location of manuals and other forms of documentation, access to score reports, Secure Browser installation, test proctor and student system interfaces, resolution of system access issues that might occur in a school, and special case processing such as test invalidations or test resets. Help Desk staff will be able to identify which questions are matters of policy and belong to the Department and which are to be answered directly by AIR staff. AIR Help Desk agents will make initial contact regarding any inquiries within 24 hours of receipt; during testing windows, the response time will be within two hours.

The majority of user contacts are resolved at the Tier 1 level with a first call resolution rate of approximately 80%, and an average call duration of under eight minutes. Ninety percent of all cases are resolved in less than 24 hours.

Tier 1 team activities are supervised by a dedicated Help Desk project coordinator who will be a subject matter expert on all New Hampshire assessment policies and procedures. Project coordinators report directly to the assistant site manager who oversees the day-to-day activities and performance of the overall Tier 1 Help Desk.

Tier 2

When a user presents a question or query that requires more in-depth analysis, the case details are promptly escalated from Tier 1 to Tier 2. This escalation process is continually monitored and, upon receipt of an escalation, a Tier 2 agent is immediately assigned as the owner of the case. The Tier 2 agent reviews the details provided within the 8x8 case notes and initiates contact with the user either by telephone or by e-mail, based on user preference. While the case is open, the Tier 2 agent will reach out to the user as needed to obtain additional information and will keep the user updated as the case progresses. This contact strategy continues through to case resolution.

Each Tier 2 team is staffed with technical support agents who possess specific areas of expertise that allow them to quickly resolve cases (e.g., network support, testing support, reporting). All Tier 2 agents are required to have a bachelor’s degree in an IT-related field.

Tier 2 team activities are supervised by team leads who provide agents additional guidance related to troubleshooting and replication activities, when needed. In addition, they monitor the status of each case to ensure that cases are progressing toward resolution satisfactorily. The team leads report to the Tier 2 senior manager who oversees the day-to-day activities and performance of the overall Tier 2 Help Desk.

Tier 3

When an issue cannot be resolved within Tier 2, the case is escalated to dedicated resources within the project team who manage and track the progression of each case with the appropriate AIR technical team (e.g., network engineer, software engineer). This team is comprised of experienced program management professionals who act as conduits between Tier 2 and the technical teams. Typically, Tier 3 cases are resolved by our most senior engineering staff.

Examples of Tier 3 events include major loss of connectivity impacting many students, loss of shipments or student data files, item/test security breach, data breach due to cyber-attack, or similar circumstances. The AIR New Hampshire project team will work with the ADE well in advance of test administration to establish notification protocols for various types of potential Tier 3 events and to ensure that the NH DOE and the New Hampshire Statewide Assessment field are notified according to the established protocols.

In the event of severe occurrences (such as administration irregularities, test security violations, etc.), AIR staff will immediately contact the Department. The Department will locate the preferred solution and provide resolution guidance or status with e-mail response and a return telephone call to the AIR Help Desk within two (2) working hours of the original service request and every 24 hours thereafter until a solution is found. AIR will work with the Department to establish these protocols and establish lines of direct contract.

In the rare event of a major system disruption, a member of the Help Desk management or the technology Network Operations Center (NOC) team will send an automated “Mission Mode” alert to our most senior AIR management and technical staff instructing those notified to immediately call into a conference bridge. This immediate alert process ensures that key stakeholders have the ability to collaboratively diagnose the issue and agree upon next steps to resolve the issue.

Errata Notice Template

AIR understands the requirement specified in the RFP to develop an errata notice template that includes a description of the issue, the timeline for resolution, and any required actions that need to be taken by district or school administrators and/or test administrators.

Please find a sample Errata Notice Template in Appendix F. The look, feel, and content of this template may be altered according the requirements of the Department and will be determined upon award of the contract during the initial kick-off meeting.

Change Management

The purpose of change management is to ensure that changes to requirements, specifications, schedules, etc., are communicated, managed, and controlled. This is accomplished by working with the Department to clearly define the change control processes during the planning stage so that when the inevitable change requests surface, there will be no confusion about roles, responsibilities, or process requirements. Otherwise, uncontrolled or mismanaged changes can negatively impact budget, scope, schedules, and possibly quality.

Weekly conference call meetings between AIR and the Department cover near-term deliverables and address issues as they arise. In addition, the AIR New Hampshire Project Manager will produce weekly reports for the Department that will include the status of items (such as deliverables), key accomplishments, upcoming activities, any changes to the schedule, or any issues. AIR will deliver the weekly status report via e-mail. Decisions are documented in the meeting minutes and in the issues log, which is updated weekly and tracks issues and near-term deliverables. It notes progress, problems, risks, and assignments.

Changes are reflected in the appropriate communication documents through a carefully coordinated process. The project manager is responsible for communicating the changes to the document owner, who reviews the changes and amends the document. The document is then routed to each stakeholder who has subscribed to the document. Changes to the document are not final until they receive sign-off from each stakeholder. The document owner is responsible for ensuring that the changes are accurately drafted, reviewed, and accepted by each stakeholder, including the Department. A change will not be implemented until the NH DOE has signed off on the change in writing.

> ***The vendor must include a plan for timely electronic notification to district and school administrators and test administrators through e-mail, posting a notice on the online system, and/or direct calling, of any issues affecting test administration.***

AIR understands the requirement to include a plan for timely electronic notification to district and school administrators and test administrators through e-mail, posting a notice on the online system, and/or direct calling, of any issues affecting test administration.

With the New Hampshire Smarter Balanced Assessments, AIR maintains an online administrative portal that is a comprehensive location to distribute accurate and timely information to district, school, and test administrators. In the event that an issue arises, AIR has the ability to post an announcement to the online administrative portal describing the issue and providing information to users as it becomes available. All verbiage included in the announcement will receive NH DOE approval before the announcement is posted to the portal. Once announcements are posted to the portal, registered users will receive an e-mail notification immediately.

Additionally, the Test Administrator Interface (described in further detail Topic 4 Assessment Delivery Platform) allows the Department and AIR to communicate immediate issues that affect test administration to Test Administrators who are logging in or are currently logged in to the system.

Topic 40 Technical Reporting

As we describe in Section D2.3 Project Plan, our team leads work with our clients to develop detailed specification documents. Specification documents function as the detailed requirements documents for each phase of the program. For example, our reporting team will work with the Department staff to develop a detailed reporting specifications document that includes lists of reports to be produced, descriptions of report recipients, packaging specifications for paper reports, calculation rules for each data element, and other details of the reporting system. This document is usually created in collaboration with the lead psychometrician and the technical, content, and leadership staff at the Department. Likewise, the lead psychometrician will work with the Department to finalize a set of scoring specifications that govern the scoring of student responses and analysis specifications that detail all procedures for calculating classical item statistics, evaluating differential item functioning, calibrating item response theory parameters, establishing the reporting scales, and so on.

Each document has an owner and a list of stakeholders. The owner is the person responsible for keeping the document up to date and ensuring that all stakeholders sign off on any changes to the document. Stakeholders include project team members whose work depends directly or indirectly on the work described in the document. In almost every case, client staff members are stakeholders in these documents as well.

Given the complex nature of assessment projects, the multitude of documents alone, not including the review and approval process of iterations of these documents, can be a challenge to maintain using traditional share-drive options found in most organizations. Therefore, AIR proposes that all major project documents, in particular management documents identified in Exhibits D2.3-1 through D2.3-4 in Section D2.3 Project Plan, be stored in KnowledgeTree.

KnowledgeTree is an online document management system on a secure cloud computing platform. KnowledgeTree manages document workflow and file sharing, alerts document followers to changes, and maintains version control, including authors and document history. Documents are checked in and checked out for editing. Access to documents can be controlled through a user roles and permissions system.

Topic 40.1 Technical Report

A technical report will be delivered to the Department each year following the administration of New Hampshire's statewide assessments. The technical report will be provided to the Department electronically for review by their TAC. Upon final approval of the technical report by the Department, a PDF version of the technical report will be provided to the Department for posting on the assessment portal. AIR commits to providing a draft report to the Department no later than three months following release of all assessment results from the operational test administration.

Exhibit D3.2-1 provides the proposed outline for the New Hampshire technical report. Appendix G provides Arizona's AzMERIT technical report based on a similar structure.

The report is structured so that some chapters require only minimal changes after the first year of the contract. For example, chapters such as Performance Standards, which documents the standard setting process, and Reporting and Interpreting New Hampshire Scores are expected to remain essentially the same for the duration of the contract. The Summary of Operational Test Administration, on the other hand, will be updated every year because it contains the results of the statistical and psychometric analyses of that year's assessment. For this reason, the structure allows for an optimal use of our and the Department's resources in reviewing and copy-editing the extensive technical documentation. Importantly, it also allows readers to focus attention each year on variable elements of the report.

The central focus of the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014) is on the validity of test score interpretations. The technical report opens with documentation of evidence for the validity of intended test score interpretations. Because validity evidence accrues over time, this chapter will be expanded as additional studies are completed. The reliability of test scores and performance-level classifications, both overall and for subscales, is documented for each test administration. As noted, procedures for configuring adaptive test administrations will be documented, as well as procedures for constructing fixed-form assessments that are equivalent across test administrations.

AIR is committed to working with the Department to provide all the technical documentation necessary to complete the peer review process successfully. To support the Department and their TAC in the development and review of peer review evidence, AIR will provide the Department with a crosswalk between the peer review guidelines and the technical report sections that provide the required evidence.

Exhibit D3.2-1: Proposed Outline for the New Hampshire Technical Report

1. **Introduction: The Validity of New Hampshire Test Score Interpretations**
   - 1.1 Overview
   - 1.2 Validity Evidence
   - 1.3 Evidence Based on Test Content
   - 1.4 Evidence for Interpretation of Performance
   - 1.5 Evidence Based on Internal Structure
   - 1.6 Evidence for Relationships with Conceptually Related Constructs
   - 1.7 Evidence for Student Growth—Overall and by Subgroup
   - 1.8 Summary of Validity of Test Score Interpretations
2. **Background of New Hampshire Statewide Assessments**
   - 2.1 Development of New Hampshire Content and Performance Standards
   - 2.2 New Hampshire Test Design
3. **Description of the Adaptive Algorithm**
   - 3.1 Objective of the Item Selection Algorithm
   - 3.2 Summary of Simulation Studies
     - 3.2.1 Bias
     - 3.2.2 Match to Blueprint
     - 3.2.3 Projected Item Exposure Rates
     - 3.2.4 Summary of Standard Errors
4. **Summary of Operational Test Administration**
   - 4.1 Student Population and Participation
   - 4.2 Classical Item Analysis
   - 4.3 Item Response Theory Analysis
   - 4.4 Summary of Overall Student Performance
   - 4.5 Student Performance by Subgroup
   - 4.6 Reliability
     - 4.6.1 Internal Consistency
     - 4.6.2 Standard Error of Measurement
     - 4.6.3 Student Classification Reliability
     - 4.6.4 Reliability for Sub-Groups in the Population
     - 4.6.5 Subscale Reliability
   - 4.7 Subscale Intercorrelations
   - 4.8 Rater Effects
5. **Item Development and Fixed-Form Test Construction**
   - 5.1 Item Development Process
     - 5.1.1 Item Writing
     - 5.1.2 Machine-Scored Constructed-Response Item Development Tools
     - 5.1.3 Item Types

Topic 41 Technical Advisory Committee

AIR acknowledges our understanding of the requirements related to the Technical Advisory Committee (TAC). AIR psychometricians regularly work with state departments of education and their TACs to identify measurement issues of concern, design and execute studies to investigate research questions, and report those results to the state and TAC. AIR's project manager, lead psychometrician, and other psychometric staff as necessary will attend all TAC meetings. To support the Department, AIR will be responsible for all planning for two TAC meetings each year. AIR will arrange travel and lodging for our staff and for the TAC members. We will handle all travel, lodging, per diem, and other logistics associated with the meeting. AIR will be responsible for all costs associated with the meetings. AIR will process per diems and reimburse committee members' expenses according to mutually agreed-on guidelines. While the Department will be responsible for selecting TAC members and facilitating TAC meetings, AIR psychometricians will be pleased to provide any assistance the Department might desire in facilitating TAC meetings.

D-4 PRICING

Topic 42 Pricing Model

AIR has included a comprehensive response to the requirements of Topic 42 Pricing Model in addition to the requirements of Section VII: Pricing Model (in response to the Appendix F worksheet included as an attachment to the RFP) under a separate Cost Proposal cover.

D-5 ASSURANCE AND TRANSITION

Topic 43 Quality Control and Sign-Offs

Sign-Offs

AIR understands the importance of documenting of NH DOE sign-offs for all deliverables. AIR will make every effort to be flexible, provide information as required, and keep this important project on track. Upon award of the contract, AIR will work with the Department to develop a Communications Plan that specifies

- who needs to receive specific information;
- when that information needs to be communicated (on a specific date or at regular intervals);
- the person responsible for communicating
- the information;
- the method used to communicate the information; and
- the processes for documentation that the information has been communicated and sign-offs have been received (if applicable).

AIR's Communications Plan is described in greater detail in described in further detail in Topic 33 Project Communication.

Quality Control

AIR understands that quality assurance and quality control are a key component of the successful delivery of these tests. We enforce quality assurance and quality control procedures through all stages of test development, test administration, scoring, and reporting of results. Below we describe our procedures associated with:

- Item development
- Test development/form building
- IT system development
- Test administration
- Test scoring
- Data file development
- IT system deployment
- Reporting

Item Development

AIR's test development staff and structure ensure the quality of items and test forms by enforcing a structured review process and formalizing feedback to our staff and item writers to continuously improve content and share ideas. We discuss our item review process in detail in Topic 2 Item Development.

Test Development/Form Building

Quality Assurance. As we discuss fully in Topic 20.3 Test Construction Evaluation, configuration of adaptive test administrations is conducted using our simulation tool which allows psychometricians to model the performance of the adaptive algorithm to ensure that all test administrations match blueprint while targeting test information to student ability, for both overall and reporting category ability estimates, as well as monitoring item usage and coverage of the content domain.

Construction of fixed form tests is equally rigorous with form construction managed through AIR's FormBuilder software which automates important form construction activities to ensure development of equated test forms. FormBuilder interfaces with AIR's Item Tracking System (ITS) to extract test information and interactively creates test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement Curves (SEMCs) as test developers build a test map. This helps our content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism. FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint have been satisfied. In addition, the FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms.

In addition, all bookmaps (test maps), key files, and conversion tables are produced directly from FormBuilder to eliminate the possibility of human error in the construction of these important files. Bookmaps, key files, conversion tables, and other critical documents are generated directly from information maintained in ITS. The information stored in ITS is rigorously reviewed by multiple skilled reviewers, to protect against errors. Automated production of these critical files (such as key files) virtually eliminates opportunities for errors.

Quality Control. AIR's test delivery system also has a real-time quality-monitoring component built in. As students are administered assessments, data flow through the test delivery system's Quality Monitor (QM) software. QM conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. QM scores the test, recalculates performance level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

QM also aggregates data to detect problems that become apparent only in the aggregate. As we describe in Topic 20.2 Operational Test Item Evaluation, QM monitors item fit and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem.

IT System Development

Although no software is being developed under this contract, AIR will be deploying our proven item banking, student registration, test delivery, quality assurance, reporting and related systems to deliver the testing services that are the subject of this project. New Hampshire can be assured that the internal software development process behind these systems conform to a structured systems development life cycle methodology:

- Requirements gathering: The requirements gathering phase is an iterative process that drives the remainder of the development process by providing the specific details of the system.
- System development: The system development phase is an iterative process that takes the larger project and divides the work into manageable pieces (iterations). All software components are unit tested and integration tested before moving them to system testing.
- System testing: This phase includes the system testing, load testing, and performance testing functionality. Testing is well documented to the overall requirements and executed using repeatable scripts, which allow iterations to be tested quickly and efficiently.
- System implementation: During the system implementation phase, the system is migrated to a production environment. This phase includes all training, system maintenance, and support plans and documentation.

Our disciplined approach is requirements-driven and iterative in nature. A significant component to this successful approach is the quality assurance activities that are integrated into each step of the process. All system requirements and software programs are thoroughly documented to ensure testability and compliance. As discussed previously, AIR also maintains a rigorous system to managing changes to software.

Test Administration

Quality Assurance

Quality assurance for online testing starts with platform review and User Acceptance Testing (UAT) of the Test Delivery System (TDS). AIR's developers and statisticians also do load testing to ensure that the online testing platform can support the anticipated number of students simultaneously testing.

Platform Review. Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on approximately 15 platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to see that it renders as expected.

User Acceptance Testing (UAT). Prior to deployment, the testing system and content are deployed to a staging server where they are subject to user acceptance testing. UAT of the test delivery system serves both a software evaluation and content approval role. The UAT period provides the NH DOE with an opportunity to interact with the exact test with which the students will interact. UAT is discussed further in Topic 37 User Acceptance Testing.

Load Testing. Load testing of AIR's servers is generally done over the summer when the fewest tests are live. AIR statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and AIR contracts for service in excess of this amount.

Quality Control

Once the tests are deployed, AIR's servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. Applications log not only errors and exceptions, but latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem.

In addition, latency data is captured for each assessed student—data about how long it takes to load, view, or respond to an item. All of this information is logged as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

AIR's Help Desk, discussed in Topics 38 Help Desk Support and 39 Support Center of this proposal, also continuously monitors any reports from the field for any issues with items or software. Any reports of issues are escalated as described in Topic 38 Help Desk Support and the AIR project manager is kept informed of the status. Issues that interfere with the delivery of tests are given top priority by the Help Desk.

Test Scoring

Quality Assurance. AIR verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the state. The ability of each of these simulated students is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they provide a check of the full range of item responses and test scores in fixed-form tests as well. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

Specifications for generating simulated data files are included in the Analysis Specifications document that AIR will submit to the NH DOE for review each year. Review of all simulated data is scheduled to be completed prior to the opening of the test administration, so that the integrity of item administration, data capture, item and test scoring and reporting can be verified before the system goes live.

Ensuring the validity of test score interpretations also requires monitoring the accuracy of machine assigned scores of students’ responses to writing performance tasks. In addition to assigning dimension scores based on the prediction model, the Autoscore software produces an index reporting the confidence of the assigned score. For example, a student’s response may be very near the threshold between two score points, and thus the assignment of a score in either direction cannot be confidently made. Moreover, a student’s response may include features that deviate significantly from those encountered in the training set, which also reduces confidence in the assigned score. AIR has proposed a process by which low confidence scores are automatically routed for verification by human readers. Test scores for flagged records will not be made available for reporting until after the human verification read has been processed and the test record is allowed to pass through to the database of record. This verification process has been implemented in multiple state assessment systems and provides an important level of quality assurance in the scoring of essay responses.

AIR utilizes experienced staff and a state-of-the-art web-based system to score essay responses routed for verification by human raters. AIR’s handscoring system integrates seamlessly with both our online and paper scoring system. The system manages the scoring process with a broad range of configurable business rules that manage the workflow, assignment of scores, and quality procedures. An AIR scoring director establishes scoring teams and assigns items to them. Each team is led by a supervisor who is responsible for managing the teams’ quality, reliability, and throughput. The system provides the supervisor with powerful tools to monitor performance and quality and to review the work of team members.

Hand-scored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the quality monitor (QM) system where the integrated record is passed for scoring. Once the integrated scores are sent to the QM, the records are rescored in the test-scoring system, a mature, well-tested real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculating performance- level indicators, subscale scores and other features, which then pass automatically to the reporting system and Database of Record (DoR). The scoring system is tested extensively prior to deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

Quality Control. To monitor the performance of the assessment system during the test administration window, a series of Quality Assurance Reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational test window. In the context of adaptive test administrations, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to specifications.

An additional set of forensic analysis reports flags unlikely patterns of behavior in testing administrations aggregated at the test administration, test administrator, and school level that may indicate cheating. The quality assurance reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the test window to ensure that test administrations conform to blueprint and items are performing as anticipated.

The suite of Quality Assurance Reports is described in detail in Topic 20.2 on Operational Test Item Evaluation.

Data File Development

Quality Assurance. The layout for data files will be determine in conjunction with the NH DOE.Once the data file layout has been locked down, AIR uses simulated test data to conduct a series of quality assurance checks on both the accuracy of scoring computations and the fidelity of the data file deliverables to the specified layout, including intake of data files by the NH DOE. These activities are completed well in advance of operational scoring and reporting activities.

Quality Control. The first operational data files delivered to the NH DOE go through a series of internal quality control steps. AIR's SAS team independently recreates the data files and compares them to the files that were automatically generated. The SAS team also checks the files against the layout specifications, verifying that all variables are within the specified ranges. Once the first files are verified as accurate, future files are generated automatically and posted for the department.

IT System Deployment

AIR is proposing existing, proven, highly configurable systems for administering New Hampshire's assessment system. AIR has automated procedures to deploy the system very quickly after the system configuration specifications are finalized. Our project team will work with department staff to make decisions about the features and use of the system, and we will deploy the system for user acceptance testing. We understand that the department must approve the system to go live. AIR will share stress-test results with the department. The user acceptance period provides an opportunity for the NH DOE to check everything about the test prior to going live.

Reporting

Quality Assurance. Scores for online assessments are assigned by automated systems in real time. For machine scored item responses, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field-testing. The review process "locks down" the item and rubric when the item is approved for web display (Web Approval).

Quality Control. During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

After passing through the series of validation checks in the QM system, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring there is only one place where the "official" record is stored. Only after scores have passed the QM checks and are uploaded to the DoR are they passed to the Online Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the Online Reporting System until it passes all of the QM system's validation checks.

Topic 44 Invoices

AIR will submit invoices according to the procedures and requirements set forth by the Department. We will agree to a payment schedule for this contract with four quarterly payments and one final payment for the services performed and deliverables provided during each period that conforms to the State fiscal year and to specific dates for the final invoice for each fiscal year and each assessment cycle.

Topic 45 Transition

The successful transition of a testing program entails transferring all required data, products, knowledge, and other state-owned assets. The transfer must be complete and correct, ensure that the tests remain accurately equated over time, and avoid disruption of services or support during the transition period.

AIR has learned from experience that a successful transition has four core requirements:

- **Identification:** The first step in developing a transition plan is to clearly identify all requisite transition information, as well as the data and format in which the data will be delivered.
- **Planning:** Highly detailed plans chronicling specific delivery deadlines for transferred information facilitate a smooth transition.
- **Delivery:** Once critical transition elements are identified and delivery timelines are established, the actual information transfer must be monitored to ensure that all information delivered is both accurate and timely.
- **Quality Control:** To ensure successful completion, the transitioning contractor must support the new contractor in replicating work and run analyses to ensure that data and trends are accurate.

Successful transition can be facilitated through an ongoing transfer of program documents, data, and content from AIR to the Department and/or new vendor. This will ensure that the Department has access to, and control of, the necessary products for your system. In addition, we recognize that integrating the Department data and content with the incoming vendor's systems will require support from AIR. We have been on both sides of assessment program transitions, and we understand the factors that can contribute to a successful transition.

In transitioning to a new vendor, AIR will work with the Department to create a transition plan that meets the requirements outlined in the RFP.

Below, we outline key transition meetings that help facilitate the transition process.

Transition Specifications and Meetings

AIR will work with the Department and the incoming contractor to successfully transition the program. The most successful transitions have incorporated the three following structures:

- **Initial Planning Meeting:** A preliminary meeting to draft a transition specification document that identifies all relevant data, test items, testing information, and other materials necessary for a successful transition. The meeting should include the Department's entire technical staff who have specific, relevant knowledge.
- **Vendor Meeting:** A follow-up meeting with the Department and the new contractor. During this meeting, both project leadership and technical/operations staff from the new contractor will work with AIR and the Department to define information format and delivery schedules and to finalize the transition specification document.
- **Transition Meetings:** Meetings specific to relevant functional areas. Staff from AIR, the Department, and the new contractor will monitor transition timeline and delivery compliance as defined in the transition specification document.

All three types of meeting work best when participants attend in person.

Delivery of Materials and Knowledge

We use a transition specification document to guide the transition, keep track of progress, and plan for contingencies around what may or may not be available. The transition document, organized by functional area, includes

- a description of the item to be transferred (e.g., item content, item statistics, data from previous administrations);
- the format in which the item is available;
- the *reference source*, which is the definitive source against which the transferred item can be verified;
- the quality assurance plan, which describes how the item will be checked against the reference source;
- the planned transition date;
- transition progress; and
- additional information and contents.

These specifications are necessarily quite detailed. For each area of activity, we complete the full transition specification for each detailed item and track it throughout the transition process.

We propose to transfer key information to the Department on an ongoing basis and work with the incoming vendor to provide the data in a format that may be integrated with its systems. If we are the selected contractor, we will commit to support the next vendor's transition.

Section V:
Corporate Qualifications

Section V: Corporate Qualifications

E-1 Required Information on Corporate Qualifications

E-1.1 Vendor and Subcontractors

This proposal is being submitted by the American Institutes for Research.

E-1.1.1 Corporate Overview

Founded in 1946, the American Institutes for Research (AIR) is a not-for-profit organization whose overriding goal is to use the best science available to bring the most effective ideas and approaches to improve lives in communities at home and around the world. We are 1,800+ people working to improve lives, and our mission calls for us to turn the best research into practice, which we do in every arena in which we work, including Early Childhood, Pre-K–12 Education and Social Development, Higher Education and Career Readiness, Adult Learning and the Workforce, Health and Wellness, and International. In AIR Assessment, we have proven able to reach beyond state-of-the-art practices to deliver innovative models, materials, and strategies that advance the field of measurement and improve the validity of student scores.

In total, AIR brings 66 years of experience in aptitude and proficiency testing, including 43 years in K–12 educational assessments. Over the years, AIR staff have worked on 40+ state and national assessments (e.g., NAEP, National Assessment of Adult Literacy, and Voluntary National Tests), on international assessments (e.g., TIMSS and PISA), and on other assessments in Brazil, Namibia, Honduras, Tajikistan, Macedonia, and other countries.

AIR has proven our ability to support 25 states with innovative standard-based assessments that meet or exceed the requirements of Every Student Succeeds Act (ESSA)/Elementary and Secondary Education Act (ESEA), groundbreaking psychometric advances, transitioning to online testing, and enhanced reporting.

Many clients are surprised to discover that AIR is now the undisputed leader in the United States for statewide summative assessment testing. In addition to delivering 43 different statewide alternate assessments, EOC assessments, science assessments, and English language proficiency testing programs, AIR delivers the core grades 3–8 ELA and mathematics testing in 17 states, serving 34% of these students nationwide. No other vendor serves more than 11 states or more than 21% of these students, and no other vendor tests anywhere near as many students online as AIR.

- AIR is the leader in online testing for statewide assessment, with approximately 30 million[1] online tests in each of the 2014–2015 and 2015–2016 school years. Approximately 18.8 million of these were secure, summative assessments, and the vast majority of these (12.3 million) were adaptive. AIR is the only company that has been successfully delivering statewide adaptive assessments at scale for five years in ELA, mathematics, and science. AIR has supported more than one-half dozen states as they moved almost instantly from a paper system to a virtually all online system, including rural states (e.g., Vermont and New Hampshire) and large states (e.g., California).
- AIR has the most accessible system. Excluding the tiny percentage of students requiring paper accommodations, AIR provides the online testing for virtually all students in California, Connecticut, Delaware, Florida, Hawaii, Idaho, Montana, New Hampshire, North Dakota,

[1] Please note that AIR only counts a complete grade/subject (e.g., grade 3 mathematics) as a test. Other companies may inflate their numbers by counting each section of an assessment as a unique test. AIR has delivered approximately 60 million online sections.

Oregon, South Dakota, Utah, Vermont, and West Virginia. This includes not only ELA and mathematics, but often science and social studies assessments as well.

- AIR is the testing system behind other companies:
  - Measured Progress has subcontracted online test delivery to AIR in Montana and North Dakota.
  - ETS has subcontracted all online test delivery to AIR in California. AIR delivers not only ELA and mathematics to 480,000 students in grades 3–11, but also ETS science assessments in grades 5 and 8 and high school; alternate assessments in ELA, mathematics, and science; and the statewide English language proficiency assessments.
  - The College Board has selected AIR to deliver the SAT online, which New Hampshire offers as their high school accountability assessment.
- As described elsewhere in the proposal, AIR's testing system has the smallest footprint in schools; supports the broadest array of item types in the industry; and has Web Content Accessibility Guidelines (WCAG) 2.0 AA certification and offers the most robust selection of accommodations and embedded supports available in the industry.

Success in New Hampshire

AIR has supported New Hampshire since 2014, beginning with the Smarter Balanced field test and continuing through the current operational test administration for New Hampshire students in grades 3–8 and 11 ELA and math. New Hampshire has tested virtually every student online, demonstrating that all of our systems work in New Hampshire schools.

E-1.1.2 Financial Strength

As a not-for-profit corporation, all of our net assets are retained by the organization to support our work and investments. Unrestricted net assets currently exceed $220 million.

AIR is a large federal contractor, audited by the U.S. Department of Education, and is registered to do business in New Hampshire and all 49 other states.

AIR's two most recent audited financial statements are included in Appendix O. As a not-for-profit corporation, our annual tax returns are also publicly available.

E-1.1.3 Litigation

AIR does not have any relevant litigation to disclose. We discuss liquidated damages as required in Topic 28 Vendor Experience.

E-1.1.4 Prior Project Descriptions

As shown below in Topic 28, AIR has more than 20 comparable, current similar projects. In this section, we provide examples of the "no more than three (3) similar projects" required by the RFP.

Arizona

Arizona State Assessment (AzMERIT)
Contract Number: #ADED-083299
Contract Value: $55.8M
Contract Term: 2014–2017
Grades/Subjects: ELA 3–11, math 3–11
Total Online Tests: 1,031,600
Peak Concurrent Testers: 70,305

When Arizona left the PARCC Consortium in 2014, AIR was awarded the contract to develop the new AzMERIT, a system of ELA and math assessments in grades 3–11. Arizona selected a combination of ICCR ELA and math item banks to go along with the development of new Arizona-owned items to create AzMERIT assessments. AIR supported a successful experience transitioning Arizona ELA and mathemathics from paper (AIMS) to the online AzMERIT. Arizona was targeting a slow rollout of online testing, but with the ease of use of AIR's systems, we expect that approximately 80% of Arizona students will take AzMERIT online, representing a much faster online adoption than the ADE expected. Arizona uses AIR's Autoscore to score essay responses in the AzMERIT English language arts assessments.

AIR developed and delivers a comprehensive technical report that documents the development, administration, scoring, and reporting of the AzMERIT assessments. The AzMERIT technical report is reviewed by the ADE and its Technical Advisory Council (TAC) and has been continually revised to support technical reporting goals of the ADE.

Beyond just technical reporting, AIR is committed to providing the ADE with the necessary research studies to support technical documentation of the validity, reliability, and fairness of Arizona's assessments in support of peer review as well as additional research studies suggested by the TAC. AIR is currently working with the ADE to investigate differential effects of glossary accommodations for limited-English-proficient students, building on an initial investigation conducted in spring 2015. We are also working to examine student perceptions of the fairness of reading passages, especially as they may be moderated by membership in important demographic subgroups, including gender and ethnic subgroups, as well as students with individualized education programs, limited-English-proficient students, and students with low socioeconomic status as indicated by their free and reduced-price lunch status.

In the AzMERIT program, as with many other states, AIR has proven our ability to develop and deliver easily understood reports based on accurate and reliable data. AIR and the ADE collaborated closely in the design of the AzMERIT online and paper score reports for English language arts and mathematics, which AIR has successfully delivered to Arizona educators and parents for the past three years.

Arizona Department of Education
1535 West Jefferson Street
Phoenix, AZ 85007
Audra Ahumada, Deputy Associate Superintendent
(602) 542-5423
Audra.Ahumada@azed.gov

Oregon

Contract Numbers: #ODE1007-07, #ODE9573
Contract Value: $31M, $24M
Contract Terms: 2007–2014, 2014–2024
Grades/Subjects: ELA 3–11; math 3–11; science in grades 5, 8, and 11; ELPA through 2015; ELPA21 2015–2024
Total Online Tests: 787,989
Peak Concurrent Testers: 16,325

AIR has supported full-year, multiple-opportunity, online adaptive summative and interim assessments for all students, including English language proficiency assessments in Oregon since 2007. At AIR, we are proud of the productive partnership we have had with Oregon. During that time, we have successfully delivered millions of tests in Oregon's schools without requiring technology upgrades; invented new ways to render tests accessible to all students; and advanced online assessment by the introduction of new item types and task models. Even with these new technologies, Oregon teachers use AIR's item banking system to develop items.

AIR's Psychometrics and Statistics team worked to help Oregon develop technical reports and design and implement studies that were submitted for the state's ultimately successful NCLB peer review. Oregon was the first state to receive peer review on the multiple-opportunity, adaptive online assessments, paving the way for the widespread use of adaptive testing in statewide assessment today.

AIR and Oregon found ways to deliver tests, including tests that deliver and capture audio, with very limited bandwidth demands and to do so reliably within real school networks. Together with Oregon, we devised a way to deliver adaptive tests to students who use Braille. While the rest of the industry has been focusing on refreshable Braille displays (RBDs), our research indicated limitations in these devices—they cannot display Nemeth Braille, nor can they display tactile graphics— that were critical for a fair test. Working together, we devised a system that incorporates real-time embossing with RBDs to fairly measure the content across subject areas for students who use Braille. Similar innovation led to a text-to-speech engine optimized for delivering tests, including the annotations typically used for read-alouds. Working with Oregon and Smarter Balanced, we continue to add embedded supports to meet the needs of English language learners, students with disabilities, and all students who need support.

Responding to Oregon's vision, AIR developed its first graphic-response items. These items were designed to support machine-scored constructed responses in order to reach deeper levels of knowledge and understanding. Similarly, we developed natural language processing engines that can recognize propositions in real student text.

Oregon Department of Education
255 Capitol Street NE
Salem, OR 97338
Kathleen Vanderwall, Manager, Test Design & Implementation
Oregon Department of Education
(503) 947-5600
kathleen.vanderwall@state.or.us

Hawaii

Contract Numbers: #CO-60111, #CO-10361, #CO-10328, #C010551
Contract Value: $75M, $5M, $4.9M, $14.6M
Contract Terms: 2007–2014, 2014–2024
Grades/Subjects: ELA 3–11; math 3–11; science in grades 5, 8, and 11; EOCs in American History, American Government, and Algebra I and II; ELPA through 2015; ELPA21 2015–2024
Total Online Tests: 288,630
Peak Concurrent Testers: 15,593

AIR has supported Hawaii since 2006 on a set of projects covering the original paper-based Hawaii State Assessments (HSA), which transitioned to online, adaptive in 2009–2010, end-of-course assessments, alternate assessments, and the transition to Smarter Balanced in 2014. AIR delivered the paper-based HSA from 2006 through 2010, before the HSA was converted online for the 2010–2011 school year. Since then and through the current Smarter Balanced testing, more than 99% of Hawaii students have tested online and fully-adaptive Braille testing has been available. AIR also developed and delivers online high school end-of-course (EOC) assessments (American History, American Government, Algebra I and II). Hawaii was the first to move 100% online and adaptive to a tablet-based alternate assessment for students with disabilities, and we also administered the assessment for students in the Hawaiian Language Immersion Program at grades 3 and 4 for several years.

AIR met the schedule requirements for this project, and actually reduced our fixed prices over time as more students tested online. AIR's Heather Hayes oversees this project and will do so for New Hampshire.

Hawaiʻi Department of Education
1390 Miller Street
Honolulu, HI 96813
Tom Saka, Director
(808) 586-3288
tom_saka@notes.k12.hi.us

E-1.1.1.4.1 Components that constitute the vendor's proposed software suite must be fully implemented and operational in at least one (1) government entity comparable in size and complexity to the State of New Hampshire.

AIR is proposing to use the same student registration, test delivery, online reporting, and related software systems that we deployed successfully in New Hampshire for the past three years, and in 24 other states and jurisdictions over the past decade. A summary is provided in Exhibit E-1.1.1.4.1-1 and described in Topic 4 Assessment Delivery Platform.

Exhibit E-1.1.1.4.1-1: Summary of Online Test Delivery Components

System	Description
Test Information Distribution Engine (TIDE)	TIDE is responsible for • registering students; • gathering of demographic data; and • ordering materials.
Test Delivery System (TDS)	The test delivery system's Test Administrator Interface provides the interface through which test administrators establish and monitor testing sessions and authenticate student users. The Student Interface is the testing system on which students take tests. The TDS delivers tests to students, records responses, and forwards data to downstream systems.
Quality Monitor (QM) System	The QM system receives the data, verifies the validity of the test administered and the scores assigned, gathers statistical data for ongoing quality reports, and forwards data to the Database of Record (DoR) and the Online Reporting System (ORS).
Database of Record (DoR)	The DoR maintains the authoritative records of tests administered and completed. Data in the DoR reflect appeals, verifications, and other post-administration adjustments to the data.
Online Reporting System (ORS)	The ORS provides a secure interface to assessment data and associated demographic information. It provides educators with a powerful tool to explore the data and turn them into actionable information.
Online Administrative Portal	The online administrative portal provides a customized, centralized location for educators and other stakeholders and provides access to the other systems as well as practice sites, training sites, tutorials, and other resources.

E-1.1.5 Subcontractor Information

AIR is not proposing any subcontractors for this effort.

AIR regularly enters into licensing agreements and engages many vendors for various services in support of assessment projects. Licensing agreements typically include copyright (e.g., publishers for permissioned passages, Copyright Clearance Center) and content agreements (e.g., licensing test content). Vendors include managed hosting services (Rackspace), printers for test materials and printed score reports (e.g., AGS), communications (e.g., multiple vendors for VoIP, WebEx, cell phones), shipping (e.g., USPS, Federal Express), internal copying (e.g., Xerox, Océ), temporary employment agencies, supplies, and many other services required to support our staff and project work. We procure these services after careful analysis and/or competition, and develop formal contractual agreements.

Section VI:
Qualifications of Key Vendor Staff

Section VI: Qualifications of Key Vendor Staff

E-2 Team Organization and Designation of Key Vendor Staff

We refer readers to the resumes of key personnel proposed to work on this project included in Appendix A. Please see Exhibit E-2-1 for a project organizational chart.

E-2.1 State Staff Resource Worksheet

A completed State Staff Resource Worksheet is included in Appendix S.

E-3 Candidates for Project Manager

We propose Tom Glorfield, B.F.A., as the project manager for New Hampshire's statewide assessments, and we present his resume in Appendix A. Exhibit E-3.1 lists reference information.

Exhibit E-2-1: Project Organizational Chart

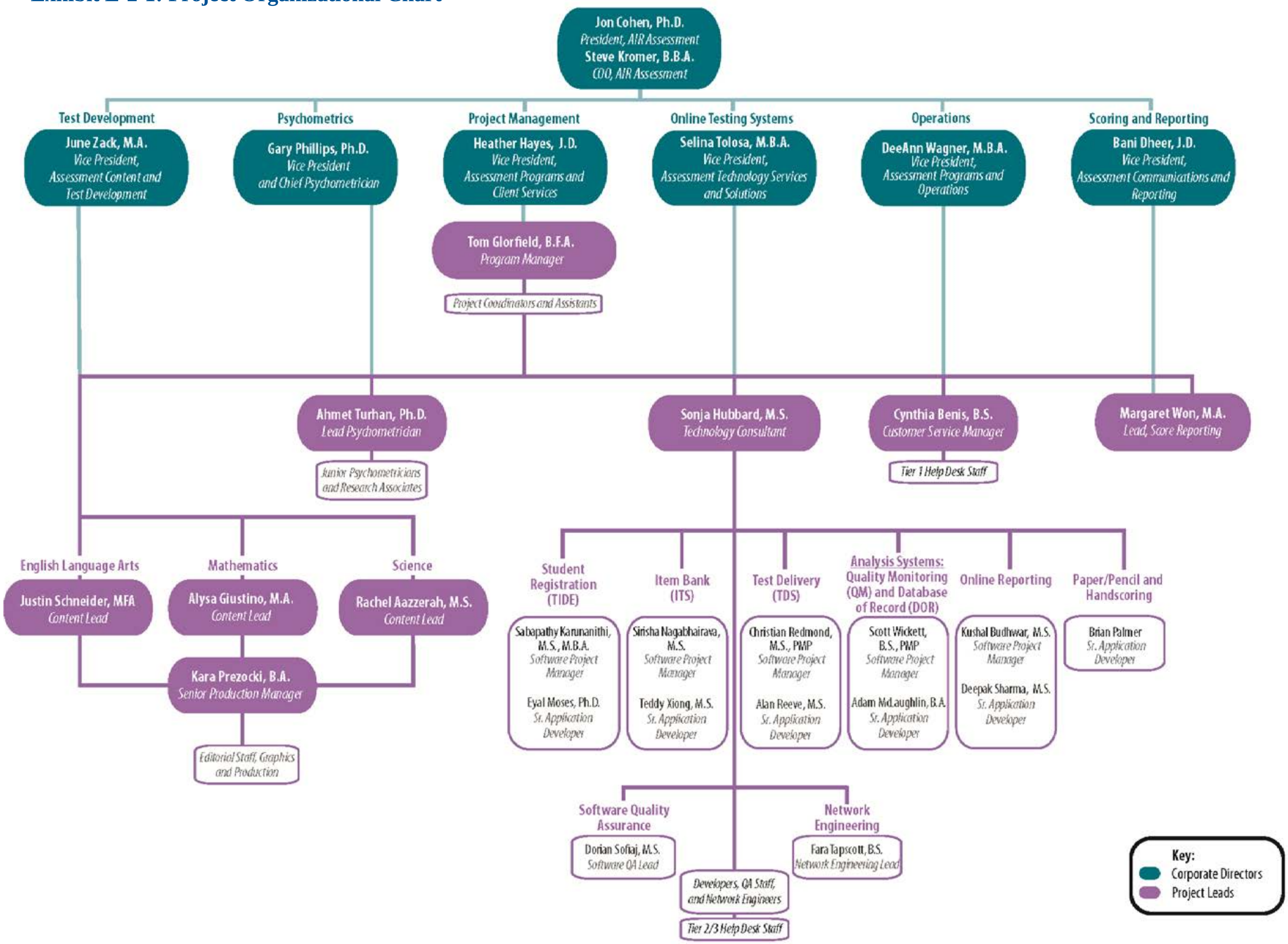


Exhibit E-3.1 References for Project Manager

AIR Staff	References
Program Manager Tom Glorfield, B.F.A.	Various state department of education staff under project officer: Brian Touchette, Director Accountability Resources Delaware Department of Education 401 Federal Street, Suite 2 Dover, DE 19901-3639 (302) 735-4090 brian.touchette@doe.k12.de.us Various state department of education staff under project officer: Michael Hock, Director of Educational Assessment Vermont Agency of Education 219 North Main Street, Suite 402 Barre, VT 05641 (802) 479-1288 michael.hock@state.vt.us Various state department of education staff under project officer: Scott J. Mantie, Ph.D., Administrator, Bureau of Accountability and Assessment New Hampshire Department of Education 101 Pleasant Street Concord, NH 03301 (603) 271-3844 scott.mantie@doe.nh.gov

E-4 Candidates for Key Vendor Staff Roles

We present resumes of AIR staff in Appendix A. Resumes include personnel's educational background, work history, project experience, and significant certifications. Exhibit E-4.1 lists reference information for our key staff.

Exhibit E-4.1: References for Key Staff Roles

AIR Staff	References
Lead Psychometrician Ahmet Turhan, Ph.D.	Various state departments of education staff under project officer: Vince Verges, Assistant Deputy Commissioner Florida Department of Education Office of Assessments 325 West Gaines Street, Turlington Building, Suite 404 Tallahassee, FL 32399 (850) 245-0513 vince.verges@fldoe.org Leslie Keng, Ph.D. Principal Researcher at Pearson (512) 534-0798 lesliekeng@gmail.com Vaughn Rhudy West Virginia Department of Education Office of Assessment 1900 Kanawha Boulevard East, Building 6, Room 825 Charleston, WV 25305

	(304) 558-2546 vrhudy@k12.wv.us

Exhibit E-4.1: References for Key Staff Roles (continued)

AIR Staff	References
Technology Consultant Sonja Hubbard, M.S.	Various state department of education staff under project officer: Jo Ellen Shaeffer, Assessment Director Utah Department of Education 250 East 500 South P.O. Box 144200 Salt Lake City, UT 84114-4200 (801) 538-7811 JoEllen.Shaeffer@schools.utah.gov Various state department of education staff under project officer: Kathleen Vanderwall, Manager, Test Design & Implementation Oregon Department of Education 255 Capitol Street NE Salem, OR 97310 (503) 947-5600 kathleen.vanderwall@state.or.us Various state department of education staff under project officer: Brian Touchette, Acting Director Accountability Resources Delaware Department of Education 401 Federal Street, Suite 2 Dover, DE 19901-3639 (302) 735-4090 brian.touchette@doe.k12.de.us
Scoring & Reporting Lead Margaret Won, M.A.	Various state department of education staff under project officer: Scott J. Mantie, Ph.D., Administrator, Bureau of Accountability and Assessment New Hampshire Department of Education 101 Pleasant Street Concord, NH 03301 (603) 271-3844 scott.mantie@doe.nh.gov Various state department of education staff under project officer: Michael Hock, Director of Educational Assessment Vermont Agency of Education 219 North Main Street, Suite 402 Barre, VT 05641 (802) 479-1288 michael.hock@state.vt.us Various state department of education staff under project officer: Abe Krisst Connecticut State Department of Education 165 Capitol Avenue Hartford, CT 06106

Exhibit E-4.1: References for Key Staff Roles (continued)

AIR Staff	References
	(860) 713-6860 abe.krisst@ct.gov
Science Content Lead Rachel Aazzerah, M.S.	Stephen Pruitt, Ph.D., Commissioner of Education Kentucky Department of Education NGSS-Achieve (Formerly) (202) 590-8944 Brett Moulding NGSS Writing Team Leader (801) 520-9744 mouldingb@ogdensd.org Mary Anderson, Ph.D., Director of Assessment Oregon Department of Education (503) 559-4157 mary.anderson@state.or.us
English Language Arts Content Lead Justin Schneider, M.F.A.	Various state department of education staff under project officer: Jo Ellen Shaeffer, Director Utah Department of Education 250 East 500 South P.O. Box 144200 Salt Lake City, UT 84114-4200 (801) 538-7811, JoEllen.Shaeffer@schools.utah.gov Various state department of education staff under project officer: Vince Verges, Assistant Deputy Commissioner Florida Department of Education Office of Assessments 325 West Gaines Street, Turlington Building, Suite 404 Tallahassee, FL 32399 (850) 245-0513 vince.verges@fldoe.org
Mathematics Content Lead Alysa Giustino, M.A.	Various state department of education staff under project officer: Vince Verges, Assistant Deputy Commissioner Florida Department of Education Office of Assessments 325 West Gaines Street, Turlington Building, Suite 404 Tallahassee, FL 32399 (850) 245-0513 vince.verges@fldoe.org

Section VII:
Pricing Model

AIR has included a comprehensive response to the requirements Section VII: Pricing Model (in response to the Appendix F worksheet included as an attachment to the RFP) in addition to the requirements of Topic 42 Pricing Model under a separate Cost Proposal cover.

Section VIII: Copy of the RFP

and any signed Addendum (a) – required in original proposal

Copy of the RFP

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

New Hampshire Department of Information Technology
RFP Cover Sheet

Name of Agency/Division: Department of Education
Contract Number/Name: RFP 2017-073 DOE New Hampshire Statewide Assessments
RFP Purpose: To contract with a qualified vendor for a statewide, computer-based assessment system based upon New Hampshire's academic standards in English language arts (ELA) and mathematics for grades 3-8, and science for grades 5, 8 & 11. The current contracts are expiring June 30th, 2017 and the State is Federally required to have an executed contract and fully tested ready for implementation State Assessment for the 2017-2018 academic school year.
DOIT ☐ or Agency X Funding

Background Information:

A. Through this request for proposals (RFP), the New Hampshire Department of Education (NH DOE) is seeking proposals from vendors experienced in statewide student assessments and assessment systems with the capacity to deliver machine-scored online summative and interim assessments. Such assessment system should include:

- Reporting capabilities to allow easy and timely access to results at various reporting levels from statewide down to the individual student/parent;
- ELA (to also include a machine-scored/artificial intelligence (AI) scored writing component) and mathematics to be administered annually in grades 3-8, and science to be administered annually in grades 5, 8, and 11;
- Interim assessment results that may be aggregated and used in lieu of summative assessments, if permitted under ESSA; and
- A paper-based assessment option for special circumstances.

B. The content of the ELA, mathematics and science assessments must align with:

- Appropriate representations of the New Hampshire's grade level academic standards for ELA, mathematics and science. (NH RSA193-C:3 III); and
- Nationally and/or internationally recognized academic standard(s) with demonstrable longitudinal data supporting both student growth and student achievement models leading to college and career readiness.

C. The assessments must meet high standards of technical quality, the requirements set forth in New Hampshire state law (RSA 193:C) and the Every Student Succeeds Act (ESSA) and corresponding United States Department of Education (US ED) regulations. Further, the assessments must meet the requirements of the US ED Peer Review Guidelines.

Special Concerns:

This RFP contains five (5) components:
Vendors may bid on all five (5) components (components A, B, C, D, E), on any combination of assessment components (A, B, C, D), or any individual assessment component (A, B, C, D). For each component selected by the vendor, the proposal must address component E, and must address each component selected completely. The NH DOE will not accept bids for pieces of individual components.

COMPONENT A: SUMMATIVE ASSESSMENTS IN ELA AND MATHEMATICS. The ELA (to include machine-scored writing component) and mathematics assessments will be administered annually in grades 3-8.

COMPONENT B: SUMMATIVE ASSESSMENT IN SCIENCE. Science assessments will be administered annually in grades 5, 8, and 11.

COMPONENT C: INTERIM ASSESSMENTS IN ELA AND MATHEMATICS. These online assessments are to be made available for local district use ***at least*** in the same grades included in the summative assessments described above. The purpose of the interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.

COMPONENT D: INTERIM ASSESSMENTS IN SCIENCE. These online assessments are to be made available for local district use ***at least*** in the same grades included in the summative assessment described above. The purpose of the interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.

COMPONENT E: REPORTING PORTAL. The reporting portal must be capable of being upgraded across time. Initially, the system must be able to organize and present assessment data in a way that is easy for all users to read and understand. The Reporting Portal should allow for differentiated access to individual student, classroom, school, district and state assessment data, and they should integrate seamlessly with each other. Assessment results must be able to be provided to parents, among others, in a timely fashion, either online or by printed report.

Machine Scored Online Assessment
The assessments will be administered and scored online where applicable and fiscally practicable.

Standards Revision and Alignment
The NH academic standards for ELA, mathematics and science are entering a revision cycle. It is therefore imperative that vendors acknowledge and include in proposals a plan for assessment revision aligned to updated standards in 2018-2019.

Comparability and Validity Studies
The NH DOE is interested in comparability, validity and research-based alignment studies completed by the vendor and/or independently for proposed RFP solutions. This should include demonstrable longitudinal support models for student growth and achievement.

Selection
The NH DOE will only select a vendor/vendors that has/have assured the state in its proposal that it will follow the *Standards for Educational and Psychological Testing* by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education. The NH DOE will select the vendor(s) that provide(s) the most technically sound and cost effective proposal. Assuming comparable technical quality and cost savings, preference will be given to vendor(s) who are fully capable of delivering the summative and interim assessments in combination.

Partnership
The vendors may partner with another firm to provide parts of the solution; however, the vendor must provide management of the partner and is responsible for all project performance. The vendor is responsible for ensuring that each subcontractor acknowledges and is contractually bound by the staffing plan and other commitments listed in this RFP.

Submitted By: Sandie MacDonald	**Current Date:** March 28, 2017
Phone: 603.271.3453	**Email:** Saundra.MacDonald@doe.nh.gov

Remainder of this page intentionally left blank

Department of Education

NH Statewide Assessments: English Language Arts, Mathematics, Science

RFP

RFP ISSUED……………………………………………………………………………..March 29, 2017

VENDOR CONFERENCE……………………………………………….April 6, 2017 at 1:00 PM
Department of Education
Room 15

STATE POINT of CONTACT…………………………………….Sandie MacDonald, Administrator
Saundra.MacDonald@doe.nh.gov
603.271.3453

CONTRACT TYPE…………………………………………………………………..Not to Exceed $3M

PROPOSALS DUE…………………………………………………………………………April 26, 2017

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

New Hampshire Department of Education
Division of Educational Improvement
101 Pleasant Street
Concord, New Hampshire 03301

March 28, 2017

REQUEST FOR PROPOSAL (RFP)

RFP 2017-073 DOE New Hampshire Statewide Assessments

Proposals Deadline: 4:00 PM, Wednesday, April 26, 2017

Proposal Inquiries: All questions about this Request for Proposal (RFP) should be submitted by email to:

Sandie MacDonald, Administrator
Bureau of Instructional Support and Student Assessment
e-mail: saundra.macdonald@doe.nh.gov

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

1. INTRODUCTION

The State of New Hampshire (State), acting through the Department of Education (NH DOE), is releasing this Request for Proposal (RFP) for NH DOE to contract with a vendor to procure an assessment, assessment system and associated support services for a computer-based summative (with paper-based option) and interim assessment system that is substantially aligned with New Hampshire academic standards in English Language Arts (ELA) and mathematics in grades 3-8 and science in grades 5, 8 and 11.

Project Overview

Through this request for proposals (RFP), the New Hampshire Department of Education (NH DOE) is seeking proposals from vendors experienced in statewide student assessments and assessment systems with the capacity to deliver machine-scored online summative, formative and interim assessments. Such assessment system should include:

- Reporting capabilities to allow easy and timely access to results at various reporting levels from statewide down to the individual student/parent;
- ELA (to also include a machine-scored writing component) and mathematics to be administered annually in grades 3-8, and in science to be administered annually in grades 5, 8, and 11;
- Interim assessment results may be aggregated and used in lieu of summative assessments, if permitted under ESSA; and
- A paper-based assessment option for special circumstances.

The assessment and assessment system must be developmentally appropriate for students and follow the *Standards for Educational and Psychological Testing* by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education. The system must also align to the requirements set forth in New Hampshire state law (RSA 193:C) and the Every Student Succeeds Act (ESSA) and corresponding United States Department of Education (US ED) regulations. Further, the assessments must meet the requirements of the US ED Peer Review Guidelines.

Contract Overview

It is expected that the contract period for the assessments described in this RFP will begin no later than July 2017, or upon NH Governor and Executive Council approval, and the contract period will continue through June 2021.

The continuation and annual renewal of all contracts are subject to continued availability of funds, State and Federal assessment requirement changes and successful implementation by the vendor.

1.1 Contract Award

The State plans to execute a *Not to Exceed* (NTE) contract as a result of this RFP. If an award is made, it shall be made based upon evaluation of the submitted proposals in accordance with the review process outlined in

Section 5 below. The award will be based upon criteria, standards and weighting identified in this RFP.

1.1.1 Non-Exclusive Contract

Any resulting contract from this RFP will be a non-exclusive contract. The State reserves the right, at its discretion, to retain other vendors to provide any of the services or deliverables identified under this procurement or make an award by item, part or portion of an item, group of items, or total proposal.

If a contract is awarded, the vendor must obtain written consent from the State before any public announcement or news release is issued pertaining to any contract award. Such permission, at a minimum, will be dependent upon approval of the contract by Governor and Executive Council of the State of New Hampshire. A contract award is contingent on approval by the Governor and Executive Council.

1.2 Contract Term

Time is of the essence in the performance of a vendor's obligations under the contract. The vendor shall be fully prepared to commence work by July 1, 2017 after full execution of the contract by the parties, and the receipt of required governmental approvals, including, but not limited to, the NH Governor and Executive Council approval ("effective date").

The vendor's initial term will begin on the effective date and extend through June 30, 2021. The term may be extended up to 3 years ("extended term") at the sole option of the State, subject to the parties' prior written agreement on applicable fees for the extended term, up to but not beyond June 30, 2024.

The vendor shall not commence work until the vendor is in receipt of the *Notice to Proceed* by the NH DOE.

1.2.1 Contract Negotiations and Vendor Notice

If a vendor is selected, the NH DOE will notify the selected vendor in writing of their selection and the State's desire to enter into contract discussions. Until the State completes discussions with the selected vendor, all submitted proposals remain eligible for selection by the State. In the event contract discussions cannot be completed with the selected vendor, the evaluation team may recommend another vendor.

In accordance with New Hampshire Statutes Chapter 21-I:13-a, no information shall be available to the public, the members of the general court or its staff, notwithstanding the provisions of RSA 91-A:4, concerning specific responses to this RFP, from the time the RFP is made public until the contract is actually awarded, in order to protect the integrity of the public procurement process. This means vendors shall not be notified until after the Governor and Executive Council have approved the resulting contract. No information can be provided to non-selected vendors until after contracts are awarded, at which time non-selected applicants may submit a written request for more information about the reasons for not being selected and recommendations that may make future applications more effective. Such requests are not considered appeals. Once an applicant has submitted a letter, the State will attempt to accommodate such requests within a reasonable time.

1.3 Subcontractors

The vendor shall identify all subcontractors to deliver required services subject to the terms and conditions of this RFP, including but not limited to, in Appendix H Section H-25: *General Contract Requirements* herein and Appendix H: *State of New Hampshire Terms and Conditions* of this RFP.

The vendor shall remain wholly responsible for performance of the entire contract regardless of whether a subcontractor is used. The State will consider the vendor to be the sole point of contact with regard to all contractual matters, including payment of any and all charges resulting from any contract.

2. SCHEDULE OF EVENTS

The following table provides the Schedule of Events for this RFP through Governor and Executive Council approval and Notice to Proceed.

EVENT	DATE	TIME
RFP released to vendors (on or about)	March 29, 2017	9:00 AM
Vendor inquiry period begins (on or about)	March 29, 2017	12:00 PM
Notification to the State of the number of representatives attending the Vendor Conference	April 3, 2017	4:00 PM
(Optional) Vendor Conference; location identified in *General Instructions*, Section 4.3	April 6, 2017	1:00 PM
Vendor inquiry period ends (final inquiries due)	April 12, 2017	4:00 PM
Final State responses to vendor inquiries	April 17, 2017	4:00 PM
Final date for proposal submission	**April 26, 2017**	**4:00 PM**
Vendor presentations/discussion sessions/interviews, if necessary	On or about May 8, 2017 through May 12, 2017	TBD
Contract negotiations and development	May–June 2017	N/A
Anticipated Governor and Executive Council approval	July 2017	TBD
Anticipated Notice to Proceed	July 2017	TBD

3. SOFTWARE, REQUIREMENTS AND DELIVERABLES

3.1 Software

Each proposal must present software that can fully support the required functionality listed in Appendix C: *System Requirements and Deliverables.*

3.2 Requirements

3.2.1 Appendix B: *Minimum Standards for Proposal Consideration*, compliance with system requirements, use of any proposed commercial off-the-shelf (COTS) software, vendor implementation experience, and proposed project team.

3.2.2 Appendix C: *System Requirements and Deliverables*

3.2.3 Appendix D: *Topics for Mandatory Narrative Responses* for software, technical services and project management topics.

3.2.4 Appendix E: *Standards for Describing Vendor Qualifications* including vendor corporate qualifications, team organization and key staff, program manager, and other key staff candidates' qualifications.

3.3 Deliverables

The State classifies deliverables into three (3) categories: *Written Deliverables, Software Deliverables,* and *Non-Software Deliverables.* Pricing and scheduling information requirements for these deliverables are provided in Appendix F: *Pricing Worksheets*. A set of required deliverables as well as a list of requirements for these deliverables is detailed in Appendix C: *System Requirements and Deliverables.* Appendix D: *Topics for Mandatory Narrative Responses* solicits responses, which will expound on the vendors' understanding of the implementation process, the manner of service delivery and experience with similar projects related to the software, technical services, and project management topics.

4. INSTRUCTIONS

4.1 Proposal Submission, Deadline, and Location Instructions

Proposals submitted in response to this RFP must be received by the Department of Education, no later than the time and date specified in Section 2: *Schedule of Events*. Proposals must be addressed to:

State of New Hampshire
Department of Education
c/o Sandie MacDonald
101 Pleasant Street
Concord, New Hampshire 03301

Cartons containing proposals must be clearly marked as follows:

RESPONSE TO PROPOSAL: NH DOE Statewide Assessments in English Language Arts, Mathematics and Science

Late submissions will not be accepted and will remain unopened. Delivery of the proposals shall be at the vendor's expense. The time of receipt shall be considered when a proposal has been officially documented by the NH DOE, in accordance with its established policies, as having been received at the location designated above. The State accepts no responsibility for mislabeled mail. Any damage that may occur due to shipping shall be the vendor's responsibility.

Vendors are permitted to submit **only one (1)** proposal in response to this RFP.

The State reserves the right to reject any and all proposals and to waive informalities and minor irregularities in proposals received. The State also reserves the right to accept any portion of a proposal or all items bid if deemed in the best interest of the State to do so.

All proposals submitted in response to this RFP must consist of:

a. One (1) original and five (5) clearly identified copies of the proposal, including all required attachments,
b. One (1) copy of the *Proposal Transmittal Form Letter* (described in Section 4.18.2: *Transmittal Form Letter*, herein) shall be signed by an official authorized to legally bind the vendor and shall be marked "ORIGINAL."
c. One (1) electronic copy on a USB Flash Drive in MS WORD format.

The original and all copies shall be bound separately, delivered in sealed containers, and permanently marked as indicated above. A vendor's disclosure or distribution of its proposal other than to the State will be grounds for disqualification.

4.2 Proposal Inquiries

All inquiries concerning this RFP, including but not limited to, requests for clarifications, questions and any changes to the RFP, shall be emailed, citing the RFP title, RFP number, page, section, and paragraph and submitted to the following RFP State Point of Contact:

Sandie MacDonald
Email: saundra.macdonald@doe.nh.gov

Vendors are encouraged to submit questions via email; however, the State assumes no liability for assuring accurate/complete email transmission/receipt and is not responsible to acknowledge receipt. All inquiries should have "NH DOE Statewide Assessments: ELA, Mathematics, Science" in the subject line.

Inquiries must be received by the RFP State Point of Contact (see above) no later than the conclusion of the *Vendor Inquiry Period* (see Section 2: *Schedule of Events*). Inquiries received later than the conclusion of the Vendor Inquiry Period shall not be considered properly submitted and may not be considered.

The State intends to issue official responses to properly submitted inquiries on or before the date specified in Section 2: *Schedule of Events*; however, this date may be subject to change at the State's discretion. The State may consolidate and/or paraphrase questions for sufficiency and clarity. The State may, at its discretion, amend this RFP on its own initiative or in response to issues raised by inquiries, as it deems appropriate. Oral statements, representations, clarifications, or modifications concerning the RFP shall not be binding upon the State. Official responses will be made in writing.

4.2.1 Restriction of Contact With State Employees

From the date of release of this RFP until an award is made and announced regarding the selection of a vendor, all communication with personnel employed by or under contract with the State regarding this RFP is forbidden unless first approved by the RFP State Point of Contact listed in Section 4.2: *Proposal Inquiries*. State employees have been directed not to hold conferences and/or discussions concerning this RFP with any vendor during the selection process, unless otherwise authorized by the RFP State Point of Contact.

4.3 Vendor Conference

An **optional** Vendor Conference will be held at the following location on the date and at the time identified in Section 2: *Schedule of Events:*

Department of Education
Room 15
101 Pleasant Street
Concord, New Hampshire 03301

All vendors who intend to submit proposals are encouraged to attend the Vendor Conference. Attendance by teleconference is permitted. Phone numbers will be emailed to registrants upon request. Vendors are requested to

RSVP via email by the date identified in Section 2: *Schedule of Events*, indicating the number of individuals who will attend the Vendor Conference. Vendors are allowed to send a maximum number of 4 representatives.

Vendors will have an opportunity to ask questions about the RFP and the State will make a reasonable attempt to answer questions it deems appropriate. Questions may include, without limitation, a request for clarification of the RFP; a request for changes to the RFP; suggestions or changes to the RFP that could improve the RFP competition or lower the offered price; and to review any applicable documentation.

Vendors are encouraged to email inquiries at least twenty-four (24) hours prior to the Vendor Conference. No responses will be given prior to the Vendor Conference. Oral answers will not be binding on the State. The State's final response to vendor inquiries and any requested changes to terms and conditions raised during the vendor inquiry period will be posted to the website by the date specified as the final State responses to vendor inquiries as specified in Section 2: *Schedule of Events*. Vendors are responsible for any costs associated with attending the Vendor Conference.

4.4 Alteration of RFP

The original RFP document is on file with the State of New Hampshire, Department of Administrative Services. Vendors are provided an electronic version of the RFP. Any alteration to this RFP or any file associated with this RFP is prohibited. Any such changes may result in a proposal being rejected.

4.5 RFP Addendum

The State reserves the right to amend this RFP at its discretion, prior to the proposal submission deadline. In the event of an addendum to this RFP, the State, at its sole discretion, may extend the proposal submission deadline, as it deems appropriate.

4.6 Non-Collusion

The vendor's signature on a proposal submitted in response to this RFP guarantees that the prices, terms and conditions, and services quoted have been established without collusion with other vendors and without effort to preclude the State from obtaining the best possible competitive proposal.

4.7 Validity of Proposal

Proposals must be valid for one hundred and eighty (180) days following the deadline for submission of proposals in Section 2: *Schedule of Events*, or until the effective date of any resulting contract.

4.8 Property of the State

All material received in response to this RFP shall become the property of the State and will not be returned to the vendor. Upon contract award, the State reserves the right to use any information presented in any proposal.

4.9 Confidentiality of a Proposal

A proposal must remain confidential until the effective date of any resulting contract as a result of this RFP. A vendor's disclosure or distribution of proposals other than to the State will be grounds for disqualification.

4.10 Public Disclosure

Subject to applicable law or regulations, the content of each vendor's proposal shall become public information upon the effective date of any resulting contract.

Any information submitted as part of a response to this request for proposal (RFP) may be subject to public disclosure under RSA 91-A. In addition, in accordance with RSA 9-F:1, any contract entered into as a result of this RFP will be made accessible to the public online via the website *Transparent NH* (http://www.nh.gov/transparentnh/). Accordingly, business financial information and proprietary information such as trade secrets, business and financials models and forecasts, and proprietary formulas may be exempt from public disclosure under RSA 91-A:5, IV. If you believe any information being submitted in response to a request for proposal, bid or information should be kept confidential as financial or proprietary information, you must specifically identify that information in a letter to the agency, and should mark/stamp the materials as such. Marking of the entire proposal or entire sections of the proposal (e.g. pricing) as confidential will neither be accepted nor honored. Notwithstanding any provision of this RFP to the contrary, vendor pricing will be subject to disclosure upon approval of the contract by the NH Governor and Executive Council.

Generally, each proposal shall become public information upon the approval of the NH Governor and Executive Council of the resulting contract, as determined by the State, including but not limited to, RSA Chapter 91-A (Right to Know law). The State will endeavor to maintain the confidentiality of portions of the proposal that are clearly and properly marked confidential. If a request is made to the State to view portions of a proposal that the vendor has properly and clearly marked confidential, the State will notify the vendor of the request and of the date and the State plans to release the records. A designation by the vendor of information it believes exempt does not have the effect of making such information exempt. The State will determine the information it believes is properly exempted from disclosure. By submitting a proposal, vendors agree that unless the vendor obtains a court order, at its sole expense, enjoining the release of the requested information, the State may release the requested information on the date specified in the State's notice without any liability to the vendors.

4.11 Security

The State must ensure that appropriate levels of security are implemented and maintained in order to protect the integrity and reliability of its information technology resources, information and services. State resources, information and services must be available on an ongoing basis, with the appropriate infrastructure and security controls to ensure business continuity and safeguard State networks, systems and data.

The State will evaluate the degree to which the proposed system is designed and architected to ensure the confidentiality and integrity of its valued asset, data.

4.12 Non-Commitment

Notwithstanding any other provision of this RFP, this RFP does not commit the State to award a contract. The State reserves the right, at its sole discretion, to reject any and all proposals, or any portions thereof, at any time; to cancel this RFP; and to solicit new proposals under a new acquisition process; or cancel this entire RFP or individual components at any time, without penalty.

4.13 Proposal Preparation Cost

By submitting a proposal, a vendor agrees that in no event shall the State be either responsible for or held liable for any costs incurred by a vendor in the preparation of or in connection with the proposal, or for work performed prior to the effective date of a resulting contract.

4.14 Oral Presentations/Interviews and Discussion

The State reserves the right to require vendors to make oral presentations of their proposals and/or to make available for oral presentations/interviews the IT consultants proposed to implement the COTS application. All costs associated with oral presentations/interviews shall be borne entirely by the vendor. Vendors may be requested to provide demonstrations of their proposed systems as part of their presentations.

4.15 Required Contract Terms and Conditions

By submitting a proposal, the vendor agrees that the State of New Hampshire terms and conditions, contained in Appendix H: *State of New Hampshire Terms and Conditions* herein, shall form the basis of any contract resulting from this RFP. In the event of any conflict between the State's terms and conditions and any portion of the vendor's proposal, the State's terms and conditions shall take precedence and supersede any and all such conflicting terms and conditions contained in the vendor's proposal.

4.16 Proposal Format

Proposals should follow the following format:

- The proposal should be provided in a three-ring binder.
- The proposal should be printed on white paper with dimensions of 8.5 by 11 inches with right and left margins of one (1) inch.
- The proposal should use Times New Roman font with a size no smaller than eleven (11).
- Each page of the proposal should include a page number and the number of total pages and identification of the vendor in the page footer.
- Tabs should separate each section of the proposal.

Exceptions for paper and font sizes are permissible for: graphical exhibits, which may be printed on white paper with dimensions of 11 by 17 inches; and material in appendices.

4.17 Proposal Organization

Proposals should adhere to the following outline and should not include items not identified in the outline.

- **Cover Page**
- **Transmittal Form Letter**
- **Table of Contents**
- **Section I:** Executive Summary
- **Section II:** Glossary of Terms and Abbreviations
- **Section III:** Responses to Requirements and Deliverables
- **Section IV:** Narrative Responses
- **Section V:** Corporate Qualifications
- **Section VI:** Qualifications of Key Vendor Staff
- **Section VII:** Pricing Model
- **Section VIII:** Copy of the RFP and any signed Addendum (a) - ***required in original proposal only***
- **Section IX:** Appendix

4.18 Proposal Content

4.18.1 Cover Page

The first page of the vendor's proposal should be a cover page containing the following text:

STATE OF NEW HAMPSHIRE
Department of Education
RESPONSE TO PROPOSAL: RFP 2017-073 DOE New Hampshire Statewide Assessments

The cover page should also include the vendor's name, contact person, contact telephone number, address, city, state, zip code, fax number, and email address.

4.18.2 Transmittal Form Letter

The vendor must submit signed Transmittal Form Letter with their response using the Transmittal Form Letter Template provided herewith. Any electronic alteration to this Transmittal Form Letter is prohibited. Any such changes may result in a proposal being rejected.

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

State of New Hampshire Proposal Transmittal Form Letter

Company Name____
Address____

To: NH Department of Education
State Point of Contact: Sandie MacDonald
Telephone: (603) 271-3453
Email: saundra.macdonald@doe.nh.gov

RE: Proposal Invitation Name: **RFP 2017-073 DOE New Hampshire Statewide Assessments**
Proposal Due Date and Time: Wednesday, April 26, 2017 at 4:00 PM

Dear Project Manager:

Company Name: ______________________ hereby offers to sell to the State of New Hampshire the services indicated in **RFP 2017-073 DOE New Hampshire Statewide Assessments** at the price(s) quoted in vendor Response Section VII: *Pricing Model*, and Appendix F: *Pricing Worksheets*, in complete accordance with all conditions of this RFP and all specifications set forth in the RFP and in the State of New Hampshire Terms and Conditions outlined in RFP Appendix H: *State of New Hampshire Terms and Conditions*.

[Company Signor:] __________________________________ is authorized to legally obligate [Company Name:] __________________________________.

We attest to the fact that:

The company has reviewed and agreed to be bound by all RFP terms and conditions including but not limited to the *State of New Hampshire Terms and Conditions* in Appendix H, which shall form the basis of any Contract resulting from this RFP; No new terms and conditions have been added and no existing terms and conditions have been deleted in this RFP Proposal.

The proposal is effective for a period of 180 days or until the effective date of any resulting contract.

The prices quoted in the proposal were established without collusion with other eligible vendors and without effort to preclude the State of New Hampshire from obtaining the best possible competitive price; and

The vendor has read and included a copy of RFP 2017-073 DOE New Hampshire Statewide Assessments and any subsequent signed Addendum (a).

Our official point of contact is ____
Title ____
Telephone_____________________Email___________________________________
Authorized Signature Printed ____

Authorized Signature __________________________ __________________
____

4.18.3 Table of Contents

The vendor must provide a table of contents with corresponding page numbers relating to its proposal. The table of contents must conform to the outline provided in Section 4.17: *Proposal Organization*, but should provide detail, e.g., numbering, level of detail.

4.18.4 Section I: Executive Summary

The executive summary, which must not exceed five (5) pages, must identify how the vendor satisfies the minimum standards for consideration, which are described in Appendix B: *Minimum Standards for Proposal Consideration*, to this RFP. The executive summary will also provide an overview of the vendor's proposed solution and services. Vendors are encouraged to highlight those factors that they believe distinguish their proposal.

4.18.5 Section II: Glossary of Terms and Abbreviations

The vendor must provide a glossary of all terms, acronyms and abbreviations used in its proposal.

4.18.6 Section III: Responses to System Requirements and Deliverables

System requirements are provided in Appendix C: *System Requirements and Deliverables.*

Using the response tables in Appendix C, the vendor must document the ability to meet the requirements and deliverables of this RFP.

4.18.7 Section IV: Narrative Responses

Section IV solicits narrative responses describing the Software, Technical Services and Project Management topics defined for this RFP Project. Appendix D: *Topics for Mandatory Narrative Responses* is organized into sections, which correspond to the different deliverables and aspects of the scoring process of the proposal. Discussion of each topic must begin on a new page.

4.18.8 Section V: Corporate Qualifications

Section V should provide corporate qualifications of all firms proposed to participate in the project. Specific information to be provided is described in Section E-1: *Required Information on Corporate Qualifications* of Appendix E: *Standards for Describing Vendor Qualifications.*

4.18.9 Section VI: Qualifications of Key Vendor Staff

This proposal section must be used to provide required information on key vendor staff. Specific information to be provided is described in Sections: E-2: *Team Organization and Designation of key vendor staff*; E-3:

Candidates for Project Manager; and E-4: *Candidates for Key Vendor Staff Roles*, of Appendix E: *Standards for Describing Vendor Qualifications*.

4.18.10 Section VII: Pricing Model

The *Pricing Model* must describe the proposed cost of the vendor proposal based on and reflected by the inclusion of the completed tables listed in Appendix F: *Pricing Worksheets*

NOTE: SECTION VII PRICING MODEL, MUST BECOME PUBLIC INFORMATION AND AS SUCH SHALL NOT BE MADE CONFIDENTIAL OR PROPRIETARY. PROPOSALS SUBMITTED WITH ALL OR PART OF SECTION VII LABELED CONFIDENTIAL OR PROPRIETARY SHALL NOT BE CONSIDERED RESPONSIVE AND SHALL NOT BE ACCEPTED.

4.18.11 Section VIII: Copy of the RFP and any signed Addendum (a) - *required in original proposal only.*

4.18.12 Section IX: Appendix- This section provided for extra materials as referenced in Appendix D such as product literature, ad hoc/Federal reporting, interface standards, testing (for UAT plan) and status meetings and reports.

5. PROPOSAL EVALUATION PROCESS

5.1 Scoring Proposals

Each proposal will be evaluated and considered with regard to the solution and services proposed, qualifications of the vendor and any subcontractors, experience and qualifications of proposed candidates, cost and the total quality of the proposed solution.

If the State, determines to make an award, the State will issue an Intent to Award Notice to a vendor based on these evaluations. Should the State be unable to reach agreement with the selected vendor during contract discussions, the State may then undertake contract discussions with the second preferred vendor and so on. Such discussions may continue at the sole option of the State, until an agreement is reached, or all proposals are rejected.

The State will use a scoring scale of **100** points, which shall be applied to the solution as a whole. Points will be distributed among the following factors**:**

28 points – Proposed Solution
21 points – Corporate Overview and Project Management
21 points – Project Execution and Ongoing Operations
24 points – Pricing Model
6 points – Overall Quality of Proposal
100 points – Total Possible Score

5.2 Rights of the State in Evaluating Proposals

The State reserves the right to:

- Consider any source of information, including but not limited to: State employees, internet research and rating agencies, in evaluating proposals;
- Omit any planned evaluation step if, in the State's view, the step is not needed;
- At its sole discretion, reject any and all proposals at any time; and
- Open contract discussions with the second highest scoring vendor, if the State is unable to reach an agreement on contract terms with the highest scoring vendor.

5.3 Planned Evaluations

The State plans to use the following process:

- Initial screening;
- Preliminary scoring of the proposals;
- Oral interviews and product demonstrations; and
- Final evaluation of proposals.

5.3.1 Initial Screening

The State will conduct an initial screening step to verify vendor compliance with submission requirements and to confirm that the proposal satisfies the conditions defined in Appendix B: *Minimum Standards for Proposal Consideration.* A proposal that fails to satisfy either submission requirements or minimum standards may be rejected without further consideration.

5.3.2 Preliminary Scoring of Proposals

The State will establish an evaluation team to initially score proposals.

5.3.3 Oral Interviews and Product Demonstrations

Preliminary scores from the initial evaluation of the proposals will be used to select vendors to invite to oral interviews and product demonstrations.

The purpose of oral interviews and product demonstrations is to clarify and expound upon information provided in the written proposals. Vendors are prohibited from altering the basic substance of their proposals during the oral interviews and product demonstrations.

For each invited vendor, the oral interview and product demonstrations will be **1 hour** in length. A highly structured agenda will be used for oral interviews and product demonstrations to ensure standard coverage of each invited vendor. Information gained from oral interviews and product demonstrations will be used to refine scores assigned from the initial review of the proposals.

5.3.4 Best and Final Offer

The State **will not** be requesting a *Best and Final Offer*. The State plans to negotiate pricing with the highest scoring vendor. If an agreement is not reached, the State reserves the right to move on to negotiations with the second-highest scoring vendor.

The State reserves the right to award this RFP in part, to multiple vendors, if so determined by the State to be in the best interest of the State.

5.3.5 Final Evaluation

The State will conduct final evaluations as a culmination of the entire process of reviewing vendor proposals and information gathering. Reference and background checks will be made for finalist or finalists, as appropriate. After making a preliminary determination of award, the State reserves the right to conduct site visits to a vendor location and/or government site(s) that utilizes the vendor software.

5.4 Scoring Detail

The State will select a vendor based upon the criteria and standards contained in this RFP.

5.4.1 Scoring of the Proposed Solution

The vendor's proposed solution will be allocated a maximum score of twenty-eight (28) points. The main purpose of this section is to measure how well the solution meets the needs outlined in Appendix C, C-1*Scope of Work* and Appendix D, D-1 *Proposed Solution*. The contribution of scoring team members representing all stakeholders will be critical in this section. Factors include but are not limited to:

Proposal

- Responds clearly, concisely and completely to all RFP priorities and requirements.
- Demonstrates understanding of the states' needs and the challenges that need to be met in order to achieve it.
- Proposes methods, procedures and strategies that are sound, innovative, and represent current research and best practice in assessment design and delivery.

Assessment

- Includes summative and interim assessments.
- Includes ELA (to also include a machine-scored/artificial intelligence (AI) writing component) and mathematics to be administered annually in grades 3-8, and science to be administered annually in grades 5, 8, and 11.
- Be substantially aligned and include appropriate representations of the New Hampshire's grade level academic standards for ELA, mathematics and science.
- Align with nationally and/or internationally recognized academic standard(s) with demonstrable longitudinal data supporting both student growth and student achievement models.
- Accommodate changes in academic standards.

Software

- Adapts to the NH DOE and school district systems.
- Supports NH DOE technical strategies and priorities.
- Accommodates changes in academic standards.
- Scalable for potential growth and or changing needs.
- User friendly and intuitive.

System Compatibility

- Employs technologies similar to other state systems.
- Ease of system modification, integration and data storage.

Protection of Data

- Project activities adhere to the highest standards of integrity particularly as they related to student information.

Compatibility with State IT Expertise and Training Approach

- Project details vendor capacity to host a cloud-based assessment delivery system, provide technical support to users and ensure the use of quality protocols.

Reporting System

- Project details vendor capacity to score student assessments with a high degree of accuracy, integrity and efficiency.
- Project activities detail vendor capacity to analyze and report student results to meet a variety of purposes and satisfy the needs of key constituent groups, including educators, policy makers and parents.

Criteria for these scores will be found in but are not limited to:

Proposal Section III: *Responses to Requirements and Deliverables* - Attachment C-2: Requirements, particularly business requirements

Proposal Section IV: *Narrative Responses*

Vendor Presentation and Demonstration

5.4.2 Scoring of Corporate Overview and Project Management

Corporate Overview and Project Management will be allocated a maximum score of twenty-one (21) points.

In this section the State will score the technical merits of how the vendor proposes to carry out the implementation and maintain the solution. The implementation of the solution will require the vendor to customize or configure the application to meet the requirements of the State, monitor and ensure its operation throughout the warranty period and, if maintenance is to be provided, to be a partner in the solution’s operation throughout its useful life. Technical details of the system, administrative procedures, how the vendor manages its team, the project and the technical environment will be critical. Factors include but are not limited to:

Corporate Qualifications

- Longevity and financial stability of business and key businesses partners, if part of this proposal.

- Standing in and evaluation of market position of proposed assessment solution(s).
- Demonstrated competence in working with the proposed product or technology, including examples of product updates (e.g., updates for standard changes) and technology over time as market needs have evolved.
- Depth of required technical skill within the company.
- Successful completion of previous similar projects.

Project Management Competence

- Expertise and experience in priority areas.
- Staff qualification and experience.
- Size and composition of the vendor team.

Proposed Work Plan

Criteria for these scores will be found in but are not limited to:

Proposal Section III: Responses to Requirements and Deliverables and Technical Requirements

Proposal Section IV: Narrative Responses

Proposal Section V: Corporate Qualifications

Vendor Presentations

5.4.3 Scoring of Project Execution and Ongoing Operations

Vendor proposed plan for project executive and ongoing operations will be allocated a maximum score of twenty-one (21) points. It must be established that the vendor company is capable of carrying out the project through implementation, the warranty period and the maintenance period and any contract extensions**.** This project has a tight implementation schedule and vendors should specifically address risks and remediation strategies to address the compressed time schedule. Factors include but are not limited to:

Project Execution

- Implementation approach
- Communication strategies
- Discovery and resolution of problems
- System support structure

Ongoing Operations

- Post warranty operation and support

- Customer solutions and support center proposal
- Technical reporting and advisory

Criteria for these scores will be found in but are not limited to:

Proposal Section III: Responses to Requirements and Deliverables and Technical Requirements

Proposal Section IV: Narrative Responses

Proposal Section V: Corporate Qualifications

Vendor Presentations

Proposed Work Plan

References

5.4.4 Scoring of Pricing Model

Vendor proposed software solution cost will be allocated a maximum score of twenty-four (24) points. The State will consider both implementation costs and subsequent year license and maintenance costs, provided in Tables F-1: *Activities/Deliverables/Milestones Pricing Worksheet*, F-5: *Software Licensing, Maintenance, and Support Pricing Worksheet* and, if appropriate, F-6: *Web Site Hosting, Maintenance, and Support Pricing Worksheet.* The cost information required in a proposal is intended to provide a sound basis for comparing costs. Factors include but are not limited to:

Price Proposal

- Cost effective budget.
- Sound fiscal management practices that meet or exceed industry standards.

Criteria for these scores will be found in but are not limited to:

Proposal Section III: Responses to Requirements and Deliverables and Technical Requirements

Proposal Section IV: Narrative Responses

Section VII: Pricing Model

Vendor Presentations

Proposed Work Plan

5.4.5 Scoring the Overall Quality of the Proposal

The vendor proposed solution provides a comprehensive, coherent and integrative response to the *Scope of Work* and other project priorities that demonstrates the capacity to implement the project on time, within budget and at a high level of quality. A maximum score of six (6) points will be allotted to the overall quality of the proposal.

Criteria for these scores will be found in but are not limited to:

Section I: Executive Summary

Section III: Responses to Requirements and Deliverables

Section IV: Narrative Responses

Section V: Corporate Qualifications

Section VI: Qualifications of Key Vendor Staff

Section VII: Pricing Model

Vendor Presentations

Proposed Work Plan

APPENDIX A: BACKGROUND INFORMATION

A-1 New Hampshire Department of Education (NH DOE) Guidance Regulations for Statewide Assessment

The NH DOE is responsible for the administration of a statewide assessment program in the areas of English language arts (ELA), mathematics and science.

A-1.1 Purpose

The purpose of the *Statewide Education Improvement and Assessment Program* is to establish what New Hampshire students should know and be able to do and to develop and implement effective methods for assessing that learning and its application so that local decisions about curriculum development and delivery can be made (NH RSA193-C:1 VI). In addition, the results of the statewide assessment program must be reported to students, parents, teachers, administrators, school board members and to all other citizens of New Hampshire in order that informed decisions can be made concerning curriculum, in-service education, instructional improvement, teacher training, resource allocation and staffing. (NH RSA193-C:1 IV).

A-1.2 Assessment Tasks

Since the program is not a minimum competency testing program, assessment instruments should be designed to reflect the range of learning exhibited by students. The assessment portion of the program shall consist of a variety of assessment tasks which can be objectively scored. The assessment instruments shall include, but not be limited to:

a) Constructed response items which require students to produce answers to questions rather than to select from an array of possible answers;
b) A writing sample; and
c) Other open-ended performance tasks (NH RSA193-C:3 II).

A-1.3 Assessment Criteria

The following criteria shall be used in the development of the *Statewide Education Improvement and Assessment Program*:

a) Academic standards specifying what students should know and be able to do shall be clearly defined before assessment procedures and exercises are developed.
b) The assessment exercises or tasks shall be valid and appropriate representations of the academic standards the students are expected to achieve.
c) At each grade level assessed, the standards and expectations shall be the same for every New Hampshire student.
d) Teachers shall be involved in designing and using the assessment system.
e) Assessment frameworks and reports shall be understandable and widely disseminated to parents, teachers, administrators, other school personnel,

school board members, teacher preparation programs, business people, government officials and community members.

f) The assessment system shall be subject to continuous review and improvement (NH193-C:3 III).

A-1.4 Assessment Data Use

The assessment system shall generate data which may be used:

- At the student level, by students, parents, and teachers, to determine what the student knows and is able to do in relationship to academic standards.
- Such assessment performance data should be comparable to demonstrable longitudinal data supporting both student growth and student achievement models leading to college and career readiness.
- At the classroom and school building levels, to monitor student progress and provide educators with tools to enhance learning.
- At the district level, to measure school and district-wide progress toward meeting goals and outcomes, to revise curriculum, to design in-service education programs, and to improve instruction.
- At the state level, to measure what students know and are able to do in relation to the attainment of goals and outcomes from the assessment frameworks, and to report the results to the citizens of New Hampshire.
- At the state level, to target services to schools, improve existing programs, develop new initiatives, and revise standards for school improvement, teacher certification, etc.
- At the college level, to integrate into teacher preparation programs instruction in state-established standards, techniques for enhancing student learning in these areas, and the use of assessment results to improve instruction.
- At all levels, to correlate, to the extent possible, with national goals and international standards.
- At all levels, to provide a basis for accountability (NH193-C:3 IV).

A-1.5 Goals and Objectives for NH Statewide Assessments Design: ELA, Mathematics, Science

The primary goals of the NH DOE with respect to assessment delivery include ensuring that all assessments measure student mastery of academic standards and achievement of expected growth; all assessments yield informative and accessible data; vendors provide a smooth transition to any new assessment system; and that this is done in a responsible cost-effective manner through:

- Efficiency and effectiveness relative to time and costs;
- Quality, consistency, and accessibility of information available to state managers, school and district leaders, and student families;
- Reporting features that support local teachers and parents; and
- Accessibility to all students.

A-1.6 Academic Standards

All assessments must include appropriate representations of the New Hampshire's grade level academic standards for ELA, mathematics and science (NH RSA 193-C:3, III) and align with nationally and/or internationally recognized academic standard(s) with demonstrable longitudinal data supporting both student growth and student achievement models leading to college and career readiness.

While individual school districts may adopt local standards, academic standards adopted by the New Hampshire State Board of Education provide a consistent, clear understanding of what students are expected to know and be able to do by the end of each grade level. The standards are designed to be robust and relevant to the real world, reflecting the knowledge and skills that students need for success in college and careers and to compete in the global economy. The standards provide a roadmap so that parents, teachers, and students have a clear understanding of the expectations in reading, writing, language, mathematics and science in school. The standards emphasize content knowledge, critical thinking and problem-solving skills.

ELA Standards:
https://www.education.nh.gov/instruction/curriculum/english_lang/index.htm
Math Standards:
https://www.education.nh.gov/instruction/curriculum/math/index.htm
Science Standards:
https://www.education.nh.gov/instruction/curriculum/science/index.htm

Standards Revision
The NH academic standards for ELA, mathematics, and science are entering the revision cycle. It is therefore imperative that the vendor acknowledge and include in proposals a plan for assessment revision aligned to updated standards in 2018-2019.

A-1.7 Test Type

It is not feasible to cover the full breadth and depth of academic standards for a given grade level with a single external (large-scale) assessment.

NH will therefore need to rely on a combination of two types of assessment strategies:

- **Summative assessments** that are implemented by the state and administered at a time determined by the NH DOE

- **Interim assessments** that are administered at a time determined by the district/school that fits the instructional sequence in the classroom.

To adequately cover all dimensions of the standards, assessment tasks will need to contain multiple components (e.g., a set of interrelated questions). These assessments may include mixed item formats and may use matrix sampling to report scores for groups. Specific components may focus on individual practices, core ideas, or crosscutting concepts, but together, the components need to support inferences about student learning.

Assessment of student mastery of academic standards can be accomplished through a range of assessments that are designed to answer different kinds of questions – with appropriate degrees of specificity – and provide results that complement one another.

A system of assessments need to measure mastery and growth and include three components:

- Assessments designed to support classroom instruction (formative);
- Assessments designed to gauge learning (interim); and
- Assessments to measure student mastery of grade level academic standards and growth (summative).

A-1.8 NH Assessment System RFP Components

- **COMPONENT A: SUMMATIVE ASSESSMENTS IN ELA AND MATHEMATICS.** The ELA (to include machine-scored writing component) and mathematics assessments will be administered annually in grades 3-8.

- **COMPONENT B: SUMMATIVE ASSESSMENT IN SCIENCE.** Science assessments will be administered annually in grades 5, 8, and 11.

- **COMPONENT C: INTERIM ASSESSMENTS IN ELA AND MATHEMATICS.** These online assessments are to be made available for local district use ***at least*** in the same grades included in the summative assessments described above. The purpose of the formative and interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.

- **COMPONENT D: INTERIM ASSESSMENTS IN SCIENCE.** These online assessments are to be made available for local district use ***at least*** in the same grades included in the summative assessment described above. The purpose of the formative and interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.

- **COMPONENT E: REPORTING PORTAL.** The reporting portal must be capable of being upgraded across time. Initially, the system must be able to organize and present assessment data in a way that is easy for all users to read and understand. The Reporting Portal should allow for differentiated access to individual student, classroom, school, district and state assessment data, and they should integrate seamlessly with each other. Assessment results must be able to be provided to parents, among others, in a timely fashion, either on line or by printed report.

- **The vendor must address completely each Assessment Component bid on.**
- **Each vendor must address Reporting Component E.**

A-1.8 Test Design Overview

All services related to test design, item development and review, item banking, test production, online administration, data files and score reports for the assessment components are the responsibility of the vendor selected under this RFP. It is the NH DOE's intention to have an assessment in Year 1 that has quality items in which the vendor can provide verification that proper alignment to academic standards is evident. The assessments used in Year 1 must allow for field testing and result in reliable and valid scores that will be reported for accountability purposes. The NH DOE expects field test items and appear alongside operational items in Year 1. All years subsequent to Year 1 must be adjusted to accommodate for any changes in academic standards. Vendors must propose a plan for customized assessments, based on updated academic standards for all components beyond Year 1.

New Hampshire currently uses the Smarter Balanced assessment. To ensure the state can continue to provide educators, families and students with longitudinal data, for school accountability and individual student growth purposes, vendors should provide assurances that they will be able to include an evidence based concordance table between proposed Assessment Component A and the *Smarter Balanced* Summative Assessment for Grades 3-8.

It is imperative that each proposed Assessment Component be ready for use in the 2017-2018 school year. While we are looking for an off-the-shelf solution, vendor-developed solutions will be considered acceptable if they align to academic standards, can be updated to revised academic standards and will be ready for implementation in the 2018-2019 school year, <u>and</u> meet all requirements as specified in the RFP.

Assessment design/development procedures should include review and revision of proposed assessment and item specifications, review of test blueprints, review and evaluation of items and texts available for use on the test (item development plan), field testing for needed items to complete the desired test design, and final operational test creation.

All phases of assessment design/development shall use accepted validity, reliability and other testing principles including Universal Test Design. A plan for complete data review of field-tested items should be delineated. The plan should include a proposed operational timeline and the resources that they would require from the NH DOE (e.g., "need X number of educators for Y number of days to do Z amount of item review, need X staff from NH DOE for Y days to review operational plans", etc…). The NH DOE must approve all items and the test forms before they become operational.

This RFP covers all tasks necessary for the following: item/assessment design/development, computer-based testing (CBT) platform, test administration; re-test administrations; technical support; psychometric analysis, research and technical activities; customer service; processing, scoring and reporting; test security; quality assurance, training and support; and general program management.

A-1.9 Professional Standards/Best Practices

The vendor shall ensure that all materials, practices and procedures developed under this contract meet relevant professional standards such as those contained in the *Standards for Educational and Psychological Testing* published by the American Education Research Association (2014 or most current version), particularly in terms of privacy, reliability, validity, opportunity to learn, accommodations, scoring, reporting and documentation.

The vendor shall inform the NH DOE when implementation practices or policies are not consistent with the best educational research and practice.

The vendor shall be responsible for clearly communicating the risks of violating conclusions of the best educational research and practice. If NH DOE concurs, the vendor shall work to make necessary corrections. The vendor shall confirm its agreement to meet this requirement.

A-1.10 State and Federal Requirements

The vendor shall ensure that all materials, processes and procedures developed under this contract meet relevant State and Federal legal requirements, including requirements under the Every Student's Succeeds Act (ESSA) and peer review, as well as the Individuals with Disabilities Education Act (IDEA). Throughout the life of the contract, and any extensions, the vendor shall communicate to the State when it concludes that the program is no longer meeting State and Federal requirements and shall provide corrective options to the State for consideration. The vendor shall confirm its agreement to meet this requirement.

A-1.11 Communication

The vendor shall assist NH DOE in explaining to the media, the public, stakeholders, the court, and/or other applicable entities why the tests are valid and reliable assessments that are appropriate for their intended purpose. The vendor shall confirm its agreement to meet this requirement.

A-1.12 Security/Confidentiality/ FERPA

The vendor shall follow FERPA, state and industry standard security policies, including the provision of confidentiality agreements for all vendor staff, subcontractors and educators participating in any aspect of this project. The vendor may include sample confidentiality agreements as an attachment. The vendor shall provide a plan detailing the implementation of security procedures. The vendor may choose to provide additional details under relevant requirements and specifications. The vendor must also indicate the base services (ex. accounting of all secure materials, sealing, forensic analysis, etc.) related to test security which it requires for its state accountability assessments. If the vendor offers a variety of services, but does not have its own base requirements regarding security, that must be indicated in the proposal.

Any breach of security that occurs through the negligence or inaction of a vendor, such as, but not limited to, failure to adhere to any security protocol or allowing raters to remove secure materials from Item Writing Meetings, Item Review Meetings, Data Review Meetings, Anchor Paper Selection, Validation Meetings, or the Scoring Center, will be considered a default on the terms of this contract.

The vendor shall provide assurance that it will meet the requirements of NH RSA 193-C:10 Accessibility of Assessment Materials:

> *"After the assessment results are released by the department, a pupil's parent or legal guardian shall have the right to inspect and review the pupil's assessment, including the questions asked, the pupil's answers, instructions or directions to the pupil, and other supplementary materials related or used to administer the pupil's assessment. A parent or legal guardian shall direct a request for inspection or review to the pupil's school, and the school shall comply with such request within 45 days of its receipt. The department of education shall make available released assessment items on the department's website as soon as possible after the statewide assessment results are released. The commissioner shall adopt rules, pursuant to RSA 541-A, to implement procedures for the review and inspection of assessment materials. These rules shall provide parents and legal guardians with no fewer rights accorded to them under the Family Educational and Privacy Rights Act, 20 U.S.C. 1232g."*

A-1.13 NH DOE Sign Off

All procedures followed in the development, production, administration, scanning, scoring and reporting of the NH assessments shall be made available for review by NH DOE and, as determined by NH DOE, may be subject to NH DOE approval. The vendor shall confirm its agreement to meet this requirement.

A-2 Department of Information Technology and Technology Status

The project will be conducted in cooperation with the New Hampshire Department of Information Technology (DoIT). DoIT coordinates the statewide information technology activities.

A-3 Related Documents Required

Vendors are NOT required to submit these certificates with their proposal. Vendors may be required to be a registered company in New Hampshire. The certificates will be requested from the selected vendor prior to contract approval.

- Certificate of Good Standing/Authority (Appendix G-2–item A) dated after April of the current year and available from the Department of State by calling (603) 271-3244 or (603) 271-3246. Forms are also available on: www.sos.nh.gov/corporate/Forms.html
- Certificate of Vote (Appendix G-2-Item B)
- Proof of Insurance compliant with Appendix H: *State of New Hampshire Terms and Conditions*.

A-4 State Project Team

State high-level staffing for the project will include:

- Frank Edelblut - Commissioner
- Heather Gage – Division Director
- Sandie MacDonald – Bureau Administrator
- Julie Couch – Assessment Administrator
- Vacant – Program Specialist for Statewide Assessments
- Barbara Hopkins – Science Education
- Erik Klardie – Department of Information Technology

A-4.1 Project Sponsor

The Project Sponsor, **Sandie MacDonald,** will be responsible for securing financing and resources, addressing issues brought to her attention by the State Project Manager, and assisting the State Project Manager in promoting the Project throughout the State. The Project Sponsor or an appropriate designee will be available to resolve issues on a timely basis.

A-4.2 State Project Manager

The State Project Manager, **Julie Couch**, will be responsible for:

- Leading the project;
- Promoting the project statewide;

- Developing project strategy and approach;
- Engaging and managing all vendors;
- Managing significant issues and risks; and
- Managing stakeholders' concerns.

APPENDIX B: MINIMUM STANDARDS FOR PROPOSAL CONSIDERATION

A proposal that fails to satisfy the requirements in this section may be rejected without further consideration.

B-1 Submission requirements

- The proposal is date and time stamped before the deadline as defined in Section 2: *Schedule of Events*.
- The vendor has sent the proper number of copies with the original version of the proposal marked "ORIGINAL" and the copies marked "COPY" as defined in Section 4.1: *Proposal Submission, Deadline and Location Instructions.*
- The original proposal includes a signed Transmittal Letter accepting all terms and conditions of the RFP without exception.

B-2 Compliance with System Requirements

System requirements and deliverables are listed in Appendix C: *System Requirements and Deliverables* in this RFP. The proposed vendor's solution must be able to satisfy **all mandatory requirements listed**.

B-3 Current Use of Vendor Proposed Software

Components that constitute the vendor's proposed software suite must be fully implemented and operational in at least one (1) government entity comparable in size and complexity to the State. See Appendix E.

B-4 Vendor Implementation Service Experience

The vendor must have completed the vendor proposed software implementation for at least one (1) government clients comparable in size and complexity to the State within the last three (3) years. The specific vendor proposed software version and functionality must be described. See Appendix D and Appendix E.

B-5 Proposed Project Team

The proposed Project Team must include individuals with substantial experience in statewide assessment design and administration, project management, data analysis and student privacy requirements.

For the purpose of evaluating compliance with this requirement, the vendor team is permitted include subcontractors. In addition, one (1) team member may be identified to fulfill the experience requirement in multiple areas. See Appendix D and Appendix E.

APPENDIX C: SYSTEM REQUIREMENTS AND DELIVERABLES

C-1 SCOPE OF WORK

The *Scope of Work* should be applied to each Assessment Component bid on:

A. Summative Assessments in ELA and Mathematics. The English language arts (to include AI scored writing component) and mathematics assessments will be administered annually in grades 3-8.

B. Summative Assessments: Science. Science assessments will be administered annually in grades 5, 8, and 11.

C. Interim Assessments: ELA and Mathematics. These online assessments are to be made available for local district use ***at least*** in the same grades included in the summative assessments described above. The purpose of the interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.

D. Interim Assessments: Science. These online assessments are to be made available for local district use ***at least*** in the same grades included in the summative assessment described above. The purpose of the interim assessments is to provide timely and useful feedback that local educators can use to gauge the growth and achievement of students throughout the school year and to predict summative learning. Such assessment results may also be aggregated and used in lieu of summative assessments, if permitted under ESSA.

E. Reporting Portal. The reporting portal must be capable of being upgraded across time. Initially, the system must be able to organize and present assessment data in a way that is easy for all users to read and understand. The Reporting Portal should allow for differentiated access to individual student, classroom, school, district and state assessment data, and they should integrate seamlessly with each other. Assessment results must be able to be provided to parents, among others, in a timely fashion, either on line or by printed report.

The vendor will be responsible for the development or procurement of all items included on the components of the NH statewide summative and interim assessments for ELA, mathematics, and/or science bid on.

The vendor's response must reflect familiarity with and understanding of the academic standards in ELA, mathematics and science, as well as the challenges and innovations associated with assessing the depth of the standards.

Accessibility for all students is a core principle for the NH DOE. The vendor's response must discuss the procedures that will be used to support inclusiveness and to ensure that the assessments are designed to be accessible to all students. The vendor's response should address the use of Universal Design (UD), Universal Design for Learning (UDL), and the use of development protocols such as the Accessible Portable Item Profile (APIP).

The assessment will include a variety of machine-scorable item types. The requirement for machine-scorable items must not preclude the use of test items that reflect the depth and complexity of the standards assessed. It is expected that the assessments at all grade levels will make the best use of available technology to include items that assess higher order thinking skills, address the multiple dimensions of the standards, and require students to produce responses as well as select from an available set of responses.

- Each Assessment Component will include a variety of items, types, including multiple choice, constructed response, writing, technology enhanced and performance tasks.
- Assessments should be online technology-based, although some students that require accommodations may require paper and pencil versions.
- The online assessments must work with a variety of devices, such as PCs, Macs, iPads, Chromebooks, and Windows tablets and using a variety of browser products, if necessary.
- Each Assessment Component bid on must address Component E: Reporting Portal. Details for Component E outlined in Appendix D-1.8.
- Results from the assessments will be used for school accountability and federal and state reporting requirements.
- Score reports will be customizable for districts, schools, teachers, parents and other audiences.
- Vendors must demonstrate the ability to be flexible to meet NH requirements and timelines.

Some of the tasks listed in the following section recur throughout the life of the contract (e.g., item development, passage selection, item analysis, technical manuals). It is the vendor's responsibility to fully understand the *Scope of Work* and to project the scope through the potential life of the contract, including any extensions.

It is imperative that each proposed Assessment Component be ready for use in the 2017-2018 school year. While this RFP seeks an off-the-shelf product, customized solutions are acceptable. All solutions must include appropriate representations of the New Hampshire's grade level academic standards for ELA, mathematics and science. (NH

RSA193-C:3 III) and align with nationally and/or internationally recognized academic standard(s) with demonstrable longitudinal data supporting both student growth and student achievement models leading to college and career readiness. Solutions must also be ready for implementation in the 2018-2019 school year, be capable of updating for changes to standards and meet all requirements as specified in the RFP.

C-2 REQUIREMENTS

Vendors shall complete the requirements checklist (Table C-2 General Requirements Vendor Response Checklist). Table C-2 is included as an attachment to this RFP.

C-3 DELIVERABLES

Vendors shall complete the response checklist below: Table C-3 *Deliverables Vendor Response Checklist*. The chart below is a sample of deliverables that represent project milestones and are to be developed by the vendor. Items listed below in the sample can be changed to represent the proposed solution. This table must align to completed pricing worksheets (Appendix F) provided by the vendor.

Table C-3 Deliverables Vendor Response Checklist

	Activity, Deliverable or Milestone	Deliverable Type	Projected Delivery Date
PLANNING AND PROJECT MANAGEMENT			
1	Conduct Project Kickoff Meeting	Non-Software	
2	Project Status Reports	Written	
3	Work Plan	Written	
4	Infrastructure Plan, including Desktop and Network Configuration Requirements	Written	
5	Security Plan	Written	
6	Communications and Change Management Plan	Written	
7	Requirements Trace ability Matrix	Written	
8	Software Configuration Plan	Written	
9	Systems Interface Plan and Design/Capability	Written	
10	Testing Plan	Written	
11	Data Conversion Plan and Design	Written	
12	Deployment and Roll-out Plan	Written	
13	Comprehensive Training Plan and Curriculum	Written	
14	End User Support Plan	Written	

15	Business Continuity Plan	Written	
16	Documentation of Operational Procedures	Written	
INSTALLATION			
17	Provide Software Licenses, if needed	Written	
18	Provide Fully Tested Data Conversion Software	Software	
19	Provide Software Installed, Configured, and Operational to Satisfy State Requirements	Software	
TESTING			
20	Conduct Integration Testing	Non-Software	
21	Conduct User Acceptance Testing	Non-Software	
22	Perform Production Tests	Non-Software	
23	Test In-Bound and Out-Bound Interfaces	Software	
24	Conduct System Performance (Load/Stress) Testing	Non-Software	
25	Certification of 3rd Party Pen Testing and Application Vulnerability Scanning	Non-Software	
SYSTEM DEPLOYMENT			
26	Converted Data Loaded into Production Environment	Software	
27	Provide Tools for Backup and Recovery of all Applications and Data	Software	
28	Conduct Training	Non-Software	
29	Cutover to New Software	Non-Software	
30	NA		
31	Provide Documentation	Written	
32	Execute Security Plan	Non-Software	
OPERATIONS			
33	Ongoing Hosting Support	Non-Software	
34	Ongoing Support & Maintenance	Software	
35	Conduct Project Exit Meeting	Non-Software	

APPENDIX D: TOPICS FOR MANDATORY NARRATIVE RESPONSES

Vendors must limit narrative responses describing the Software, Technical, Services and Project Management topics defined for this Project. The following table identifies specific topics for narratives.

D-1 Proposed Solution
D-1.1 Test Construction
Topic 1 – Test Design
Topic 2 – Item Development
D-1.2 Solution Technology
Topic 3 – Technology Requirements
Topic 4 – Assessment Delivery Platform
Topic 5 – Data Exchange and Process
Topic 6 – Data Privacy and Security
Topic 7 – Technical Compatibility
D-1.3 Security and Protection of Data
Topic 8 – Security and Forensics
Topic 9 – System Security
Topic 10 – Back-up and Recovery
Topic 11 – Assurance of Business Continuity
D-1.4 Training and Support
Topic 12 – User Manuals and Guides
Topic 13 – Training Materials
Topic 14 – Practices Test and Student Materials
Topic 15 – Software Implementation Training
D-1.5 Assessment Scoring, Analysis and Equating
Topic 16 – Machine Scored Items
Topic 17 – Analysis and Psychometric Support
Topic 18 – Calibration and Scaling
Topic 19 - Equating
Topic 20 – Assessment Evaluation
D-1.6 Reporting
Topic 21 – Assessment Scores
Topic 22 - Reports
Topic 23 – Data Files
D-1.7 Standard Setting
Topic 24 – Performance Levels
Topic 25 – Cut Scores
Topic 26 – Standard Setting Report
D-1.8 Reporting Portal
D-2 Corporate Overview and Project Management
D-2.1 Corporate Qualifications

Topic 27 - Corporate Overview
Topic 28 – Vendor Experience
D-2.2 Project Management
Topic 29 – Management Team
Topic 30 - Staff Qualifications and Experiences
D-2.3 Project Plan
Topic 31 – Work Plan
Topic 32 – Management Meetings
Topic 33 – Project Communication
Topic 34 – Program Improvement Plans
Topic 35 – Risk Management and Quality Assurance
D-3 Project Execution
D-3.1 Implementation and Operation
Topic 36 – Implementation Approach
Topic 37 – User Acceptance Testing
D-3.2 Ongoing Operations
Topic 38 - Help Desk Support
Topic 39 – Support Center
Topic 40 – Technical Reporting
Topic 41 – Technical Advisory Committee
D-4 Pricing Model
Topic 42 – Pricing Model
D-5 Assurance and Transition
Topic 43 – Quality Control Sign-Offs
Topic 44 - Invoices
Topic 45 - Transition

D-1 PROPOSED SOLUTION

D1.1 Test Construction

The vendor will propose a plan for the construction of test forms for the Spring 2018 assessment administration and subsequent operational administrations of the NH assessments. The vendor's response must include a description of the processes and procedures that will be used to select items to be included on each assessment form and ensure that each assessment form and the assessment as a whole meets the requirements described in the test specifications. The vendor's response should address the manner in which the NH Instructional Support Teams will be involved in the process of test construction. The vendor's response must include a description of the quality assurance, quality control, and any other review processes and procedures that will be used to ensure the accuracy and technical quality of each of the assessment forms and the assessment as a whole.

Topic 1 Test Design

The vendor must propose a plan for the design of the Spring 2018 field test and the subsequent operational NH assessments that best meets the purposes and intended uses of the results of the NH assessments. The vendor's response must reflect an awareness of the breadth, depth, and complexity of academic standards; the State requirements in RSA 193:C, Federal requirements for assessment as expressed in the Every Student Succeeds Act and US ED assessment regulations; the high level Test Design Framework and preliminary Test Specifications provided by the NH DOE; the NH DOE constraints related to testing time, test format, and the use of machine-scorable items described throughout this RFP, including the AI scored ELA writing assessment, and the NH DOE's desire for a cost effective and efficient assessment program which includes interim and summative components. The proposed field test and operational test designs must demonstrate the vendor's ability to balance those factors to produce an assessment program that meets the following high-level priorities:

- The NH summative assessments are being designed intentionally as relatively short (i.e., 2 hours of testing per summative assessment area, and easy to score (i.e., machine-scorable).

- The summative assessments will align to the corresponding academic standards in ELA, mathematics and science.

- The summative assessments will use a common-matrix design to support a) the reporting of student-level overall science performance in terms of performance levels and scaled scores and b) the reporting of school- and district-level scores in a manner that reflects the depth and breadth of the academic standards.

The vendor shall create a proposed test blueprint for the assessments to be developed in each content area. Interim assessment frameworks should be based on the frameworks developed for the summative assessments. The successful vendor will provide refined blueprints to be reviewed by the NH DOE for approval. The blueprint must specify the numbers of each type assessment item to be used at each grade level in each content area, depth of knowledge, the numbers of items in each eventual test form (field and operational) and the total amount of testing time. The vendor must take into consideration the fact that the assessments must generate valid information throughout the data distribution in both tails.

Reading assessments may include a combination of literary and informational passages. Literary texts include fiction, literary nonfiction and poetry. Informational texts include exposition, argumentation, persuasive texts, procedural texts and documents.

Writing assessments may include a combination of narrative, informative and persuasive purposes.

The vendor must provide its plan for ensuring a high level of consistency between the summative and interim item types and blueprints with adjustments made due to time and item type constraints.

Topic 1.1 Standards Revision

The NH academic standards for ELA, mathematics and science are entering the revision cycle. It is therefore imperative that the vendor acknowledge and include in proposals a plan for assessment revision aligned to updated standards in 2018-2019.

All test design items must be updated to align to academic standards in years subsequent to the 2017-2018 administration. Vendor’s response must provide a full description of its proposed plan for realignment and blue print updates.

Topic 1.2 Test Administration

The summative assessment should be designed to take approximately two hours of testing time, per content area, for the vast majority of students. The test will be loosely timed, and allow for the accommodation of additional testing time into the schedule.

Additionally, each summative assessment component should be structured to be divided into two sessions. Both test sessions should also be available to schools to administer in one day if they chose to do so.

Topic 1.3 Student Registration

The vendor shall be responsible for managing the student/organization registration process. This process shall include the registration of students in private out of school district placements.

The vendor will be able to accommodate any student in home education or private school environments that wish to participate in the assessment process by identifying their registration separate from public school student registrations for school accountability purposes. (Per NH RSA 193-C:6: Home educated students will contact their local school districts if they wish to participate in the statewide assessment. Private schools will contact the department of education to participate in the statewide assessment.)

After the window for registration is complete, the NH DOE shall approve the registration counts. The vendor must propose a plan for allowing the NH DOE the opportunity to review and amend registration information.

The vendor’s response should include a recommended timeline for the registration process (relative to the test administration window) to ensure that accurate information is captured, appropriate quality checks occur, and to allow the NH DOE a sufficient window of time to review and amend the registration information.

The vendor’s response should include a plan for allowing the registration of students enrolling in schools after the end of the test registration window and for students moving between schools during the testing window.

Topic 1.4 Accessibility and Fairness

Accessibility is a core principle of the NH DOE. The NH DOE is committed to ensuring that all students are able to have equitable and fair access to the NH assessments, including access to assessment items, training materials and supports. Information about students’ required accessibility features and accommodations must be gathered and maintained in accordance with Federal laws.

The vendor’s response must include a draft list of appropriate assessment accommodations separately for students with disabilities and English language learners.

Accommodations listed must be supported by the most current research. The list shall describe the test accommodations and supports that allow access for students with disabilities and English learners to most fully participate in each assessment without interfering with the measurement of the constructs. Vendors shall also discuss accommodations which would threaten the validity of the assessment by interfering with the construct being measured.

Vendors are asked to review the accessibility features and accommodations policies for the current state NH assessments in English language arts, mathematics and science as well as best practices for ensuring accessibility with computer-based tests. The vendor's response must detail their plans for meeting accessibility requirements. The vendor's response should address how their proposed assessment system compares to the states' current systems and explain how it will address accessibility, accommodations, and fairness – while maintaining data privacy and security.

Vendors must include a description of how students with visual impairment will access on-line assessments or be provided with other accommodation, as appropriate.

Vendors must include a description of how students unable to access online assessments will participate in the assessments.

Topic 2 Item Development

This section of the RFP addresses tasks related to the development, evaluation, and selection of the items that will be included in each Assessment Component (A, B, C, D). Each of the assessments may include a combination of items developed specifically for the assessment program and items procured from other sources.

While this RFP seeks an off-the-shelf solution, such a solution as well as a customized solution, may/will involve the creation of specific test items (defined below). This section dealing with Item Development applies to the development of items, whether used to tailor an off-the-shelf solution in response to this RFP or in a customized solution. Per NH RSA 193-C:3, III (d), "teachers shall be involved in designing and using the assessment system."

Throughout this section of the RFP, the term item development will be used to refer to both items that are custom-developed for the NH assessments and items procured by the vendor from other sources. Whether custom-developed for the NH assessments or obtained from other sources, all items included on the assessments will be subject to the review processes described in the RFP.

The vendor's response should describe the vendor's general procedures for item development including the use of outsourced or free-lance item writers, the use of items previously developed by the vendor, and/or the procurement of developed items from other sources. The vendor's response should include a description of how all item writers are trained on the content of the academic standards, in general, and on any specific content criteria related to the assessment component.

In addition to addressing the specific tasks described in the RFP, the vendor's response must also meet the following requirements:

- The vendor's response must reflect familiarity with the academic standards as well as current best practices and recommendations regarding the assessment of student achievement in each content area bid on.
- The vendor's response must reflect familiarity with computer-based testing and the use of a variety of item types, including technology enhanced items (TEI) to assess students' higher order cognitive skills as well as their knowledge of core ideas and concepts.
- The vendor's response must discuss the procedures that will be used to ensure that all assessment components are accessible to all students. The vendor's response should address the use of Universal Design (UD), Universal Design for Learning (UDL), and the use of development protocols such as the Accessible Portable Item Profile (APIP). The vendor's response should also address how technology will be used to enhance accessibility.
- The vendor's response must indicate how the security of items will be maintained throughout the development and item review process, including procedures that will be taken to ensure the secure transfer of items and item information to/from states during the development process.
- If vendors believe that tasks not specified in this RFP are critical to the development of quality items and a testing program, they should identify and describe the significance of those tasks in their response.
- The writing component of the English language arts assessment must include constructed response item(s) as it relates to a reading passage to be included in the students overall score. The writing component must not simply ask students to reply to a writing prompt.

Topic 2.1 Item Development Team

The successful vendor will appoint an Item Development Team, as necessary, responsible for the development of items for the each Assessment Component bid on. The vendor's response will identify key personnel and describe the proposed composition of the item development team, including describing the responsibilities and time commitments of the proposed members.

The vendor's response will include a description of how the specific needs of students with disabilities and English language learners will be accounted for within the proposed item development team. If the vendor proposes that a single person will fulfill multiple roles within the team (e.g., lead developer and grade level developer; developer at multiple grade levels) that must be clearly described in the vendor's response. The vendor's response must include a rationale to support the proposed composition of the item development team.

The vendor's response must include, as necessary, a description of the procedures, including the use of technology that will be used to facilitate interactions among the vendor's Item Development Team and the NH Instructional Support Team.

The vendor's response must describe the type, number, and duration of in-person and virtual meetings between the item development team and the NH Instructional Support Team that will be needed throughout the development cycle to produce quality items for the tests. All costs for proposed in-person development meetings (including travel costs and lodging) should be included in the proposal.

Topic 2.2 Item Development and Review

The vendor shall propose and describe the process that will be used to interact with the NH Instructional Support Teams throughout the item development and review process. The vendor's response should reflect an understanding of the responsibilities of the NH DOE staff and NH educators and propose a process that avoids unnecessary travel, makes the most efficient use of their time, and allows sufficient turnaround time for review and approval of all items and related materials.

The vendor will describe the type and number of in-person and virtual meetings that will be held during an annual development cycle.

The vendor's response should include a proposal for an initial in-person meeting between the NH Instructional Support Team for each assessment area (ELA, mathematics, science) and the vendor's Item Development Team at the beginning of the project.

Topic 2.3 Item Review Committees

The NH DOE believes that the use of Item Review Committees consisting of educators from the state is a critical part of the item development process. The Item Review Committees not only provide a unique and important perspective during item review, but also enhance the transparency of the assessment program and increase understanding and buy-in for the program.

The primary purpose of the Item Review Committees is to provide feedback on the content of items proposed to be included for field testing on the next administration of the assessment, including the alignment of content to academic standards, accuracy of content, and appropriateness of content for the grade levels being tested. The committee may also be asked to provide feedback on the use and appropriateness of specific item types and to provide suggestions for future item development.

The vendor will support grade-level item review committees (3-8 ELA; 3-8 mathematics; 5, 8, and 11 science) for each of the assessments proposed. The committees will consist primarily of grade-appropriate teachers recruited and selected by the NH DOE. Additional committee members may include local curriculum coordinators, content specialists, and ESOL or special education specialists. The NH DOE will determine the composition of committees.

Meetings of the Item Review Committees will be jointly facilitated by the NH Instructional Support Team and the vendor's Item Development Team. The committee will meet annually in NH during the summer months.

The NH DOE is proposing an annual summer meeting due to the availability of educators and the timing of the meeting within the development cycle. If the vendor thinks that additional meetings of the Item Review Committee would be necessary and/or that the timing of the meetings should be changed, the vendor's response should include a proposal for an alternative meeting schedule. In particular, the vendor should indicate if they think that additional meetings will be needed during the initial year of the contract in preparation for the spring 2018 assessments.

The vendor's response should indicate the steps that will be taken to maximize efficiency throughout the item review process and, in particular, should describe how the vendor will make the most efficient use of the limited time available with the members of the Item Review Committees. This could include proposing a structure for the meeting that would provide for the most effective and efficient use of people and resources (e.g., organized by grade level, organized by content).

In preparing a response, the vendor should plan on supporting the item review committee meeting with the following specifications:

- Each grade level committee will consist of 3-6 members.
- Committee members will be paid a stipend of $150 per day for participation in the summer meetings. (If the vendor proposes meetings during the school year, the stipend will be replaced by a corresponding payment to the committee members' school district for substitute reimbursement).
- Representatives from NH will also attend the meeting, including the NH Instructional Support Teams (up to a maximum of three (3) people total).
- The vendor will support and arrange for lodging for committee members and NH representatives. The vendor's response should presume that all participants will require 3 nights lodging for the meeting (beginning one night prior to the meeting).
- The vendor will also be responsible for travel expenses (e.g., mileage, airfare) for all participants to attend any out of state meetings.
- The vendor will provide breakfast and lunch each day of the meeting and be responsible for dinner expenses ($25 per day) on days which require an overnight stay.

Meetings will be held at a hotel, conference center, or similar suitable location in NH. The meeting location may vary around the state or may be held in a central location to minimize travel requirements and expenses for committee members. The vendor will work with the NH DOE to develop a timeline to ensure that the NH Instructional Support Teams have sufficient time to review and provide feedback on all materials and items prepared for the Item Review Committee meetings.

The vendor will schedule an additional meeting day following the conclusion of the Item Review Committee meeting for a meeting of the vendor's Item Development Team and the appropriate NH Instructional Support Team to reconcile item feedback.

The vendor will produce a written report documenting the meeting and recommendations within two weeks of each committee meeting.

Topic 2.4 Bias/Sensitivity Review Committee

The vendor will support a Bias/Sensitivity Review Committee consisting of external educators and experts recruited and selected by the NH DOE to review the content of passages, other stimuli, and test items for potential bias and sensitivity. The NH DOE will determine the composition of the committee.

There will be a single committee responsible for reviewing materials across grade levels for each content area; ELA, mathematics, science (i.e., there will not be separate bias/sensitivity committee for each grade assessed). Each committee (ELA, mathematics, science) will contain five (5) members.

The meetings will be facilitated by the vendor in coordination with a representative from the NH DOE. The committee will meet in NH annually during the summer. The vendor may propose additional meetings, if necessary during the initial year of the project.

In preparing a response, the vendor should plan on supporting the bias/sensitivity review committee meeting with the following specifications:

The committee will consist of five (5) members.

- Committee members will be paid a stipend of $150 per day plus travel expenses for participation in the summer meeting. (If additional meetings are proposed during the year, the $150 stipend will be replaced by a corresponding payment to districts for substitute reimbursement for any committee members who are employed by local education agencies.)
- The meeting will also be attended by a representative of the NH DOE (1 person).
- The vendor will support and arrange for lodging for committee members and the state's representative attending the meeting.
- The vendor will provide breakfast and lunch each day of the meeting and be responsible for dinner expenses ($25 per day) on days which require an overnight stay.

The Bias/Sensitivity Review Committee will focus on review of stimuli proposed for the development of new field test items, review of newly developed items recommended for field test, and beginning in 2018 after the initial field test, review of items recommended for inclusion in the operational item bank that have been flagged for Differential Item Functioning (DIF) analysis.

If feasible, the Bias/Sensitivity Review Committee meetings may be scheduled concurrently with the Item Review Committee. The vendor's response should propose a process that will help avoid the costs and lack of efficiency of having assessment items go through development and be flagged for bias/sensitivity (content review, not empirical DIF analyses) only after substantial investment in development effort.

The vendor will produce a written report documenting the committee meeting and recommendations within two weeks of each meeting.

Topic 2.5 Content Review

The vendor must ensure that the content of all items recommended for field-testing is accurate and reflects the current state of knowledge in the appropriate field. The vendor's response must describe their methods and procedures for meeting this requirement within the item development process.

The vendor shall work in collaboration with the NH DOE to ensure all assessments adhere to current and future standards for ELA, mathematics and science. The State has begun a process for review and revision of the ELA, mathematics and science and expects to field test in the 2018-2019 school year.

Topic 2.6 Item Types and Number of Items

As described previously, all items on the ELA, mathematics and science assessments must be machine/AI-scorable. The requirement for machine/AI-scorable items, however, must not limit the use of item types to traditional multiple-choice or selected-response items. The vendor's response must describe the variety of item types that could be included on the assessment components bid on, including item types that require students to generate or produce a response as well as select a response. The vendor's response must include a description of the vendor's experience with each of the item types proposed and provide access to sample items to allow the NH DOE to review proposed item formats.

- Single Correct Response Items:

  Selected-Response – These are items in which students are presented with several answer choices, one of which is correct. Students are asked to select the correct answer. With an online assessment, this could include items such as drag and drop and hot spot items. Selected-response items should be designed to assess the highest level of content knowledge and cognitive complexity that can be appropriately measured through this type of item appropriate to the evidence outcome being assessed.

  "Griddable" Response – Although "griddable" response items are too cumbersome or costly for paper/pencil administration, online administration provides for single number or single word answers making use of the item type more feasible. These items should be easy for the computer to score without requiring extensive engine training.

- Short Constructed Response – These are items in which students write in a response to a question or a prompt. Student responses consist of 1) one to five sentences, 2) written work in solving a mathematics problem, 3) drawing, or 4) another response that can typically be provided in five minutes or less.
- Extended Constructed Response - These are items in which students write in a response to a question or a prompt. Student responses consist of one to several paragraphs. Student responses can be typically be provided in 20 minutes or less in ELA, mathematics, and science.
- Performance Items – These items are performance-based. They require students to engage in higher order thinking and to provide a response that is more complex than a simple correct or incorrect answer. Student responses may consist of written or illustrated integrated answers. Response time will vary across grades and content areas from 20-60 minutes.
- Innovative, Interactive Technology-Enhanced Items – These are sophisticated, simulation-based items that require students to respond with a virtual performance. The items utilize advanced simulation engines featuring online tools and content not previously assessed through selected-response or constructed-response item formats. Students' responses are expected to be computer-scored.

Topic 2.7 Number of Items

The vendor's response must include a proposed plan for the number of items of various types that will need to be developed for the Spring 2018 administration and subsequent operational test administrations. The vendor's response should address the number of items that will be administered to an individual student as well as the total number of items that will be administered across matrix-sampled forms on each Assessment Component bid on.

The vendor's response should reflect an understanding of the NH DOE's intended use of the results from each Assessment Components, the type of scores that will be reported, the plan to release items, as well as an understanding of academic standards. The vendor's response should also reflect an awareness of the testing time and cost constraints discussed throughout this RFP.

Topic 2.8 Item Release

The vendor will propose a plan for the annual release of a representative sample of test items. The purpose of releasing items to provide local educators with information about the type and level of knowledge and skills assessed on the assessments, the variety of item types used on the assessments, and the rigor of the items on the assessments.

The annual release of items will consist of up to 25 percent of the items (points) on a single student test form. The vendor's response should address how common items in a common-matrix design may be used to support the release of items.

Release of items will begin with the first operational administration of the summative assessment. All item types should be represented in the release of items. The balance of item types should be proportional to their use on the assessment. Items will be released in a digital format that enables local educators to interact with the items in the same manner that the items would be encountered on the assessment. Released items will be accompanied by supporting materials including relevant item statistics, information about the knowledge and skills assessed by the item, information on how the item was scored, and information regarding correct and common incorrect responses to the item.

Topic 2.9 Rotation of Common Items

The vendor will propose a plan for the rotation of common items across years. The plan should address issues related to security, item exposure, maintaining content balance, and stability of assessment forms across years. At a minimum, the proposed plan should address:

- The number (or percentage) of items that should remain in place for consecutive years.
- The number (or percentage) of items that should be replaced after each test administration (including released items).
- The number of years before the common items on an assessment are totally refreshed.
- The maximum number of years, if any, that an item should be included in the operational test bank.

Topic 2.9.1 Item Bank

The vendor will develop and maintain an item bank of all items developed for and included on the assessments. The item bank will include a database that provides electronic access to each item (text and graphics) as well as pertinent information for each item, including item information, item history (placement, item statistics for all administrations of the item, editing, and context), and current item status.

The NH Instructional Support Teams and Management Teams should have appropriate access to generate reports and/or view items and item information, as needed. The vendor's response will include a description of its existing software to meet this requirement or describe plans to develop or procure the software necessary to meet this requirement.

The vendor's response will describe steps that will be taken to ensure the security of the items.

The vendor's response will describe how items and item information developed for the NH assessments through contracts awarded by the NH DOE will be accessible by the NH DOE at the conclusion or termination of those contracts.

Topic 2.9.2 Use of items from other sources

The vendor will describe how the interests of the NH DOE will be assured and protected if items from other sources are included on the assessments. In particular, the vendor's response will describe:

- How items eligible for use on the NH assessments will remain secure, including any procedures in place to ensure that items are not released by other assessment programs or used for any other non-secure purposes.
- How license agreements will be structured to ensure that items may be used on the NH assessments for multiple administrations.
- The vendor's experience in handling any restrictions that may be placed on the use of items from other sources that would negatively impact the NH DOE.

D1.2 Solution Technology

Topic 3 Technology Requirements

The vendor will provide the test delivery platform, hosting site, test administration application, server, and application management services for the NH summative and interim assessments.

The vendor will be responsible for the maintenance of the full system; including code updates and/or patches, technical support, hosting, management, coordination, and support for customer-facing administration activities.

The vendor's response must provide a full description of its proposed computer-based test administration solution. The vendor's response should address each of the following:

1. Requirements for the use of any software (and supporting devices) should be clearly documented and explained.
2. The minimum and preferred technology infrastructure needed to support online testing should be documented and explained.
3. The technical support documents should include information about suggested computer lab configurations.
4. Information on computer-based assistive technologies should be provided to the client so that the client can determine which they may allow; data on use of these technologies should be collected.
5. Practice and training tests should be provided to allow students to become familiar with keyboarding and navigation techniques and tools that will be used during the live assessment.
6. Procedures for uploading student demographic data in the online assessment system, including any necessary accessibility tools and supports, should be provided, as well as instructions and procedures for modification of enrollment data, where permitted by the client.
7. Procedures for maintaining the security of the online testing environment should be documented.
8. Descriptions of training protocols to be provided at the local level on the test administration procedures should be provided.
9. In the first two years of the program, the vendor will be responsible for providing up to four (4) one-half day regional trainings on system use and test administration procedures, to be supplemented by an on-line webinar and other online training materials (e.g., slide deck from webinar, FAQ document). In subsequent years, in-person training sessions may be replaced by a series of webinars.
10. Technical support should be available via telephone and electronically with tools such as help desk and/or email. (see additional details in the Support Center section below).
11. Metrics for monitoring and documenting systems performance should be identified and described.
12. Documentation should be provided regarding the capacity of the system to support the current and potential future range of item types.
    Provide documentation regarding the application's capacity to import and export as applicable: items, student item response data, student registration, demographics, and data regarding eligible and utilized accommodations.

Topic 4 Assessment Delivery Platform

The vendor will ensure that the assessment delivery platform provides the technical infrastructure necessary to manage and administer assessments across the state. The vendor's response will address each of the following subcomponents and functionalities:

- Administrative portal;
- Test registration and scheduling;
- Test administration (administrator interface);
- Test delivery (student interface);
- Test client;
- Key-based and rule-based machine scoring;
- Assessment delivery data storage; and
- Student toolset (e.g., virtual calculators, protractor, ruler, notepad, highlighter).

The vendor's response will also address each of the following functionalities:

- Authentication/User Identity Management: internal user management, user authentication, role-based authorization.

- Logging and Audit: A centralized capability for logging, log analysis and audit support, capturing and recording all system and testing activities at sufficient detail to detect conformance and compliance issues, and track errors. Logging is also used to capture data for analytics and secondary analyses.

- System Monitoring and Alerting: A centralized system for monitoring all processes and systems (network, hardware, software) in the assessment system and sending alert notifications whenever behavior fall outside of nominal ranges. Also, a system for monitoring and alerting support system data and test security.

- Common ID system: A centralized system for assigning and managing persistent, unique identifiers to all users (educators and students) of the NH assessment system. The purpose of this service is to assure the integrity of student data, including to avoid multiple creation of the same ID number, and to prevent the mismatching of students to assessment results.

The vendor will provide a detailed description of the interfaces and the System components used for processing.

- Describe the software platform that the system operates on (code base, database, etc.). Note any third party platform components. Indicate the need for the State to purchase licenses.

- Include details of the proposed software Solution including the database management system, licensed software suggested for data retrieval and reporting, proposed approach to developing any custom-built software components.
- Discuss plans for anticipated future release of System software. Address any impact on System users or interfaces.

Topic 5 Data Exchange and Process

The vendor will provide a detailed description of the mechanism and tools included in the proposed System to enable the specified data sharing between the vendor and the State.

- Identify the type of interface/mechanism/tool and the frequency of data exchange between the State and the vendor with a full explanation of the processes involved in the exchange.
- Identify the format of the data the vendor will provide.

The vendor will provide a detailed description of how State staff will track status of the data submissions and follow to view a record of:

- Administrative actions: Login, Logout, Password reset, IP address, batch file transmission;
- Data submission including user, date, time, and IP address;
- Users viewing validations by date, time, file, IP address;
- Users viewing completeness measures by date, time, file, IP address;
- Users making corrections by date, time, file, record, element, and IP address.

The vendor will provide a detailed description of the process the State will follow to submit special requests for research. The vendor may wish to include a sample scenario with the format of returned results.

The vendor will describe and provide samples of the available documentation supporting the system and the asset verification service.

Topic 6 Data Privacy and Security

All hardware and software components and all services and processes must ensure the highest levels of auditable security for the NH assessments, data, and data access at all times and at all levels of state, district, and school use. The vendor will be expected to comply with Federal laws data privacy and security that include how data are accessed, stored, and exchanged; and how the vendor's employees are managed and trained regarding data security protocols

The vendor's response must detail their privacy and security plans. The vendor's response should address how the proposed solution and associated activities will employ security protocols and design features to meet the states' rigorous security needs for data encryption, identity management, data access, and redundant layers of data protection.

Topic 6.1 Information Technology Standards

The vendor will provide a detailed description of how the proposed solution complies with established information technology standards. The proposed solution must comply with Open Standards and Open Data Formats as mandated by RSA 21-R (HB418 2012). Include the following:

- A description of the degree to which the solution complies with information technology standards on the State of NH web site: http://www.nh.gov/doit/internet/vendors.php
- A description of Open Source Software
- A description of Open Data Format?
- A statement of compliance with privacy and confidentiality standards including HIPAA, NIST, FERPA.

Topic 7 Technical Compatibility

The NH DOE is committed to an assessment system that utilizes solutions that recognize the heterogeneity of technology capacities in the states districts and schools, while supporting the leading-edge assessment methodologies and technologies. Solutions need to provide optimal performance in high-technology capability settings that have current generation computers and large bandwidth networks, but that still function without sacrificing performance in low-technology capability settings. This core principle includes a "device agnostic" approach to assessment content and assessment technology development. All assessment components must be designed to function comparably across a range of devices using commonly deployed web browsers, including desktops, laptops, netbooks, and tablets (9.5" or larger) running Windows, Mac, Linux, Apple iOS, Android, and Chrome operating systems.

The vendor's response must describe how it will ensure and verify that its system functions comparably across a range of devices. The vendor's response should address how it will ensure that the system is not impacted by upgrades or other changes to devices or operating systems.

Topic 7.1 Interoperability

Interoperability is a core design principle for NH assessments technology development and operations. The NH DOE is committed to the application of open technology interoperability standards in order to make assessments, assessment items, and assessment data formats portable across organizations, systems, and districts. Reliable and flexible interchanging of data between components that are both internal and external to the assessment system, and across diverse networks, are key requirements of the NH Assessment System.

The vendor's response must detail their plans for ensuring interoperability. The vendor's response should address its compliance with industry-recognized, open-licensed interoperability standards and the processes and procedures used to verify and validate interoperability.

D1.3 Security and Protection of Data

Topic 8 Security and Forensics

The vendor must propose a plan and describe procedures for maintaining and monitoring security of test items, other secure materials, and student data both within and external to the computer-based test administration system before, during, and after test administration, including ensuring security throughout the test design and development process.

Topic 8.1 Test Security

The vendor's response should address the following areas in general test security:

- Develop and implement a comprehensive plan to ensure the security of test items, materials, and student data throughout the assessment cycle.
- Develop and implement training procedures and materials regarding test security, and confidentiality of student data and personally identifiable information
- Develop protocols for the secure collection, storage and destruction of secure and confidential teacher and student information.

- Develop and implement uniform policies and procedures for identifying and dealing with possible security breaches and testing irregularities.
- Develop implement procedures to account for and protect secure materials at all stages of distribution, receipt, storage, and return. Note: This requirement has general implications, but applies specifically to paper-based test forms.
- Chain of Custody for materials shipped or transported: Develop and implement policies, guidelines and sign-off procedures for State, District, and School officials to establish and document a chain of custody for hand-offs to ensure that documents are received, accounted for, and distributed and returned.
- Provide a secure architecture to protect the development and administration environment from network-based attacks.

Topic 8.2 Data Forensics

The vendor will apply procedures to monitor, detect, and evaluate the assessments for potential cheating, and provide documentation to the NH DOE. The vendor's response should describe plans and procedures to provide continuous updates that capture a variety of data including but not limited to:

- Time of testing,
- All student answer choices including the final choice used for scoring;
- Response latency;
- Tracking the movement of the examinee through the test;
- Student response times;
- Accessibility options used by the student; and analysis of student gains over time; and
- Differential performance on common and matrix-sampled items, if applicable.

Topic 8.3 Test Monitoring

The vendor shall describe in detail the steps that it would take to monitor the fidelity with which the test administration and security procedures are being applied.

In New Hampshire the district test coordinators are expected to report violations of administration and security procedures to the NH DOE. Should a district test coordinator contact the vendor, the vendor shall redirect the district test coordinator to verify with NH DOE that the issue was appropriately reported.

Topic 9 System Security

The vendor shall provide a detailed description of the security design and architectural features incorporated into the proposed system. At a minimum, the vendor will discuss the following:

- Describe the practices employed to ensure that your system and staff comply with FERPA regulations.
- Describe the system assurance provisions incorporated into the proposed system. At a minimum, discuss the following:
    a) What process or methodology is employed within the proposed system to ensure data integrity?
    b) To what degree does the approach rely on system assurance capabilities of the relational database management system (RDMS)?
    c) If multiple databases are employed, what extra procedures are employed to ensure synchronization among databases?
- Discuss the company's practices pertaining to the following security testing:
    a) The identification and authentication methods used to ensure that users and any interfacing applications are identified and that their identities are properly verified.
    b) The authorization methods used to ensure that users and client applications can only access data and services for which they have been properly authorized.
    c) The immunity methods used to ensure that unauthorized malicious programs (e.g., viruses, worms and Trojan horses) do not infect the application.
    d) The methods used to ensure that communications and data integrity are not intentionally corrupted via unauthorized creation, modification or deletion.
    e) The methods used to ensure that the parties to interactions with the application cannot later repudiate or rebut those interactions.

f) The intrusion detection methods used to ensure the detection, recording and review of attempted access or modification by unauthorized individuals.

g) The privacy methods used to ensure that confidential data and sensitive communications are kept private.

h) The system maintenance methods used to ensure that unauthorized system maintenance does not unintentionally disrupt the security mechanisms of the application or supporting hardware.

i) The testing methods conducted to load and stress test your system to determine its ability to withstand Denial of Service (DoS) attacks.

j) Your software patch schedule employed to protect the software from new security vulnerabilities as they arise.

k) The ability of your system's software to be installed in a "locked-down" fashion so as to turn off unnecessary features (user accounts, operating system services, etc.) thereby reducing the software's security vulnerabilities and attack surfaces available to system hackers and attackers.

Topic 10 Backup and Recovery

The vendor shall provide a detailed description of the backup and recovery processes used to protect mission-critical data. The State seeks a sound backup and recovery provision as part of the solution.

The vendor will:

- Describe the tools used for backup and recovery of applications and data.
- Describe the impact of the proposed backup process on the operation of the System.
- The vendor will address the following:
    - Use of and method for logging and journalizing;
    - Single points of failure and recommended approaches for their elimination; and
    - Approach to redundancy.

- Describe options to have the collected data stored at the vendor's site in addition to sending results along to the State. Include a proposed retention schedule

Topic 11 Assurance of Business Continuity

The vendor shall provide a detailed description of the business continuity plan that mitigates risk to the State.

- Provide information on business continuity plans in the event that the hosting site becomes unavailable.
- Discuss plans for moving operations to a remote site if the hosting site is incapacitated.
- Discuss recovery time objectives and how the company will continue to meet federally required response metrics.
- The State believes that additional software license fees solely related to redundancy for assurance of business continuity would be inappropriate. If the proposal differs from this standard, describe and provide rationale for the difference.

D1.4 Training and Support

The vendor will be responsible for providing the training and support required to ensure the administration of the NH assessments, including maintaining a support center to provide quality customer service and support to districts and schools throughout the registration, testing, and reporting cycles. The vendor will develop test coordinator and test administrator manuals to ensure effective administration of the NH assessments.

The manuals will be provided in formats that will permit them to be accessed via the internet. Posted documents must be available for viewing and downloading and must be provided in ADA compliant format.

Topic 12 User Manuals and Guides

The vendor shall develop and produce an *Online User's Guide*. The guide shall provide technical specifications for use of the online platform used for testing.

Information shall include but not be limited to: hardware specifications, proctor caching requirements if needed, student data upload process, data editing information, detailed information on the use of the assessment tools, and other technical guidelines as necessary. Thumbnail art shall be included as much as possible. Separate guides may be provided with focuses for technical and assessment staff. The guide shall be provided in PDF format for posting to NH DOE and vendor websites and in Word for use by the NH DOE in creating training and informational materials.

Topic 12.1 Test Coordinator Manual

The *Test Coordinator Manual* will focus on the tasks that must be completed at the district and school level, including scheduling, meeting technology requirements, student registration, accessibility, maintaining security, and the training of Test Administrators on test administration policies and procedures as well as security policies and protocols.

Topic 12.2 Test Administration Manual

The *Test Administration Manual* (TAM) will provide all directions needed by the test administrator to prepare for and administer the assessments, including security procedures.

The vendor shall develop and produce a TAM for each assessment per administration. This document is provided to assist the test administrator during the testing session. Administration requirements for all grade levels and/or content areas assessed for each assessment shall be included in the TAM. Separate TAMs may be developed for field test administrations, online assessments, and paper assessments.

The TAM shall contain general instructions for administering the assessment including, but not limited to, information about checking materials, planning testing schedules, organizing classrooms, preparation of students, use of standardized testing procedures, administering practice activities, security of materials, completing the student demographic portion of the response documents, accessibility instructions, test directions, administration of the test, assembly of materials for scanning and processing, checklists for class, school and district level procedures and information for returning materials. Thumbnail to full scale images of documents, forms, and other ancillary materials as needed with illustrations and explanatory diagrams shall be used extensively. The TAM will include scripts necessary to administer the assessments, and procedures and scripts necessary for accommodated testing outside of the assessment delivery system.

The TAM shall be reviewed prior to each administration and revisions shall be made to reflect changes related to the program, State and/or federal guidelines. The vendor shall make these documents available in printed form and for downloading from the Internet on a secure site.

Topic 13 Training Materials

The vendor will provide training and training materials to support the efficient and secure handling of materials as well as standardized administration activities. All proposed training materials and activities will be subject to NH DOE approval.

The vendor must provide training and training materials for district/school assessment coordinators, test administrators and district/school technology coordinators. As appropriate, the training must include information about student registration procedures; administration protocols; security policies, protocols, and procedures; the assessment delivery system; and accessibility and accommodations policies and protocols. The vendor should design training modules to enhance efficiency across types of users.

All training materials will be provided in formats that will permit them to be accessed via the internet. Posted documents must be available for viewing and downloading and must be provided in ADA compliant format.

Training materials should for test administrators should include the opportunity to practice all steps necessary to administer the assessment, including experiencing the assessment from the student's perspective.

The vendor will be responsible for providing annually up to four (4) one-half day regional trainings throughout NH on system use and test administration procedures, to be supplemented by an on-line webinar and other online training materials (e.g., slide deck from webinar, FAQ document). In subsequent years, in-person training sessions may be replaced by a series of webinars.

In addition to the regional training described above, the vendor's response must describe the type and amount of training that the vendor feels is necessary to ensure the administration of the NH assessments (Summative, Formative, Reporting). The vendor's response should address the type of training materials that will be used including narrated PowerPoint web presentations, WebEx or other similar webinar tool, or videos, in addition to hard copy documents.

The vendor's response must propose recommended methods and procedures for ensuring that test coordinators, test administrators, and technology coordinators have accessed the relevant training materials, have participated in and completed the required training, and/or are certified to fulfill their responsibilities in administering the assessment.

Topic 13.1 Teacher Directions

The vendor shall develop and produce teacher's directions for each assessment per administration. The teacher's directions shall contain specific instructions for the administration of each grade level and/or content area per assessment. The teacher's directions shall include information related to test administration including but not limited to, test security, the timing of tests and/or subtests, and the number of items on each assessment part.

A script for the administration of each content area shall be included to ensure consistent and appropriate directions are given to students to begin the test. The teacher's directions shall be reviewed prior to each administration and revisions shall be made to reflect changes related to the program. Directions shall be provided electronically.

Topic 14 Practice Tests and Student Materials

The vendor shall provide practice tests for each of the Summative Assessment Components bid. Sufficient opportunity for students to become familiar with and comfortable in the online testing environment prior to testing is critical to ensuring validity and allowing students to demonstrate what they know and are able to do.

A key purpose of the practice tests is to allow students to experience and become familiar with the computer-based test experience prior to testing.

- The practice tests will include all item types and/or response formats that a student may encounter during testing.
- The practice tests will include all support and accessibility features and functionalities that a student will have access to during testing.
- The items on the practice tests will include a range of content, depth of knowledge, and rigor.
- The practice test should require approximately 30 minutes, but no more than 45 minutes, for students to complete.
- Student scores on each item should be provided to students at the conclusion of the practice test.
- The practice tests will be updated, as needed, to incorporate new item types, response formats, or other assessment features and functionalities.

In addition to the practice tests, the vendor's response should describe written materials, online tutorials, or other supports that may be developed to ensure that students are prepared to function within the online testing environment.

Topic 15 Software Implementation Training

The vendor will provide a detailed summary of proposed training approach to include:

- Recommended training approach (instructor led vs. computer based)

- Training evaluation tools
- Training coordination
- Description of training materials and plans for revision
- Training timeline

Topic 15.1 Training and Professional Development

Training and support for the NH assessments shall be provided by the vendor to NH educators as needed for each assessment component. The vendor must include in its proposal a detailed plan of action and timeline that describe how and when each of the training and support tasks will be accomplished.

Topic 15.2 Technology Director Training

The vendor shall describe its training plan for district technology directors. This training may include training on the operation and features of the online assessment system. It may include training on the physical and electronic security of assessments, system requirements for implementing the online assessment and troubleshooting of technology issues at the school or district site. Training may include a visual as well as oral presentation and may include other types of interactive technology. The delivery mode for these activities must be identified. The vendor shall discuss the role NH DOE will have in previewing each training session and webinar.

Topic 15.3 Assessment Administration Training

The vendor shall describe its assessment administration training plans for district test coordinators and test administrators. This training may include how to sign up for the interim assessment program, as well as how to enroll students in the summative assessment. Training related to the actual test administration should also be discussed. Training may include a visual as well as oral presentation and may include other types of interactive technology. The delivery mode for these activities must be identified. The vendor shall discuss the role NH DOE will have in previewing each training session and webinar.

Topic 15.4 Assessment Results Training

The vendor shall describe its plans for providing educators with tools to evaluate and analyze assessment results in order to make informed instructional and programming decisions. Training may include a visual as well as oral presentation and may include other types of interactive technology. The delivery mode for these activities must be identified.

The vendor shall discuss the role NH DOE will have in previewing each training session and webinar.

Topic 15.5 Online Training Support

The vendor shall describe its plan to provide training and customer support specific to online assessment. The description should include training with an easy to understand set of directions, including screenshots, for operating the online assessment software. The vendor may also include other beneficial training materials in its response such as e-learning modules and online tutorials for users.

<u>D1.5 Assessment Scoring, Analysis and Equating</u>

The vendor is responsible for the accurate and efficient scoring of all items on the NH assessments. The vendor's response must include a description of the methods used to ensure and verify that each student's response has been captured and scored accurately.

The NH DOE is interested in providing real time scoring to students and teachers for the interim assessments. The efficiencies of automated and artificial intelligence scoring must be maximized.

Topic 16 Machine Scored Items

The vendor's response must include a description of their experience scoring all item types proposed for use on the NH assessments as well as a detailed description of the methods that will be used to ensure and verify the accuracy of scores from each type of item.

The vendor's response must include a description of type of information that will be collected and available to states related to scored student responses, particularly for items that require students to generate a response, make multiple selections, or have complex scoring algorithms.

Topic 16.1 Automated Scoring of Student-Generated Responses

The vendor's response should address the current and near-term feasibility of using automated scoring to score student-generated text responses of varying lengths (e.g., single word, 1-2 sentences, paragraph, and extended essay). In addition to issues related to technical quality and accuracy of scoring, the vendor's response should address, if applicable, issues such as cost, development time required, testing time required, and impact on the breadth and depth of content coverage on the assessment.

The vendor must indicate in detail its experience in developing or using artificial intelligence (AI) software in scoring student responses. The description shall illustrate the vendor's experience with using AI scoring for each of the proposed item types, as well as limitations to the use of the vendor's artificial intelligence scoring engine for each of these item types. This includes past and current projects, the software used in each, the manner in which the vendor proposes to use its software for this assessment component, the issues that it anticipates in using its software in NH, as well as the areas in which it anticipates that its software will not be effective. The vendor shall describe how its AI engine functions, including how it is trained in relationship to content. The vendor shall provide its projected plans, if any exist, for improving its AI scoring capacity, including a description of why the company believes that this is a realistic goal. This description shall include specific time frames and must be considered within the context of the projected online implementation schedule of each content area in the summative assessment component.

For all constructed response items it is proposing to bring to the NH Assessment System, the vendor shall present its current procedures for development and selection of training papers for scoring of constructed response items and training of the artificial intelligence scoring system. The vendor shall also present its plan for development and selection of training papers for scoring of constructed response items and training of the artificial intelligence scoring system for items developed specifically for NH. The role, if any, of NH educators in validating the rubrics and scoring of the training papers should be discussed.

Topic 17 Analysis and Psychometric Support

The vendor is responsible for designing and conducting all analyses necessary to report student, school, district, and state results from the NH assessments and for ensuring that the NH assessments meet standards of technical quality for high-quality state assessments. In particular, the vendor is responsible for designing and conducting all analyses necessary to provide evidence that the assessment program meets relevant U.S. Department of Education Peer Review requirements.

The vendor shall describe how the different types of scores it is proposing, individual scale scores and subscores, will be produced and verified. The vendor must include scores produced strictly on items which are computer-scoreable and scores produced based on a combination of the computer-scored and hand-scored items. The limitations in interpretation of both of these scores must be discussed.

The vendor shall also describe how the Interim Assessment reports will provide predictive information regarding expected performance on the summative assessment. The timeline for providing this information must be discussed.

Topic 18 Calibration and Scaling

The vendor will calibrate test items and develop a scale(s) for each of the NH assessments using appropriate item response theory model(s).

The vendor's response must propose a recommended model(s) for item calibration and scaling and provide a rationale for the recommendation that includes:

- A discussion of the benefits/advantages and limitations of the proposed model(s);
- Its appropriateness for the type of items that will be included on the NH assessments;
- Its appropriateness based on anticipated initial student performance on items aligned to NH Academic Standards for ELA, Math and Science; and
- Its appropriateness for the type of scores that will be reported from the NH assessments.

The vendor's response must identify the software that will be used to perform item calibration and scaling and include a description of the vendor's familiarity and experience with the software. If the vendor is proposing the use of proprietary or open-source software, the vendor's response must include a description of the steps that will be taken to ensure and verify the accuracy and reliability of the software.

Topic 18.1 Calibration Plan

The vendor's response must include a description of how items from the Spring 2018 Field Test will be calibrated and placed on a common scale. The vendor's response should address how a matrix-sampling test design will impact and be accounted for in the calibration process for the Spring 2018 Field Test and future operational test administrations.

The vendor's response must include a description of how embedded field-test items on operational administrations of the NH assessments will be calibrated and placed on NH assessment scoring scales.

The vendor's response must demonstrate an understanding of the NH DOE's desire to take advantage of the matrix-sampling design to produce school-level results. The vendor's response should address the feasibility of producing independent subscales for particular domains or dimensions within each of the NH assessments.

The vendor's response must include a description of the calibration process and/or established concordance table that will align the student results on the proposed NH Summative Assessments for ELA and mathematics with the current Smarter Balanced reporting scale.

Topic 19 Equating

The vendor will design and conduct all analyses required to equate the NH assessments from year to year.

Topic 19.1 Equating Plan

The vendor's response must include a description of how it proposes to equate the NH assessments from year to year. The vendor's response should address how a common-matrix sample test design will factor into the equating design and also indicate the equating model that will be used.

Topic 19.2 Equating Verification

The vendor's response will include a description of the steps that will be taken to ensure the accuracy of equating results.

The vendor will support an independent real-time review of the equating process, analyses, and results by independent vendor(s), identified by the NH DOE. The vendor(s) will support this effort by providing the consultant(s) with the necessary data files and other materials in a timely manner during the equating process.

Topic 19.3 Equating Report

The vendor will produce an annual report documenting the equating process and results. The report will be available for use by the NH DOE in evaluating and approving the results of the equating process prior to reporting.

Topic 20 Assessment Evaluation – Item Evaluation

The vendor will design and conduct all analyses required to evaluate the quality and performance of all items developed for and/or included on the NH assessments. The vendor's response must include a description of item statistics that will be generated and other analyses that will conducted. The vendor's response should address how the appropriateness of items for all students will be examined and how the use of matrix-sampling may impact item evaluation.

Topic 20.1 Field-Test Item Evaluation

The vendor's response must include a description of the processes that will be used to generate appropriate information to support the evaluation of field test items.

Topic 20.2 Operational Test Item Evaluation

The vendor's response must include a description of the processes that will be used to generate appropriate information to support the evaluation of items that will be used to generate student and school scores and items that will be used to equate tests from year to year.

Topic 20.3 Test Construction Evaluation

The vendor will conduct analyses and provide psychometric support necessary to support the construction of technically sound test forms that meet all of the purposes and intended uses of results from the NH assessments.

The vendor's response should address how it proposes to use item statistics and information from psychometric analyses to support test construction.

The vendor will design and conduct all analyses necessary to produce accurate results student, school, district, and state reports.

Topic 20.4 Additional Assessment Analyses

The vendor shall describe its proposed procedures for assuring that the assessments will be scored in a reliable and valid manner. Reliable and valid scoring for subgroups must also be discussed.

The vendor shall describe its system's capabilities in identifying unusual responses such as those that indicate abuse or potential for student self-harm and in flagging those responses on Interim Assessments for the teacher to review, and on Summative Assessments to notify the NH DOE.

D1.6 Reporting

The vendor is responsible for the accurate and timely reporting of results of the NH assessments. The assessment results are to be reported in a "user friendly" format. The vendor shall describe the process it has used or will use to develop the formatting of the reports.

The NH DOE is especially interested in reporting approaches that provide actionable information for students, parents, and classroom teachers. The reporting system should be designed to complement instruction and to facilitate the use of assessment results to improve student achievement. Reports should reflect areas of strength as well as areas that need to be targeted for instruction.

The NH DOE is also interested in collecting and reporting annual longitudinal student achievement data for the purpose of measuring growth and setting academic goals. The vendor shall describe the process it has used for collecting, monitoring and reporting longitudinal growth data and provide sample reports.

Results from the first operational administration will be reported following standard setting.

Complete results from subsequent operational administrations of the NH assessments will be reported in a timely fashion at the completion of testing.

The vendor's response must include a detailed description of the processes that will be used to ensure the production of accurate color reports at the student, school, district, and state levels, including information on quality assurance and quality control procedures that will be used to ensure and verify the accuracy of reported results.

The vendor's response must also include a plan for the design, review, approval, and production of color reports. The plan should describe how the vendor will interact with the NH DOE throughout the design process.

The vendor must discuss the availability of reports in languages other than English.

The vendor will describe and provide examples of the various report forms distributed to students, schools, and districts. It will contain supportive information related to interpreting the test results, including but not limited to: reporting categories assessed and definitions for technical assessment terms.

The comprehensive interpretive guides shall be developed for use by schools and districts and shall be posted on the vendor's information portal and available electronically for the NH DOE, schools and districts to download.

The vendor shall collaborate in the development of the interpretive guides to ensure accurate information related to the assessment design is clearly provided. From the comprehensive guide, a smaller Parent Brochure shall be developed containing information pertinent to student level reports. The parent brochure shall be translated into one or more languages for distribution to non-English background parents as requested by the NH DOE. The parent brochure shall be distributed with the printed test scores and posted on the vendor's information portal.

The specifications for the interpretive guides and Parent Brochure include but are not limited to:

- Available in electronic format that is accessible via the Internet.
- Include thumbnails and larger images of selected reports.
- Interpretive guides are developed for the purpose of providing schools and districts with an understanding of the reports that are available.
- Interpretive guides include training information for use of online reporting tools.
- Parent Brochure: developed for the purpose of providing test awareness for parents and students, shall include training information for use of online reporting tools.
- Parent Brochure: provided in print format on a 1:1 ratio to student reports per content area per assessment.

Topic 21 Assessment Scores

The primary student scores reported on each of the NH assessments will be an overall achievement score and performance level. A student's achievement score will be based on her/his performance on all operational items (common and matrix) included on the test form and reported on a vertically scaled learning continuum.

The vendor's response should propose and describe options for the type of additional student scores that could be supported by the proposed design of the assessment.

The NH DOE is interested in providing reporting measures that contain actionable information, such that teachers and parents can use results to connect students with targeted instructional and academic materials that meet and challenge the student's abilities, interests and learning objectives. Should the vendor propose to utilize the scores of a subcontractor that can help identify appropriate reading materials, those costs must be listed as an option.

Topic 21.1 Aggregate School, District, and State Scores

A primary focus of the NH Assessment System is to provide detailed information about student achievement at the school and district level. An intended benefit of the use of matrix sampling is the ability to provide more detailed information about performance at the school, district, and state levels than can be provided at the student level.

In addition to aggregate student scores such as mean scaled score and the percentage of students performing at each achievement level, school and district reports should contain detailed information about performance on critical aspects of the NH academic standards.

The vendor's response should describe how matrix sampling will be used to produce reliable subscores that provide useful information and support valid inferences about school and district performance at one or more levels below overall achievement.

The vendor's response should also indicate whether such school and district scores will be reported on their own subscales which can be linked across years to allow comparisons in performance from one year to the next.

Topic 22 Reports

The vendor shall propose a process for the design of reports that includes participation of the NH DOE management team and the NH Content Teams. The vendor's response shall describe the process and procedures that will be used to generate initial design specifications and concepts, to facilitate review and revision, and for the approval of report designs.

The vendor's response should address the feasibility of obtaining external feedback on proposed report designs.

Topic 22.1 Types of Reports

The vendor will be responsible for producing a variety of reports intended for use by a variety of audiences.

Topic 22.2 Student Reports

The vendor will produce hard copy student reports (one per student) that will be shipped directly to schools.

The vendor will also produce a printable, digital version of the student report that may be printed by the district or school.

Topic 22.3 School, District, and State Reports

The vendor will produce school-, district-, and state-level reports in printable, digital format.

The vendor's response will propose a system for providing schools and districts with efficient and secure access to confidential and non-confidential reports.

Examples of the school-, district-, and state-level reports to be produced include:

- **Rosters** providing individual student-level results at the school or classroom level (dependent upon data available). Rosters may also include item-level results for released items.
- **Summary aggregating results** from the Roster at the school or classroom level. May include school, district, and state comparisons.
- **School Report Package** containing information on school participation and performance including performance level results, use of accommodations, subgroup results as required by the USED and subscore results. The report may also include selected results from the released items, district and state comparisons, and comparisons with previous years.
- **District Report Package** containing the same information as the school report aggregated at the district level.
- **State Report Package** containing the same information as the school report aggregated at the state level.
    - **School Summary Report** providing summary participation and performance information across grade levels tested within the school.
    - **District Summary Report** providing the same information as the school summary report aggregated at the district level.
    - **State Summary Report** providing the same information as the School Summary Report aggregated at the state level.

Topic 23 Data Files

In addition to printable, digital reports, the vendor will provide the information contained in all reports in a data file in an agreed upon format that can be imported into the NH DOE and schools' reporting systems.

Interpretive Material

To the extent possible, all reports should contain embedded information to support and promote the proper interpretation and use of the results provided on the report.

The vendor will also propose the development of supplemental materials to assist in the interpretation and use of NH assessment reports by the parents and students, local educators, and the public.

Interpretive materials will be developed in digital form for web-based delivery. Vendors may propose options for printable text materials as well as materials in other media such as videos or interactive graphics.

The vendor will also support two in-person reporting workshops following the first two operational administrations of the NH assessments.

D1.7 Standard Setting

Student results from the NH assessments will be reported according to performance levels. Details on the number and names of performance levels will be determined during 2017-2018.

The vendor shall describe its standard setting methodology for summative and interim assessments. Use of empirical data, including summative data, as well as the any inclusion of NH educators in the process must be discussed.

The vendor's plan to ensure continuity between grade levels must be described.

Cut scores indicating the level of student performance required to attain each performance level classification will be determined in the summer following the first operational administration of the NH assessments.

The vendor shall recommend methods of validating cut scores across time, including approaches to revising as evidence indicates. In addition, the vendor shall suggest how to approach communication with the field regarding potentially changing cut scores.

The vendor will support the NH DOE in all activities related to establishing performance standards for the NH assessments. Major activities that are the responsibility of the vendor are described in the tasks that follow.

Topic 24 Performance Levels

The vendor will be responsible for proposing, organizing, and supporting a process for supporting the NH DOE in determining the number and names of performance levels appropriate for the NH assessments. Such performance levels shall be set, at a minimum, to meet federal assessment reporting requirements.

The vendor will be responsible for proposing, organizing, and supporting a process for developing appropriate performance level descriptions for the NH assessments.

The vendor's response will include a description of the processes that are proposed to determining the number and names of performance levels and then to develop appropriate performance level descriptions. The vendor's response should include a timeline of major activities and/or milestones in the process.

The vendor will be responsible for all costs associated with proposed meetings, including expenses, stipends and/or reimbursement costs for external panelists. Vendors should use $150 per day per panelist for stipends/substitute reimbursement when budgeting for these meetings.

Topic 25 Cut Scores

The vendor will be responsible for proposing, organizing, and supporting a process for determining performance level cut scores (i.e., thresholds) on each of the NH assessments.

The vendor's response will propose a standard setting method that is appropriate for use with the NH assessments. The vendor's response will include a rationale for the use of the proposed method and will address how the method will be applied with the matrix-sampled design of the assessments.

The vendor's response will include a description of the processes and procedures necessary to implement the proposed standard setting method. The vendor will be responsible for all costs associated with standard setting meetings, including expenses, stipends and/or reimbursement costs for standard setting panelists. Vendors should use $150 per day per panelist for stipends/substitute reimbursement when budgeting for standard setting meetings.

The vendor's response should describe the role of the states in setting performance level cut scores before, during, and after any proposed standard setting panel meetings.

Topic 26 Standard Setting Report

The vendor will prepare a report describing and documenting the entire Standard Setting Process. The report will be delivered in digital format no later than one month following the completion of the standard setting process.

Topic 26.1 Standard Setting Validation

The vendor's response should include a plan for conducting analyses to validate the performance standards following the second operational administration of the NH assessments.

D1.8 Reporting Portal

The NH Assessment Reporting Portal should provide a platform that seamlessly integrates data from state summative and interim assessments, providing stakeholders with a user-friendly interface that increases access to results. The NH DOE prefers a system that would allow users to customize particular aspects of their individual dashboard profiles. The vendor should describe the features of its Reporting Portal, including the extent to which its system includes the preferred features. The system must be designed to allow the state access to high level information and would ideally allow students and parents to access detailed information. In addition, educators, school administrators and district administrator roles must be included. For costing purposes included in the *Pricing Model*, state costs for making the system operable to the educator level should be provided.

Vendors may choose to provide a separate *Pricing Model* for a parent portal option and a separate *Pricing Model* for student portal option.

The optional parent portal should provide parents with a user- friendly platform that allows them to access their child's assessment results, as well as other classroom and school information. The vendor shall indicate whether or not its system includes features for parents. The vendor shall describe the parent- related features of its dashboard system, including the extent to which its system includes the preferred features. Pricing for the parent portal should be included separately as an option.

Vendors may describe how the student portal will allow for students to customize their individual pages. Pricing for the student portal should be included separately as an option.

D-2 CORPORATE OVERVIEW AND PROJECT MANAGEMENT

D2.1 Corporate Qualifications

Topic 27 Corporate Overview

See Appendix E

Topic 28 Vendor Experience

The vendor must present a description of corporate capabilities. The vendor shall provide the company's history, including the number of years that it has been in business, buyouts, takeovers, IPO's, bankruptcies, litigations and claims, etc. within the last five (5) years, or for that period which the firm has been in business, if less than five (5) years. Situations arising in assessed liquidated damages (LDs) must be described with their resolution, along with the amount of the LDs or provided additional services.

The description shall also identify the number of employees in the company and the company's location(s), including any presence in New Hampshire. The overall capacity of the vendor's organization(s) and the resources that it will commit to the work for the project (by name and role in project) shall be discussed.

The description shall also outline the vendor's overall position in the State assessment market, including the length of time, states served, addition/loss of states over the past five (5) years.

A general description of the vendor's capabilities and capacities related to development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities shall be included. Responses must demonstrate that the vendor meets, at a minimum, the mandatory qualifications presented at the beginning of this component.

Specific examples of the vendor's work products such as test and item specifications, items, forms, technical manuals, research reports, technical services, etc. should be identified under the relevant requirements and specifications and provided in attachments as appropriate. NH DOE expects to receive the same or better quality of work throughout the contract, including any extensions, as the examples that are provided in the proposal.

Topic 28.1 Relevant Experience

In tabular format, the vendor shall provide a listing and descriptions of all work in similar projects that it and its proposed subcontractors have carried out or are carrying out for other clients. The table shall include client, program name, content area, grades, administration mode (paper-pencil or computer-based), use of scoring (human and artificial intelligence), length of contract and number of students.

For computer-based testing, the vendor shall include the total number of tests administered and the highest number of on current testers. For each such project, the vendor must provide the name of the state or other organization, name of client contact person, this individual's telephone, email and fax numbers, and e-mail address

Current Use of Vendor Proposed Software – Current Implemented Sites of Vendor Proposed Software

Components that constitute the vendor's proposed software suite must be fully implemented and operational in at least one (1) government entity comparable in size and complexity to the State of New Hampshire.

D2.2 Project Management

The vendor will be responsible for the effective and efficient management of the project. The vendor's response must include a description of the procedures and processes that will be used to ensure the effective and efficient management of the project. The vendor's response should address areas such as scheduling, communication (internal, with the states, with districts and schools), and coordination across tasks and parties.

Throughout this document, the terms "vendor" and "vendor(s)" are also assumed to include subcontractors where appropriate and applicable. If the vendor proposes to subcontract any part of the work, the vendor's response must refer to the subcontractors where appropriate. Within the relevant requirements and specifications, a description of each proposed subcontractor's role in the project, qualifications to perform that role, management structure, key staff assignments and qualifications of assigned staff shall be included.

If the vendor has discovered fault with a subcontractor named in this RFP, the vendor has the obligation to inform NH DOE immediately and the appropriate steps must be taken by either the subcontractor or the vendor to correct the problem prior to that problem resulting in substandard performance or non-compliance. The vendor shall remain responsible for the performance of its subcontractors.

Topic 29 Management Team

The vendor shall provide a list of key staff, including but not limited to, the program manager, program coordinator, lead psychometrician, content development lead, content specific area lead, technology lead, special populations consultant, scoring manager(s), production manager(s), and publication staff, as well as all staff assigned 0.20 FTE or greater to each component. Each staff member's assigned responsibilities and time allocated to the project must be provided. Time expected to be allocated to other projects must also be indicated. In no case should an individual be assigned to more than one full-time equivalent position.

The vendor shall affirm in the response to this request for proposal that should the contract be awarded, all key personnel proposed shall be released from any concurrent responsibilities that would impede their availability to assume the work as proposed.

The vendor shall assign one person to function as the **Program Manager**. That person must be responsible for all activities required by the project and will serve as the main contact person between the vendor and NH DOE. The Program Manager shall have the authority to make decisions and commitments on behalf of the vendor, subject to NH DOE approval.

NH DOE requires that a **Technology Consultant** be identified. This individual shall be responsible for a number of tasks, including but not limited to, assisting NH DOE and the districts with the transition to assessment system; working with NH DOE information staff to ensure the accurate and efficient transfer of data to and from NH DOE; creating, defining and reviewing file layouts; providing assistance in the verification of demographic data; and assisting NH DOE and district assessment coordinators with the use of vendor websites and functionality.

NH DOE reserves the right to interview and approve all key staff including subcontractor staff. Throughout the life of this contract, and any extensions, changes to the assigned program manager, program coordinator, lead psychometrician, content development lead, content specific area lead, special populations consultant, and technology consultant, except for those resulting from separation of services, will require prior written consent by NH DOE. In the event that NH DOE requests removal of specific vendor personnel, the vendor shall provide acceptable replacement(s) with no impact to the project. Replacement(s) shall have qualifications which meet or exceed the original staff member proposed or the staff member holding the position previously and shall be approved by NH DOE.

All personnel who will work on-site at NH DOE or school sites may be required to be pre-approved for site access via a criminal background check paid for by the vendor.

Topic 30 Staff Qualifications and Experiences

Qualifications of all key personnel shall be presented in the vendor's proposal, including subcontractors. Supporting resumes outlining education/training, employment history, and experience in conducting work similar to what is expected under this contract shall be included as an appendix.

NH DOE requires a **psychometric team that** will not only execute routine functions, but will also be able to provide a sophisticated level of expertise to guide the psychometric decisions that will need to be made and re-evaluated as the program evolves. The expectation is that the team will be able to provide psychometric options with strengths and challenges and its recommendations along with rationale. In addition, especially in the event of unexpected challenges, the team must include someone with both extensive experience and psychometric knowledge, as well as the decision-making authority to quickly address and remedy the situation.

For all meetings involving educators, the vendor must indicate the qualifications of the facilitators. General qualifications for training and meeting facilitators must be included in the response to this request for proposals. Facilitators must be familiar with best practices, as well as state and federal laws, procedures and regulations concerning assessment. As applicable, facilitators must also be familiar with academic instruction of students and the educational and assessment landscape. Facilitators must be able to clearly articulate spoken English and create easily understood written materials and visual training aids. Facilitators must have demonstrated experience in leading large-group trainings including webinars and meetings as fit their responsibilities.

Organizational charts, including identification of Program Manager and key personnel, for the vendor as a whole and for the NH DOE project team specifically, including subcontractors where applicable, must be provided. The charts shall clearly indicate lines of authority and communication within and among the vendor's departments and subcontractors, where appropriate.

The vendor shall also describe its escalation process for resolving any vendor/client disagreements.

The Program Manager directly in charge of overseeing the NH project shall be identified. This manager shall be available both during and outside of normal business hours to assist with any urgent situations. Contact information for this individual shall be provided at the time of contract award.

Changes to the assigned Program Manager, except for those resulting from separation of services, require prior written consent by NH DOE. The replacement shall have qualifications which meet or exceed the original staff member proposed or the staff member holding the position previously and shall be approved by NH DOE.

Qualifications of key executive personnel must be presented. A supporting resume outlining education/training, employment history, and experience in conducting work similar to what is expected under this contract shall be included as an appendix.

The vendor shall fulfill this requirement and all requirements listed in Appendix E and Appendix H

D2.3 Project Plan

The vendor shall describe the planned project management activities as they pertain to the three phases, planning, Implementation, and operations. In addition to addressing the components listed in *Appendix D2.3 Work Plan*, the vendor shall provide an example of status reports prepared for another similar project. Names of the project and of any individuals involved may be removed. ***Topic 31 Work Plan***

The State requires vendors to present a thorough project Work Plan in the proposal addressing all work offered in their proposal. The State will evaluate the proposed project Work Plan contained in the Proposal to determine how well it will serve the needs of State Project leaders.

The State sees a Work Plan as essential to reaching a comprehensive agreement with a vendor. Consequently, the State will seek to refine the proposed Work Plan during contract finalization with the selected vendor and to incorporate the refined Work Plan by reference into a contract. In addition, the State will require the selected vendor to update the Work Plan in consultation with the State during the term of the project.

Topic 31.1 Preliminary Work Plan

Provide a preliminary Work Plan for the planning and implementation phases of the engagement. The vendor's preliminary proposed Work Plan includes a description of the schedule, tasks, deliverables (with pricing), major milestones, task dependencies, and a payment schedule. The Work Plan shall also address resource allocations (both State and vendor team members). Include sufficient detail that the State will be able to identify departures from the plan in sufficient time to seek corrective action. In particular provide information about staffing. Identify and discuss the following:

- All assumptions upon which the work plan is based;
- Descriptions of recommended roles by activity and time required for both State and vendor members of the project team;
- Assignments of members of the vendor's team identified by role to specific tasks; and
- Critical success factors for the project.

The vendor Work Plans should include information pertaining to resource allocation, update frequency, financial check points and a graphic overview.

Topic 31.2 Project Plan and Schedule

Proposals shall include a detailed schedule reflective of the Work Plans that describe how each of the requirements and specifications described in the proposal will be accomplished. The schedule shall at a minimum identify the tasks, subtasks, beginning date, end date and the party/functional group responsible for each step in the process. The schedule must be included as a separate attachment to the proposal.

The proposed plans and schedule shall clearly identify and include:

- Key activities related to the field (ordering of materials, receipt of materials, test dates, return of materials, demographic clean-up window, release of individual student scores, final individual student, school and district score file release, and receipt of paper reports)

- Key transfer dates between the vendor and NH DOE related to development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities.

At the beginning of the project and by the beginning of each fiscal year, the vendor will develop a detailed project plan and schedule for the coming fiscal year.

Activities related to the development for the next year's assessment and reporting for the prior year's assessment must be clearly distinguishable from activities related to the current year's assessment.

Joint review of this schedule followed by NH DOE's approval for the first contract period should occur within two weeks of the contract award. The vendor and NH DOE shall mutually agree upon final dates. Joint monitoring of the schedule shall occur on an on-going basis. The vendor shall ensure that all schedule adjustments allow for final deliverable dates to be met. If necessary, timelines and schedules may be revised with prior approval of NH DOE and an executed contract amendment for all deliverables subject to liquidated damages.

A revision of a timeline on the part of the vendor exempts the vendor from meeting a contractual deadline only if (1) the vendor and NH DOE mutually agree upon and document through a contract amendment an extension of the deadline as executed through a contract amendment or (2) the vendor is able to prove that the deadline was not met due to NH DOE's failure to meet a contractual deadline resulting in the vendor's inability to adhere to the schedule for delivery of products and services.

The vendor shall alert NH DOE as soon as it believes a deliverable subject to liquidated damages is at risk of not meeting its delivery date.

NH DOE must be notified whenever the New Hampshire contract is included in vendor's internal meetings focused on programs at-risk.

For the contract beginning after July 1, 2017, the review of the schedule should occur within the first two weeks of the initial contract. For each following contract year, by May 1, the vendor shall provide an updated detailed Work Plan and project schedule that specifies all activities leading to products or services deliverable to either NH DOE or local school districts for the following assessment year.

The development of the project plan and schedule will follow a review of the current project status and contract specifications by the vendor and NH DOE. Any foreseeable changes to contract requirements and/or costs will be discussed and agreed upon during this process and reflected in the project plan and schedule.

Topic 32 Management Meetings

The vendor will be responsible for organizing and supporting regular management meetings with the NH DOE project management team. The vendor will be responsible for costs associated with management meetings.

An initial two-day, in-person management meeting will be held shortly after the contract is awarded. Participants will include key vendor staff and State project leaders. This meeting will enable leaders to become acquainted and establish any preliminary project procedures.

Additional meetings shall include:

- Status Meetings: Participants will include project leaders from the vendor and the State. These meetings, which will be conducted at least twice monthly, will address overall project status and any additional topics needed to remain on schedule and within budget. A status report from the vendor will serve as the basis for discussion.
- Special Meetings: Need may arise for a special meeting with State leaders or project stakeholders to address specific issues.
- In Year 1 of the project, weekly phone calls between pertinent NH DOE staff and the vendor's Program Manager and other key vendor staff shall be held between in-person project meetings to keep NH DOE current on project status, discuss issues as they arise, and to plan upcoming activities.
- Exit Meeting: Participants will include Project leaders from the vendor and the State. Discussion will focus on lessons learned from the Project and on follow-up options that the State may wish to consider.

The vendor shall include the following when describing the meeting process:

- Timing, duration, recommended participants and agenda for the kickoff meeting;
- Frequency and standard agenda items for status meetings;
- Availability for special meetings; and
- Agenda for the exit meeting.

The State expects the vendor to prepare agendas and background for and minutes of meetings. Background for each status meeting must include an updated Work Plan. Drafting of formal presentations, such as a presentation for the kickoff meeting, will also be a vendor responsibility.

Vendor shall submit reports in accordance with the schedule and terms of the contract. All reports shall be prepared in formats approved by the State. The vendor's project manager shall produce reports related to project management as reasonably requested by the State. Vendor shall produce project status reports, which shall contain, at a minimum, the following:

- Project status as it relates to Work Plan
- Deliverables status
- Accomplishments during weeks being reported
- Planned activities for the upcoming two (2) week period
- Future activities
- Issues and concerns requiring resolution
- Report and remedies in case of falling behind schedule

Topic 33 Project Communication

Effective and efficient communication is critical to the operation of the project.

The vendor will propose a communication plan to ensure effective communication among key project stakeholders.

Topic 33.1 Ongoing Communication

Communication between the vendor and NH DOE personnel is essential. Telephone calls, telephone conference calls, emails, overnight courier service, facsimile correspondence, and other communication procedures will be at the vendor's expense.

Toll-free numbers shall be provided by the vendor for telephone communication including conference calls and webinars.

The vendor shall make all written communication or summaries of communications with any subcontractor(s) identified in this proposal available to NH DOE at its request. In addition, NH DOE expects to be able to participate during all appropriate and applicable meetings and trainings between the vendor and any subcontractor(s) identified in this proposal.

Topic 33.2 Timeliness of Communication

The Program Manager shall return calls from NH DOE staff and respond to email messages within 24 hours. If the Program Manager is not available to take calls and return messages, NH DOE shall be notified in advance. In the event that the Program Manager is not available, the vendor shall notify NH DOE as to whom to contact in his or her absence, and shall provide contact information for such individual. The vendor shall confirm its agreement to meet this requirement.

The vendor's response should address any technology that will be proposed to support effective communication, any regular written communication or reports that are proposed, and processes and procedures that will be taken to monitor and evaluate the effectiveness of project communication.

Topic 33.3 Monthly Reports

The vendor shall provide a monthly report that summarizes actions taken, issues that arose, issue resolution that occurred, outstanding issues and when they will be resolved, upcoming deadlines, work that will occur in the next month and beyond, and so forth. These reports shall be sent monthly to NH DOE by the third business day of the following month.

Topic 34 Program Improvement Plans

For each phase of the program including development, production, shipping and receipt, administration (of paper- based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities, the vendor shall provide a report that addresses the relevant phase by detailing the activities completed and by providing recommendations for improvement for the next assessment cycle. The report shall also detail errors, problems and/or discrepancies by district and by school. The report will allow NH DOE to detect any patterns in the errors, problems, or discrepancies noted in the report and to use that information to clarify instructions in the Assessment Administration and/or Coordinator Manuals. This report shall be completed within one month of completing the relevant phase.

Topic 35 Risk Management and Quality Assurance

The vendor shall provide a detailed description of the proposed approach to timely identification and effective action on issues and risks.

Vendors shall specifically address timeline issues, risks, and mitigation and contingency plans for all aspects of the project. These plans should include:

- Description of the proposed approach to managing risks and issues.
- A sample tracking document.
- Methodology to ensure that the State staff is involved in the process.
- Description of known risks and proposed steps to mitigate them.

Additional details may be provided in the response to relevant requirements and specifications. The vendor should highlight its and its proposed subcontractors proven ability to document and enact risk management strategies – especially as they relate to the development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, reporting and psychometric activities of high-visibility assessments.

The vendor should submit sample risk assessment documentation used in an existing program to demonstrate the comprehensiveness of its ability to conduct contingency planning for a variety of conditions. This risk assessment documentation may be submitted as an attachment to the proposal. This documentation should also highlight internal procedures and protocols for quality assurance in all aspects of delivering large-scale, statewide assessments – including test development, production, shipping and receipt, administration (of paper-based and online assessments), scanning, scoring (human and artificial intelligence), data processing, and reporting.

D-3 PROJECT EXECUTION

D3.1 Implementation and Operation

The vendor shall provide a detailed description of the roles and responsibilities of vendor staff and State staff during pre-implementation, Implementation, and operational phases of the engagement. The description shall include the amount of time required of each staff member and when their time is needed during the implementation and operational phases of the project.

Topic 36 Implementation Approach

The vendor shall provide one or more feasible implementation plans and user readiness. For each plan provided:

- Identify timeframes for major milestones.
- Discuss cost implications of the plan, including impact on maintenance fees; and
- Address the level of risk associated with the plan.

Topic 37 User Acceptance Testing

The vendor shall provide a detailed description of the support the vendor will supply to assist State during user acceptance testing of the configured System for New Hampshire.

State staff will conduct Acceptance Testing, but support from the selected vendor is required. To define the type of support that will be provided, address the following questions:

- Describe your testing methodology and include a proposed test plan.
- Include the time the State will need to complete User Acceptance Testing of a component.
- Include a description of the support will be provided to prepare State staff during Acceptance testing.
- Include the preparation required for testing the configured Software.
- Include the documentation that will be available to the testing team for the configuration.
- Include any defects likely to be encountered. This information should be based on previous experience and include metrics from other projects to support the response.
- Include time frames for investigation of planned or suspected defects.
- Include time frame for defect correction.
- Provide a sample User Acceptance Test Plan from a completed project as an appendix.

D3.2 Ongoing Operations

Topic 38 Help Desk Support

The vendor shall provide a detailed description of support available to the State to help them with the process of uploading and receiving files and other aspects of data validation and correction. The vendor shall:

- Describe support for the State to assist with the process of uploading files and receiving files. Include hours of operation, response times, problem classification, and escalation procedures.
- Describe your electronic problem tracking process and tools used.
- Describe how user account management will be handled.
- Describe how general support and maintenance skills are transferred to State technical support personnel for knowledge sharing.
- Describe how are support and maintenance issues are tracked detailing methodology and if any additional software is required.
- Describe process for maintenance of the general knowledge base.
- Describe any particular procedures required to handle escalation and emergency calls
- Detail the plan for preventive maintenance and for upgrade installations
- Detail the types and frequency of support tasks required

Topic 39 Support Center

The vendor's response must describe processes and procedures used to ensure timely and accurate assistance; measures used to monitor and document the efficiency and accuracy of the service provided; expected standards for performance and customer service (e.g., wait time, quality of service); and procedures to measure customer satisfaction with the services provided.

The vendor's response should address the processes, procedures, or systems that will be used to ensure that all interactions with districts and schools are documented and maintained in a system that allows for efficient access and review.

The vendor's plan for maintaining a support center must meet the requirements described below.

The vendor will provide for provide customer support to districts and schools throughout the registration, testing, and reporting cycles, with an emphasis on service provided at key periods such as registration of students and test administration.

1. The vendor will guarantee that help desk staffing will increase and/or decrease based on call volume and wait time/caller. When staffing increases/decreases will be determined in consultation with and with approval from the NH DOE management team.

2. The vendor will provide help desk and technical support via toll-free phone, e-mail, and/or other online methods Monday through Friday from 7:00 a.m. EST/EDT through 5:00 p.m. EST/EDT. This includes a dedicated technical support line for NH districts, schools, and state representatives.

3. The vendor will provide tiered levels of customer support to district and school administrators and educators. The vendor and states will agree upon the type of questions and issues that will be addressed by the vendor, what actions the support center and other vendor staff will take to resolve and/or answer those questions and issues, and the type of questions and issues that will be forwarded to the NH DOE for resolution. Support center staff must have the ability to reopen accidentally closed tests.

4. The vendor will ensure that all support center staff and other vendor staff are qualified and have been trained to provide the level of support required by their position.

5. The vendor must develop a Service Level Agreement (SLA) to ensure that the system specifications, performance, and support are appropriate and acceptable. The SLA should have Level 1 (basic level), Level 2 (intermediate level), and Level 3 (technical level) services. The SLA and support processes, shall include at a minimum the following:

    - Availability;
    - Reliability;
    - Latency;
    - Disaster recovery plan;
    - Server backup plan;
    - Recovery point objective;
    - Issue resolution times;

- Maintenance windows;
- Service reporting;
- Support hours;
- Support contact information;
- Escalation;
- Errata notice template suitable for electronic posting and distribution (subject to state approval); and
- Change management.

The vendor must include a plan for timely electronic notification to district and school administrators and test administrators through email, posting a notice on the online system, and/or direct calling, of any issues affecting test administration.

The vendor must develop an errata notice template that includes a description of the issue, the timeline for resolution, and any required actions that need to be taken by district or school administrators and/or test administrators.

Topic 40 Technical Reporting

The vendor will produce and maintain adequate documentation of all technical processes, procedures, and analyses conducted on an ongoing basis throughout the registration, testing, and reporting cycles. One purpose of the documentation will be to enhance quality assurance and quality control. The technical documentation will be produced in a format that is accessible to the NH DOE and conveys useful information to the NH DOE about the technical quality of the assessment program.

Topic 40.1 Technical Report

The vendor will design, develop, and produce an annual Technical Report that documents and provides the necessary evidence to demonstrate the quality of the technical processes and procedures related to the design, development, administration, and reporting of results from the NH assessments. As appropriate, the annual Technical Report must also provide evidence that the planned processes and procedures were implemented for the given year.

The Technical Report is one piece of evidence produced to demonstrate that each of the NH assessments and the assessment program as a whole serve their intended purposes and meet accepted professional standards for educational testing.

The NH DOE will approve the table of contents, design, and format for the Technical Report.

The annual Technical Report will not replace or fulfill the general requirement of ongoing technical documentation of the NH Assessment Program or for task-specific technical documentation specified in this RFP.

A final draft of the annual Technical Report will be delivered to the NH DOE no later than three months following the release of assessment results from operational assessments or three months following the completion of the administration of the Spring 2018 Field Test.

The annual Technical Report will be delivered to the NH DOE in a digital format suitable for posting and distribution through the NH DOE website.

The vendor's response must include a Technical Report that it has prepared for a large-scale state assessment program. If applicable, a link to a publicly available Technical Report can be provided in the vendor's response to fulfill this requirement.

Topic 41 Technical Advisory Committee

The vendor will support two meetings per year of a NH Assessment System Technical Advisory Committee.

The vendor will be represented at the meetings by the project director, lead psychometrician assigned to the project, and additional staff as needed based on the agenda for the meeting.

The NH DOE will select members of the Technical Advisory Committee and will be responsible for facilitating all meetings of the Technical Advisory Committee.

The vendor will be responsible for all activities related to planning for the meeting and for all costs associated with the meeting and activities, including reimbursements and payments made to TAC members.

D-4 PRICING

Topic 42 Pricing Model

The vendor shall provide a detailed description of the *Pricing Model* for the proposed solution that addresses the following components:

- Fixed prices;
- Recurring prices;

- Price per transaction;
- Price per batch;
- Implementation pricing;
- Operations pricing;
- Transition services pricing; and
- Other applicable prices.

The vendor shall include all components found in **Appendix F. and D-5 ASSURANCE AND TRANSITION.**

Topic 43 Quality Control and Sign-Offs

Reviews and signoffs for all deliverables shall be documented and available to NH DOE upon request. The vendor shall document the steps, timeline, and staff involved in the quality control procedures for each phase and deliverable of the project.

Topic 44 Invoices

The vendor shall submit invoices according to the procedures and requirements set forth by NH DOE. It is expected that the payment schedule for this contract will be four quarterly and one final payment for the services performed and deliverables provided during each period. The fiscal year for the State of New Hampshire runs from July 1 to June 30. The last invoice for each fiscal year must be received by June 15. The final invoice for each assessment cycle must be provided by September 1.

Topic 45 Transition

Proposals must include an end of service transition plan detailing the transfer of relevant assessment documents and materials. An organized transition that ensures the continuity of the state assessment program is of the essence. The Transition Plan must address the transfer of materials, both pre-existing and newly developed, from the vendor to NH DOE or another vendor upon termination or expiration of the contract.

The vendor shall assist NH DOE with all activities required to transfer all assessment documents and materials during the transition phase. Draft transition plans shall include procedures for the transition of documents and materials.

The vendor shall ensure that all relevant documents and materials, including but not limited to those identified in the following list are transferred efficiently among NH DOE, the current vendor, and NH DOE's future vendor(s):

- Test development - all critical documents and materials used in the test development process;
- Item and test specifications – all item format details, test map requirements, test blueprints, and technical reports;
- Test books –all paper and electronic test booklets and electronic answer documents from previous test administrations; test maps for each form from the previous year's administration with keys and metadata;
- Passages and artwork – all photocopies of the original passages with source documentation, copies of contracts, original electronic art files and applicable permission information;
- Item bank, item and test statistics – all item-level metadata and previous usage statistics, available test-level statistics, previous anchor range finding papers, rubrics, constructed-response materials such as training material protocols, previous operational and field test usage of each item year and form item position status;
- Program administration - all critical documents and materials used with the test administration process;
- General program documentation –all critical documents and materials used for general program documentation and summary reports;
- Reports –sample copies of all reports provided to districts and schools;
- Manuals/guides –sample copies of all guides and manuals (hard copy and electronic versions) for the operational test administrations, and copies of all electronic materials posted on the state website during the operational test administration;
- Scoring information - all critical documents and materials used in the scoring process;
- Scoring/reporting specifications – all documentation regarding scoring rules, aggregation rules, roll-up algorithms, and tables used to calculate student, school, district, and state results;
- Psychometric and related assessment information required for the program - all critical documents and materials used for psychometric analyses and related procedures;
- Professional development – all critical documents and materials used for professional development;

- Editing Specifications – all documentation that outlines how the state would like answer documents edited during the scanning process
- Equating data files –all documentation that outlines layouts for files including item statistics, master file, pre-id, school/district score data and state-level score data;
- Performance scoring specifications – all training papers, anchor sets, calibration papers, rubrics, and constructed-response scoring rules; previous year's score distributions for each item and historical reader agreement rates;
- Technical reports and other validity and reliability reports - all electronic copies of past technical reports produced by the previous vendor and electronic copies of any other reports that discuss the validity or reliability of the assessments;
- Project plan - all documents that outline the tasks/deliverables and corresponding schedule for those tasks/deliverables;
- Schedules - all previous project schedules containing dates/durations for the following tasks:
- Developing items, forms, and materials
- Enrollment and pre-identification
- Receiving and scanning
- Scoring and reporting
- Packaging specifications - all documentation concerning packaging algorithms and shipping points; and
- Print specifications - all spreadsheets detailing print specifications for test booklets, scannables, answer documents, labels, envelopes, and manuals.

Draft Transition Plans shall include procedures for the transition of documents and materials related to the following:

- Program administration - The vendor shall ensure that all critical documents and materials used with the test administration process are transferred efficiently between NH DOE and/or vendors.
- Test development - The vendor shall ensure that all critical documents and materials used in the test development process are transferred efficiently between NH DOE and/or vendors.

- Scoring information - The vendor shall ensure that all critical documents and materials used in the scoring process are transferred efficiently between NH DOE and/or vendors.
- Psychometric and related assessment information required for the program - The Vendor shall ensure that all critical documents and materials used for psychometric analyses and related procedures are transferred efficiently between NH DOE and/or vendors.
- General program documentation – The vendor shall ensure that all critical documents and materials used for general program documentation and summary reports are transferred efficiently between NH DOE and/or vendors.
- Professional development – The vendor shall ensure that all critical documents and materials used for professional development are transferred efficiently between NH DOE and/or vendors.

The vendor must describe the process for the safe handling of State data during the transition phase.

APPENDIX E: STANDARDS FOR DESCRIBING VENDOR QUALIFICATIONS

Qualifications are important factors in selecting a vendor and accompanying implementation and follow on support services. To facilitate evaluation of vendor qualifications, the State seeks information about:

(1) corporate qualifications of each vendor proposed to participate in the Project;
(2) proposed team organization and designation of key staff;
(3) individual qualifications of candidates for the role of Project Manager; and
(4) individual qualifications of candidates for other key staff roles.

This appendix identifies specific information that must be submitted.

E-1 Required Information on Corporate Qualifications

Information is required on all vendors who will participate in the project. Vendors submitting a proposal must identify any subcontractor(s) to be used.

E-1.1 Vendor and Subcontractors

The vendor submitting a proposal to this project must provide the following information:

E-1.1.1 Corporate Overview

Identify the proposed role of the firm on the project. Describe the major business areas of the firm. Provide a high-level description of the firm's organization and staff size. Discuss the firm's commitment to the public sector, experience with this type of project implementation and experience in New Hampshire.

E-1.1.2 Financial Strength

Provide at least one of the following:

1 The current Dunn & Bradstreet report on the firm; or
2 The firm's two most recent audited financial statements; and The firm's most recent un-audited, quarterly financial statement;
or
3 The firm's most recent income tax return

E-1.1.3 Litigation

Identify and describe any claims made by clients during the last ten (10) years. Discuss merits, current status and, if available, outcome of each matter.

E-1.1.4 Prior Project Descriptions

Provide descriptions of no more than three (3) similar projects completed in the last three (3) years. Each project description should include:

1. An overview of the project covering type of client, objective, project scope, role of the firm and outcome;
2. Project measures including proposed cost, actual project cost, proposed project schedule and actual project schedule;
3. Names and contact information (name, title, address and current telephone number) for one or two references from the client; and
4. Names and project roles of individuals on the proposed team for the New Hampshire project that participated in the project described.

E-1.1.4.1 Components that constitute the vendor's proposed software suite must be fully implemented and operational in at least one (1) government entity comparable in size and complexity to the State of New Hampshire.

E-1.1.5 Subcontractor Information

Vendors must provide information on any subcontractors proposed to work on this project. Required information shall include but not be limited to:

1. Identification of the proposed subcontractor and a description of the major business areas of the firm and their proposed role on the project;
2. A high-level description of the subcontractor's organization and staff size;
3. Discussion of the subcontractor's experience with this type of project;
4. Resumes of key personnel proposed to work on the project; and
5. Two references from companies or organizations where they performed similar services (if requested by the State).

E-2 Team Organization and Designation of Key Vendor Staff

Provide resumes of key personnel proposed to work on the project and an organizational chart depicting the vendor project team. This chart should identify key staff required from the vendor, any subcontractors, and the State.

Define the responsibilities and length of assignment for each of the roles depicted in the organizational chart. Identify the positions that should be designated key staff. Ensure that designation of key vendor staff includes subject matter experts in the following areas:

A single team member may be identified to fulfill the experience requirement in multiple areas.

E-2.1 State Staff Resource Worksheet

Append a completed *State Staff Resource Worksheet* to indicate resources expected of organization. Expected resources must not exceed those outlined in Section A 4.2. The required format follows.

Table E-2: Proposed State Staff Resource Hours Worksheet

State Role	Initiation	Configuration	Implementation	Close Out	Total
Project Manager					
Technology Consultant					
Position 2					
Position 3					
Position 4					
Position 5					
State Total					

E-3 Candidates for Project Manager

Although the State recognizes that staff availability is somewhat uncertain, qualifications of the Project Manager are particularly critical. Therefore, the State requires that the Project Manager be identified with some degree of certainty.

For the Project Manager candidate, provide a resume not to exceed five (5) pages in length addressing the following:

- The candidate’s educational background;
- An overview of the candidate’s work history;
- The candidate’s project experience, including project type, project role and duration of the assignment;
- Any significant certifications held by or honors awarded to the candidate; and
- At least three (3) references, with contact information, that can address the candidate’s performance on past projects.

E-4 Candidates for key vendor staff Roles

Provide a resume not to exceed 2 pages for each key vendor staff position on the project team. Each resume should address the following:

- The individual's educational background;
- An overview of the individual's work history;
- The individual's project experience, including project type, project role and duration of the assignment;
- Any significant certifications held by or honors awarded to the candidate; and
- At least three (3) references, with contact information, that can address the individual's performance on past projects.

APPENDIX F: PRICING WORKSHEETS

A vendor's *Pricing Model* must be based on the worksheets formatted as described in this appendix.

F-1 Activities/Deliverables/Milestones Pricing Worksheet – Deliverables List

The vendor must include, within the not-to-exceed for IT service activities, tasks and preparation of required deliverables, pricing for the deliverables required based on the proposed approach, and methodology and tools. The following format must be used to provide this information.

Table F-1: Activities/Deliverables/Milestones Pricing Worksheet

	Activity, Deliverable, or Milestone	Deliverable Type	Projected Delivery Date	Price
PLANNING AND PROJECT MANAGEMENT				
1	Conduct Project Kickoff Meeting	Non-Software		
2	Project Status Reports	Written		
3	Work Plan	Written		
4	Infrastructure Plan, including Desktop and Network Configuration Requirements	Written		
5	Security Plan	Written		
6	Communications and Change Management Plan	Written		
7	Requirements Trace Ability Matrix	Written		
8	Software Configuration Plan	Written		
9	Systems Interface Plan and Design/Capability	Written		
10	Testing Plan	Written		
11	Data Conversion Plan and Design	Written		
12	Deployment and Roll-out Plan	Written		
13	Comprehensive Training Plan and Curriculum	Written		
14	End User Support Plan	Written		
15	Business Continuity Plan	Written		
16	Documentation of Operational Procedures	Written		
INSTALLATION				

17	Provide Software Licenses, if needed	Written		
18	Provide Fully Tested Data Conversion Software	Software		
19	Provide Software Installed, Configured, and Operational to Satisfy State Requirements	Software		
TESTING				
20	Conduct Integration Testing	Non-Software		
21	Conduct User Acceptance Testing	Non-Software		
22	Perform Production Tests	Non-Software		
23	Test In-Bound and Out-Bound Interfaces	Software		
24	Conduct System Performance (Load/Stress) Testing	Non-Software		
25	Certification of 3rd Party Pen Testing and Application Vulnerability Scanning	Non-Software		
SYSTEM DEPLOYMENT				
26	Converted Data Loaded into Production Environment	*Software*		
27	Provide Tools for Backup and Recovery of all Applications and Data	*Software*		
28	Conduct Training	*Non-Software*		
29	Cutover to New Software	*Non-Software*		
30	NA			
31	Provide Documentation	*Written*		
32	Execute Security Plan	*Non-Software*		
OPERATIONS				
33	Ongoing Hosting Support	Non-Software		
34	Ongoing Support & Maintenance	Software		
35	Conduct Project Exit Meeting	Non-Software		

F-2 Proposed vendor Staff, Resource Hours and Rates Worksheet

Use the proposed vendor staff position, resource hours and rates worksheet to indicate the individuals that will be assigned to the project, hours and applicable rates. Names must be provided for individuals designated for key roles, but titles are sufficient for others. Information is required by phase.

Table F-2: Proposed Vendor Staff, Resource Hours and Rates Worksheet

Title	Name	Initiation	Implementation	Project Close out	Hourly Rate	Hours X Rate
Project Manager						
Technology Consultant						
Position #2						
Position #3						
TOTALS						

F-3 Future Vendor Rates Worksheet

The State may request additional services from the selected vendor and requires rates in the event that additional service is required. The following format must be used to provide this information. "SFY" refers to State Fiscal Year. The New Hampshire fiscal year runs from July 1 through June 30 of the following calendar year. Positions not identified in the *Proposed Position Worksheet* may be included in the future vendor rates worksheet.

Table F-3: Future Vendor Rates Worksheet

Position Title	SFY 201X	SFY 201X	SFY 201X	SFY 201X
Project Manager				
Technology Consultant				
Position #2				
Position #3				

F-4 Software Licensing, Maintenance, and Support Pricing Worksheet

Table F-4: Software Licensing, Maintenance, and Support Pricing Worksheet

Software Name	Initial Cost	Maintenance Support and Upgrades				
		Year 1	Year 2	Year 3	Year 4	Year 5

F-5 Web Site Hosting, Maintenance, and Support Pricing Worksheet

Table F-5: Web Site Hosting, Maintenance, and Support Pricing Worksheet

HOSTED SERVICES	Year 1	Year 2	Year 3	Year 4	Year 5	TOTAL
Web Site Hosting Fee						
Technical Support and updates						
Maintenance and Updates						
GRAND TOTAL						

APPENDIX G-1 SECURITY

Application Security

IT security involves all functions pertaining to the securing of State data and systems through the creation and definition of security policies, procedures and controls covering such areas as identification, authentication and non-repudiation.

This shall include but is not limited to the:

- Development of software applications based on industry best practices and incorporating information security throughout the software development life cycle.
- Following change control process and procedures relative to release of code.
- Developing applications following security-coding guidelines as set forth by organizations such as, but not limited to Open Web Application Security Project (OWASP) Top 10, SANS Common Weakness Enumeration (CWE) Top 25 or CERT Secure Coding.
- Making available for review and audit purposes all software development processes and certification for application developers on secure coding techniques.

APPENDIX G-2 TESTING REQUIREMENTS

All testing and acceptance addressed herein shall apply to testing the system. This shall include planning, test scenario development, data, and system preparation for testing, and execution of unit testing, system integration testing, conversion/migration testing, installation testing, performance, and stress testing, security review and testing, and support of the State during User Acceptance Testing (UAT).

G-2.1 Test Planning and Preparation

The overall test plan will guide all testing. The vendor provided, State approved, test plan will include, at a minimum, identification, preparation, and documentation of planned testing, a requirements traceability matrix, test variants, test scenarios, test cases, test scripts, test data, test phases, unit tests, expected results, and a tracking method for reporting actual versus expected results as well as all errors and problems identified during test execution.

It is crucial that client training and testing activities not be abbreviated in order to meet project implementation schedules. Therefore, the State requires that the testing activities be represented both in terms of effort and duration.

Vendors must disclose in their proposals the scheduling assumptions used in regard to the client resource efforts during testing.

State testing will commence upon the vendor Project Manager’s certification, in writing, that the vendor’s own staff has executed all prerequisite vendor testing, along with reporting the actual testing results, prior to the start of any testing executed by State staff.

The State will commence its testing within five (5) business days of receiving certification from the vendor that the State’s personnel have been trained and the system is installed, configured, complete, and ready for State testing. The testing will be conducted by the State in an environment independent from the vendor’s development environment. The vendor must assist the State with testing in accordance with the test plan and the work plan, utilizing test and live data to validate reports, and conduct stress and performance testing, at no additional cost.

G-2.2 Testing

Testing begins upon completion of the software configuration as required and user training according to the work plan. Testing ends upon issuance of a letter of UAT acceptance by the State.

Vendor must demonstrate that their testing methodology can be integrated with the State standard methodology.

Unit Testing	Application components are tested on an individual basis to verify that the inputs, outputs, and processing logic of each application component functions without errors. Unit testing is performed in either the development environment or a testing environment. The goal is to find errors in the smallest unit of software if subsequent integration testing should only reveal errors related to the integration between application components.
System Integration Testing	a.) Validates the integration between the individual unit application components and verifies that the new System meets defined requirements and supports execution of interfaces and business processes. The Systems Integration Test is performed in a test environment. b.) Emphasizes end-to-end business processes and the flow of information across applications. It includes all key business processes and interfaces' being implemented, confirms data transfers with external parties, and includes the transmission or printing of all electronic and paper documents. c.) The State will conduct System Integration Testing, utilizing scripts developed, as identified in the Test Plan, to validate the functionality of the System and its interfaces. The State will also use System Integration Testing to validate modifications, fixes and other System interactions with the vendor supplied software solution.
Conversion /Migration Validation Testing	The Conversion/Migration Validation Testing should replicate the entire flow of the converted data through the Software Solution. As the Software Solution is interfaced to legacy or third-party applications, the testing verifies that the resulting converted legacy data performs correctly.
Installation Testing	Application components are installed in the system test environment to test the installation routines and are refined for the eventual production environment. This activity serves as a dry run of the installation steps in preparation for configuring the production system.
User Acceptance Testing (UAT)	The User Acceptance Test (UAT) is a verification process performed in a copy of the production environment. The UAT verifies system functionality against predefined acceptance criteria that support the execution of approved business processes. a.) The vendor's Project Manager must certify in writing, that the vendor's own staff has executed all prerequisite vendor testing, along with reporting the actual testing results prior to the start of any testing executed by State staff. b.) The State will be presented with a State approved test plan, test scenarios, test cases, test scripts, test data, and expected results, as well as written certification of the vendor's having completed the prerequisite tests, prior to the State staff involvement in any testing activities

	c.) UAT will also serve as a performance and stress test of the System. It may cover any aspect of the new System, including administrative procedures such as backup and recovery. The results of the UAT provide evidence that the new system meets the user acceptance criteria as defined in the work plan. d.) Upon conclusion of UAT and system deployment, the State will issue a letter of UAT acceptance and the respective warranty period shall commence as described in **Section H-25.10.1: Warranty Period**.
Performance Tuning and Stress Testing	Vendor shall develop and document hardware and software configuration and tuning of system infrastructure as well as assist and direct the State's system administrators and database administrators in configuring and tuning the infrastructure to support the software throughout the project **Performance Tuning and Stress Testing** **<u>Scope</u>** The scope of performance testing shall measure the system level metrics critical for the development of the applications infrastructure and operation of the applications in the production environment. It will include the measurement of response rates of the application for end-user transactions and resource utilization (of various servers and network) under various load conditions. These response rates shall become the basis for changes and retesting until optimum system performance is achieved. The application transactions shall be identified with specific roles and selected transactions shall be recorded for the performance measurements. These will be compared to baselines to determine if object and/or system performance increases as changes are made. Performance testing shall consider the full scope of the application infrastructure with emphasis on the most heavily used or shared transactions. Performance testing of the application will profile the identified user transactions and assist in the identifying performance gaps to improve the most critical parts of the applications. Performance testing and tuning shall occur in the final production environment and shall use a copy of the final production database to provide the best results. Vendor must lead this effort. Responsibilities include identifying appropriate tunable parameters and their default and recommended settings, developing scripts, which accurately reflect business load and coordinating reporting of results.

	<u>Test types</u> Performance testing shall use two different types of tests to determine the stability of the application. They are baseline tests and load tests **<u>Baseline Tests</u>** Baseline tests shall collect performance data and load analysis by running scripts where the output is broken down into business transactions or functions. The test is like a single user executing a defined business transaction. During baseline testing, each individual script is run to establish a baseline for transaction response time, throughput and other user-based metrics. Usually each business transaction is executed multiple times during a single test run to obtain an average for the user-based metrics required for the performance testing evaluations. It must be noted that changes made to the code after baseline testing is completed will skew the results collected to date. All effort will be made to provide a code test base that is tested in the environment for problems prior to the establishment of the baseline, which are used in future testing and tuning efforts. Any changes introduced into the environment after performance testing has started can compromise the accuracy of the results and will force a decision to be made whether baseline results need to be recreated. **<u>Load Tests</u>** Load testing will determine if the behavior of a system can be sustained over a long period of time while running under expected conditions. Load tests helps to verify the ability of the application environment under different load conditions based on workload distribution. System response time and utilization is measured and recorded. **<u>Tuning</u>** Tuning will occur during both the development of the application and load testing. Tuning is the process whereby the application performance is maximized. This can be the result of making code more efficient during development as well as making tuning parameter changes to the environment. For infrastructure tuning, parameters will be identified for all components prior to undertaking the load testing efforts. This should include a list of the variables, their definitions, the default settings, range of acceptable settings and the settings as testing begins. This will permit the team to identify the areas of most potential gain and a starting point. Tuning is a process which is repeated until the team feels that the systems are running at or near optimum performance.

	<u>Implementing Performance and Stress Test</u> Performance and stress test tools must be provided by the vendor for this effort. Consideration must be given to licensing with respect to continued use for regression testing. If the vendor is familiar with open source low/no cost tools for this purpose those tools should be identified in your response. **<u>Scheduling Performance and Stress Testing</u>** Vendor shall perform test planning. The steps for planning include identification of application functionality as well as what percentage of normal daily use is represented by each function. This information will become the foundation for scripting so that tests closely represent what loads in production will look like. Vendor shall provide definition and expectations from testing. This definition should include who is in charge of testing and coordinating results, anticipated run times, logs required for tracking, their locations and which technician is responsible to track and provide them following each test to the team. Initial test runs shall be completed to establish that the tests and data sets can be run to completion without errors. The ratio of types of transactions which makeup the test shall be reviewed prior to the beginning of testing and then again once testing has begun to make sure that testing accurately reflects the system performing in production. Initial tests shall be used to establish a baseline from which all subsequent tests will be compared. Tests will be considered for baseline status once two of them have been run within 2 percent of each other in key and overall performance areas. No changes to the test scripts or data sets (with the exception of restores after each test) can be done to the test environment once tuning has begun so as to not damage the comparison to baseline results. The systems must be restarted prior to each test run to assure all cache is cleaned out. All effort will be made to run these tests at a time when system and network infrastructure utilization doesn't impact the results. Tests will be run in close proximity to our infrastructure to eliminate the public network from our environment. Post-test reporting and result assessment will be scheduled following each test. The team will compare these results to the baseline and a

	determination must be made to make additional changes to the parameter being tuned or return to the prior configuration and select another parameter to tune while keeping in mind that significant changes to any one parameter may require the retesting of some others. Careful work on identifying dependencies up front should minimize this impact. If defects are identified in the application during testing, they will be recorded; however, changes to the application code should be avoided if possible so as not to affect baseline comparisons. If a change to the application is required new baselines will be established (and possibly the execution of prior tests to validate changes with the new application) before testing can continue. When performing capacity testing against a GUI the focus will be on the ability of the interface to respond to user input. During stress/load testing the tester will attempt to stress or load an aspect of the system to the point of failure. The goal being to determine weak points in the system architecture. The tester will identify peak load conditions at which the program will fail to handle required processing loads within required time spans. During performance testing the tester will design test case scenarios to determine if the system meets the stated performance criteria (i.e. A Login request shall be responded to in one (1) second or less under a typical daily load of 1000 requests per minute.). In both cases, the tester will determine the capacity of the system under a known set of conditions.
Regression Testing	As a result, of the user testing activities, problems will be identified that require correction. The State will notify the vendor of the nature of the testing failures in writing. The vendor will be required to perform additional testing activities in response to State and/or user problems identified from the testing results. Regression testing means selective re-testing to detect faults introduced during the modification effort, both to verify that the modifications have not caused unintended adverse effects, and to verify that the modified and related (possibly affected) system components still meet their specified requirements. a.) For each minor failure of an acceptance test, the acceptance period shall be extended by corresponding time defined in the test plan.

	b.) The vendor shall notify the State no later than five (5) business days from the vendor's receipt of written notice of the test failure when the vendor expects the corrections to be completed and ready for retesting by the State. The vendor will have up to five (5) business days to make corrections to the problem unless specifically extended in writing by the State. c.) When a programming change is made in response to a problem identified during user testing, a regression test plan should be developed by the vendor based on the understanding of the program and the change being made to the program. The test plan has two objectives: **1.** validate that the change/update has been properly incorporated into the program; and **2.** validate that there has been no unintended change to the other portions of the program. d.) The vendor will be expected to: 1. Create a set of test conditions, test cases, and test data that will validate that the change has been incorporated correctly; 2. Create a set of test conditions, test cases, and test data that will validate that the unchanged portions of the program still operate correctly; and 3. Manage the entire cyclic process. e.) The vendor will be expected to execute the regression test, provide actual testing results, and certify its completion in writing to the State prior to passing the modified software application to the users for retesting. In designing and conducting such regression testing, the vendor will be required to assess the risks inherent to the modification being implemented and weigh those risks against the time and effort required for conducting the regression tests. In other words, the vendor will be expected to design and conduct regression tests that will identify any unintended consequences of the modification while taking into account Schedule and economic considerations. In their proposals vendors must acknowledge their responsibilities for regressio testing as described in this section.
Security Review and Testing	IT security involves all functions pertaining to the securing of State data and systems through the creation and definition of security policies, procedures and controls covering such areas as identification, authentication and non-repudiation. All components of the software shall be reviewed and tested to ensure they protect the State's hardware and software and its related data assets.

Service Component	Defines the set of capabilities that:
Identification and Authentication	Supports obtaining information about those parties attempting to log onto a system or application for security purposes and the validation of users
Access Control	Supports the management of permissions for logging onto a computer or network
Encryption	Supports the encoding of data for security purposes
Intrusion Detection	Supports the detection of illegal entrance into a computer system
Verification	Supports the confirmation of authority to enter a computer system, application or network
Digital Signature	Guarantees the unaltered state of a file
User Management	Supports the administration of computer, application and network accounts within an organization
Role/Privilege Management	Supports the granting of abilities to users or groups of users of a computer, application or network
Audit Trail Capture and Analysis	Supports the identification and monitoring of activities within an application or system
Input Validation	Ensures the application is protected from buffer overflow, cross-site scripting, SQL injection, and unauthorized access of files and/or directories on the server

In their proposal, the vendors must acknowledge their responsibilities for security testing. Tests shall focus on the technical, administrative and physical security controls that have been designed into the system architecture in order to provide the necessary confidentiality, integrity and availability. Tests shall, at a minimum, cover each of the service components. Test procedures shall include 3rd party penetration tests (pen test) or code analysis and review.

Prior to the system being moved into production, the vendor shall provide results of all security testing to the Department of Information Technology for review and acceptance. All software and hardware shall be free of malicious code (malware).

		Penetration Testing shall include: **(delete requirements that are specific to PCI testing)** **11.3** Implement a methodology for penetration testing that includes the following: · Is based on industry-accepted penetration testing approaches (for example, NIST SP800-115) · Includes coverage for the entire NH DOE perimeter and critical systems · Includes testing from both inside and outside the network · Includes testing to validate any segmentation and scope-reduction controls · Defines application-layer penetration tests to include, at a minimum, the vulnerabilities listed in Requirement 6.5 · Defines network-layer penetration tests to include components that support network functions as well as operating systems · Includes review and consideration of threats and vulnerabilities experienced in the last 12 months · Specifies retention of penetration testing results and remediation activities results.
		11.3.1 Perform *external* penetration testing at least annually and after any significant infrastructure or application upgrade or modification (such as an operating system upgrade, a sub-network added to the environment, or a web server added to the environment).
		11.3.2 Perform *internal* penetration testing at least annually and after any significant infrastructure or application upgrade or modification (such as an operating system upgrade, a sub-network added to the environment, or a web server added to the environment).
		11.3.3 Exploitable vulnerabilities found during penetration testing are corrected and testing is repeated to verify the corrections.
		11.3.4 If segmentation is used to isolate the NH DOE from other networks, perform penetration tests at least annually and after any changes to segmentation controls/methods to verify that the segmentation methods are operational and effective, and isolate all out-of-scope systems from in-scope systems.

APPENDIX G-3: CERTIFICATES

A. Certificate of Good Standing

As a condition of Contract award, the vendor, if required by law, must furnish a Certificate of Authority/Good Standing dated after April 1, 2017, from the Office of the Secretary of State of New Hampshire. If your company is not registered, an application form may be obtained from:

Secretary of State
State House Annex
25 Capitol Street
Concord, New Hampshire 03301
603-271-3244

If your company is registered, a certification thereof may be obtained from the Secretary of State.

Note: Sovereign states or their agencies may be required to submit suitable substitute documentation concerning their existence and authority to enter into a contract.

B. Certificate of Authority/Vote

The Certificate of Authority/Vote authorizes, by position, a representative(s) of your corporation to enter into an agreement or amendment with the State of New Hampshire. This ensures that the person signing the agreement is authorized as of the date he or she is signing it to enter into agreements for that organization with the State of New Hampshire.

The officer's signature must be either notarized or include a corporate seal that confirms the title of the person authorized to sign the agreement. The date the Board officer signs must be on or after the date the amendment is signed. The date the notary signs must match the date the Board officer signs.

You may use your own format for the Certificate of Authority/Vote as long as it contains the necessary language to authorize the agreement signatory to enter into agreements and amendments with the State of New Hampshire as of the date they sign.

CERTIFICATE OF AUTHORITY/VOTE CHECKLIST

SOURCE OF AUTHORITY

Authority must come from the **governing body,** either:

(1) a **majority voted** at a meeting, or
(2) the body provided **unanimous consent in writing**, or
(3) the organization's **policy or governing document** (bylaws, partnership agreement, LLC operating agreement) authorizes the person to sign

SOURCE OF AUTHORITY WAS IN EFFECT ON DAY AGREEMENT OR AMENDMENT WAS SIGNED

Certificate must show that the person signing the contract **had authority when they signed the Agreement or Amendment,** either:

(1) Authority was **granted the same day** as the day the Agreement or Amendment was signed, or
(2) Authority was **granted after** the day the agreement or amendment was signed and the governing body ratifies and accepts the earlier execution, or
(3) Authority was **granted prior** to the day the agreement or amendment was signed and it has not been amended or repealed as of the day the contract was signed.

APPROPRIATE PERSON SIGNED THE CERTIFICATE

The person signing the certificate may be the same person signing the agreement or amendment only if the certificate states that the person is the **sole director** (for corps) or **sole member** (for LLCs).

APPENDIX H – STATE OF NEW HAMPSHIRE TERMS AND CONDITIONS

AGREEMENT

The State of New Hampshire and the vendor hereby mutually agree as follows:

GENERAL PROVISIONS

1. IDENTIFICATION.

1.1 State Agency Name New Hampshire Department of Education		1.2 State Agency Address 101 Pleasant Street Concord, NH 03301	
1.3 Vendor Name		1.4 Vendor Address	
1.5 Vendor Phone Number	1.6 Account Number	1.7 Completion Date	1.8 Price Limitation
1.9 Contracting Officer for State Agency Saundra MacDonald		1.10 State Agency Telephone Number 603.271.3543	
1.11 Vendor Signature		1.12 Name and Title of Vendor Signatory	
1.13 Acknowledgement: State of , County of On , before the undersigned officer, personally appeared the person identified in block 1.12, or satisfactorily proven to be the person whose name is signed in block 1.11, and acknowledged that s/he executed this document in the capacity indicated in block 1.12.			
1.13.1 Signature of Notary Public or Justice of the Peace [Seal]			
1.13.2 Name and Title of Notary or Justice of the Peace			
1.14 State Agency Signature Date:		1.15 Name and Title of State Agency	
1.16 Approval by the N.H. Department of Administration, Division of Personnel (if applicable) By: Director, On:			
1.17 Approval by the Attorney General (Form, Substance and Execution) (*if applicable)* By: On:			
1.18 Approval by the Governor and Executive Council *(if applicable)* By: On:			

2. EMPLOYMENT OVENDOR/SERVICES TO BE PERFORMED. The State of New Hampshire, acting through the agency identified in block 1.1 ("State"), engages vendor identified in block 1.3 ("Vendor") to perform, and the Vendor shall perform, the work or sale of goods, or both, identified and more particularly described in the attached EXHIBIT A which is incorporated herein by reference ("Services").

3. EFFECTIVE DATE/COMPLETION OF SERVICES.
3.1 Notwithstanding any provision of this Agreement to the contrary, and subject to the approval of the Governor and Executive Council of the State of New Hampshire, if applicable, this Agreement, and all obligations of the parties hereunder, shall become effective on the date the Governor and Executive Council approve this Agreement as indicated in block 1.18, unless no such approval is required, in which case the Agreement shall become effective on the date the Agreement is signed by the State Agency as shown in block 1.14 ("Effective Date").
3.2 If the Vendor commences the Services prior to the Effective Date, all Services performed by the Vendor prior to the Effective Date shall be performed at the sole risk of the Vendor, and in the event that this Agreement does not become effective, the State shall have no liability to the Vendor, including without limitation, any obligation to pay the Vendor for any prices incurred or Services performed. Vendor must complete all Services by the Completion Date specified in block 1.7.

4. CONDITIONAL NATURE OF AGREEMENT.
Notwithstanding any provision of this Agreement to the contrary, all obligations of the State hereunder, including, without limitation, the continuance of payments hereunder, are contingent upon the availability and continued appropriation of funds, and in no event shall the State be liable for any payments hereunder in excess of such available appropriated funds. In the event of a reduction or termination of appropriated funds, the State shall have the right to withhold payment until such funds become available, if ever, and shall have the right to terminate this Agreement immediately upon giving the Vendor notice of such termination. The State shall not be required to transfer funds from any other account to the Account identified in block 1.6 in the event funds in that Account are reduced or unavailable.

5. CONTRACT PRICE/PRICE LIMITATION/ PAYMENT.
5.1 The contract price, method of payment, and terms of payment are identified and more particularly described in EXHIBIT B which is incorporated herein by reference.
5.2 The payment by the State of the contract price shall be the only and the complete reimbursement to the Vendor for all expenses, of whatever nature incurred by the Vendor in the performance hereof, and shall be the only and the complete compensation to the Vendor for the Services. The State shall have no liability to the Vendor other than the contract price.
5.3 The State reserves the right to offset from any amounts otherwise payable to the Vendor under this Agreement those liquidated amounts required or permitted by N.H. RSA 80:7 through RSA 80:7-c or any other provision of law.
5.4 Notwithstanding any provision in this Agreement to the contrary, and notwithstanding unexpected circumstances, in no event shall the total of all payments authorized, or actually made hereunder, exceed the Price Limitation set forth in block 1.8.

6. COMPLIANCE BY VENDOR WITH LAWS AND REGULATIONS/ EQUAL EMPLOYMENT OPPORTUNITY.
6.1 In connection with the performance of the Services, the Vendor shall comply with all statutes, laws, regulations, and orders of federal, state, county or municipal authorities which impose any obligation or duty upon the Vendor, including, but not limited to, civil rights and equal opportunity laws. This may include the requirement to utilize auxiliary aids and services to ensure that persons with communication disabilities, including vision, hearing and speech, can communicate with, receive information from, and convey information to the Vendor. In addition, the Vendor shall comply with all applicable copyright laws.
6.2 During the term of this Agreement, the Vendor shall not discriminate against employees or applicants for employment because of race, color, religion, creed, age, sex, handicap, sexual orientation, or national origin and will take affirmative action to prevent such discrimination.
6.3 If this Agreement is funded in any part by monies of the United States, the Vendor shall comply with all the provisions of Executive Order No. 11246 ("Equal Employment Opportunity"), as supplemented by the regulations of the United States Department of Labor (41 C.F.R. Part 60), and with any rules, regulations and guidelines as the State of New Hampshire or the United States issue to implement these regulations. The Vendor further agrees to permit the State or United States access to any of the Vendor's books, records and accounts for the purpose of ascertaining compliance with all rules, regulations and orders, and the covenants, terms and conditions of this Agreement.

7. PERSONNEL.
7.1The Vendor shall at its own expense provide all

personnel necessary to perform the Services. The Vendor warrants that all personnel engaged in the Services shall be qualified to perform the Services, and shall be properly licensed and otherwise authorized to do so under all applicable laws.
7.2 Unless otherwise authorized in writing, during the term of this Agreement, and for a period of six (6) months after the Completion Date in block 1.7, the Vendor shall not hire, and shall not permit any subcontractor or other person, firm or corporation with whom it is engaged in a combined effort to perform the Services to hire, any person who is a State employee or official, who is materially involved in the procurement, administration or performance of this Agreement. This provision shall survive termination of this Agreement.
7.3 The Contracting Officer specified in block 1.9, or his or her successor, shall be the State's representative. In the event of any dispute concerning the interpretation of this Agreement, the Contracting Officer's decision shall be final for the State.

8. EVENT OF DEFAULT/REMEDIES.
8.1 Any one or more of the following acts or omissions of the Vendor shall constitute an event of default hereunder ("Event of Default"):
8.1.1 failure to perform the Services satisfactorily or on schedule;
8.1.2 failure to submit any report required hereunder; and/or
8.1.3 failure to perform any other covenant, term or condition of this Agreement.
8.2. Upon the occurrence of any Event of Default, the State may take any one, or more, or all, of the following actions:
8.2.1 give the Vendor a written notice specifying the Event of Default and requiring it to be remedied within, in the absence of a greater or lesser specification of time, thirty (30) days from the date of the notice; and if the Event of Default is not timely remedied, terminate this Agreement, effective two(2) days after giving the Vendor notice of termination;
8.2.2. give the Vendor a written notice specifying the Event of Default and suspending all payments to be made under this Agreement and ordering that the portion of the contract price which would otherwise accrue to the Vendor during the period from the date of such notice until such time as the State determines that the Vendor has cured the Event of Default shall never be paid to the Vendor;
8.2.3. set off against any other obligations the State may owe to the Vendor any damages the State suffers by reason of any Event of Default; and/or
8.2.4. treat the Agreement as breached and pursue any of its remedies at law or in equity, or both.

9. DATA/ACCESS/CONFIDENTIATLITY/RESERVATION.
9.1 As used in this Agreement, the word "data" shall mean al information and things developed or obtained during the performance of, or acquired or developed by reason of, this Agreement, including, but not limited to, all studies, reports, files, formulae, surveys, maps, charts, sound recordings, video recordings, pictorial reproductions, drawings, analyses, graphic representations, computer programs, computer printouts, notes, letters, memoranda, papers, and documents, all whether finished or unfinished.
9.2 All data and any property which has been received from the State or purchased with funds provided for that purpose under this Agreement, shall be the property of the State, and shall be returned to the State upon demand or upon termination of this Agreement for any reason.
9.3 Confidentiality of data shall be governed by N.H. RSA chapter 91-A or other existing law. Disclosure of data requires prior written approval of the State.

10. TERMINATION. In the event of an early termination of this Agreement for any reason other than the completion of the Services, the Vendor shall deliver to the Contracting Officer, not later than fifteen (15) days after the date of termination, a report ("Termination Report") describing in detail all Services performed, and the contract price earned, to and including the date of termination. The form, subject matter, content, and number of copies of the Termination Report shall be identical to those of any Final Report described in the attached EXHIBIT A.

11. VENDOR'S RELATION TO THE STATE. In the performance of this Agreement the Vendor is in all respects an independent vendor, and is neither an agent nor an employee of the State. Neither the Vendor nor any of its officers, employees, agents or members shall have authority to bind the State or receive any benefits, workers' compensation or other emoluments provided by the State to its employees.

12.ASSIGNMENT/DELEGATION /SUBCONTRACTS.
The Vendor shall not assign, or otherwise transfer any interest in this Agreement without the prior written notice and consent of the State. None of the Services shall be subcontracted by the Vendor without the prior written notice and consent of the State.

13. INDEMNIFICATION. The Vendor shall defend, indemnify and hold harmless the State, its officers and employees, from and against any and all losses suffered by the State, its officers and employees, and any and all claims, liabilities or penalties asserted against the State, its officers and employees, by or on behalf of any person, on account of, based or resulting from, arising out of (or which may be claimed to arise out of) the acts or omissions of the Vendor. Notwithstanding the foregoing, nothing herein contained shall be deemed to constitute a waiver of the sovereign immunity of the State, which immunity is hereby reserved to the State. This covenant in paragraph 13 shall survive the termination of this Agreement.

14. INSURANCE.
14.1 The Vendor shall, at its sole expense, obtain and maintain in force, and shall require any subcontractor or assignee to obtain and maintain in force, the following insurance:
14.1.1 comprehensive general liability insurance against all claims of bodily injury, death or property damage, in amounts of not less than $1,000,000per occurrence and $2,000,000 aggregate ; and
14.1.2 special cause of loss coverage form covering all property subject to subparagraph 9.2 herein, in an amount not less than 80% of the whole replacement value of the property.
14.2 The policies described in subparagraph 14.1 herein shall be on policy forms and endorsements approved for use in the State of New Hampshire by the N.H. Department of Insurance, and issued by insurers licensed in the State of New Hampshire.
14.3 The Vendor shall furnish to the Contracting Officer identified in block 1.9, or his or her successor, a certificate(s) of insurance for all insurance required under this Agreement. Vendor shall also furnish to the Contracting Officer identified in block 1.9, or his or her successor, certificate(s) of insurance for all renewal(s) of insurance required under this Agreement no later than thirty (30) days prior to the expiration date of each of the insurance policies. The certificate(s) of insurance and any renewals thereof shall be attached and are incorporated herein by reference. Each certificate(s) of insurance shall contain a clause requiring the insurer to provide the Contracting Officer identified in block 1.9, or his or her successor, no less than thirty (30) days prior written notice of cancellation or modification of the policy.

15. WORKERS' COMPENSATION.
15.1 By signing this agreement, the Vendor agrees, certifies and warrants that the Vendor is in compliance with or exempt from, the requirements of N.H. RSA chapter 281-A *("Workers' Compensation")*.
15.2 To the extent the Vendor is subject to the requirements of N.H. RSA chapter 281-A, Vendor shall maintain, and require any subcontractor or assignee to secure and maintain, payment of Workers' Compensation in connection with activities which the person proposes to undertake pursuant to this Agreement. Vendor shall furnish the Contracting Officer identified in block 1.9, or his or her successor, proof of Workers' Compensation in the manner described in N.H. RSA chapter 281-A and any applicable renewal(s) thereof, which shall be attached and are incorporated herein by reference. The State shall not be responsible for payment of any Workers' Compensation premiums or for any other claim or benefit for Vendor, or any subcontractor or employee of Vendor, which might arise under applicable State of New Hampshire Workers' Compensation laws in connection with the performance of the Services under this Agreement.

16. WAIVER OF BREACH. No failure by the State to enforce any provisions hereof after any Event of Default shall be deemed a waiver of its rights with regard to that Event of Default, or any subsequent Event of Default. No express failure to enforce any Event of Default shall be deemed a waiver of the right of the State to enforce each and all of the provisions hereof upon any further or other Event of Default on the part of the Vendor.

17. NOTICE. Any notice by a party hereto to the other party shall be deemed to have been duly delivered or given at the time of mailing by certified mail, postage prepaid, in a United States Post Office addressed to the parties at the addresses given in blocks 1.2 and 1.4, herein.

18. AMENDMENT. This Agreement may be amended, waived or discharged only by an instrument in writing signed by the parties hereto and only after approval of such amendment, waiver or discharge by the Governor and Executive Council of the State of New Hampshire unless no such approval is required under the circumstances pursuant to State law, rule or policy.

19. CONSTRUCTION OF AGREEMENT AND TERMS. This Agreement shall be construed in accordance with the laws of the State of New Hampshire, and is binding upon and inures to the benefit of the parties and their respective successors and assigns. The wording used in this Agreement is the wording chosen by the parties to express their mutual intent, and no rule of construction shall be applied against or in favor of any party.

20. THIRD PARTIES. The parties hereto do not intend to benefit any third parties and this Agreement shall not be construed to confer any such benefit.

21. HEADINGS. The headings throughout the Agreement are for reference purposes only, and the words contained therein shall in no way be held to explain, modify, amplify or aid in the interpretation, construction or meaning of the provisions of this Agreement.

22. SPECIAL PROVISIONS. Additional provisions set forth in the attached EXHIBIT C are incorporated herein by reference.

23. SEVERABILITY. In the event any of the provisions of this Agreement are held by a court of competent jurisdiction to be contrary to any state or federal law, the remaining provisions of this Agreement will remain in full force and effect.

24. ENTIRE AGREEMENT. This Agreement, which may be executed in a number of counterparts, each of which shall be deemed an original, constitutes the entire Agreement and understanding between the parties, and supersedes all prior Agreements and understandings relating hereto.

.

H-25 GENERAL CONTRACT REQUIREMENTS

H-25.1 State of NH Terms and Conditions and Contract Requirements

The contract terms set forth in **Appendix H: State of New Hampshire Terms and Conditions** shall constitute the core for any contract resulting from this RFP.

H-25.2 Vendor Responsibilities

The vendor shall be solely responsible for meeting all requirements, and terms and conditions specified in this RFP, its proposal, and any resulting contract, regardless of whether or not it proposes to use any subcontractor.

The vendor may subcontract services subject to the provisions of the RFP, including but not limited to, the terms and conditions in **Appendix H: *State of New Hampshire Terms and Conditions*.** The vendor must submit with its proposal all information and documentation relating to the Subcontractor necessary to fully respond to the RFP, which must include terms and conditions consistent with this RFP. The vendor shall remain wholly responsible for performance of the entire contract regardless of whether a subcontractor is used. The State will consider the vendor to be the sole point of contact with regard to all contractual matters, including payment of any and all charges resulting from any contract.

H-25.3 Project Budget/Price Limitation

The State has funds budgeted for this project, subject to **Appendix H: *State of New Hampshire Terms and Conditions*, Section 4: *Conditional Nature of Agreement* and Section 5: *Contract Price/Price Limitation/Payment***.

H-25.4 State Contracts

The State of New Hampshire intends to use, wherever possible, existing statewide software and hardware contracts to acquire supporting software and hardware.

H-25.5 Vendor Staff

In the proposal, the vendor shall assign and identify a Project Manager and key vendor staff, in accordance with the requirements and deliverables of **Appendix C: *System Requirements and Deliverables*** and **Appendix E: *Standards for Describing Vendor Qualifications.***

The vendor’s selection of a Project Manager will be subject to the prior approval of the State. The State’s approval process may include, without limitation, at the State’s discretion, review of the proposed Project Manager’s resume, qualifications, references and background checks, and an interview. The vendor’s Project Manager must be qualified to perform the obligations required of the position under the contract, have full authority to make binding decisions, and shall function as the vendor’s representative for all administrative and management matters. The Project

Manager must be available to promptly respond during normal working hours within two (2) hours to inquiries from the State, and be at the site as needed. The vendor must use his or her best efforts on the project.

The vendor shall not change key vendor staff and Project Manager commitments (collectively referred to as "project staff") unless such replacement is necessary due to sickness, death, termination of employment, or unpaid leave of absence. Any such changes to the vendor's project staff shall require the prior written approval of the State. Replacement project staff shall have comparable or greater skills with regard to performance of the project as the staff being replaced and be subject to the provisions of this RFP and any resulting contract.

The State, at its sole expense, may conduct reference and background checks on the vendor's project staff. The State shall maintain the confidentiality of reference and background screening results. The State reserves the right to reject the vendor's Project Staff as a result of such reference and background checks. The State also reserves the right to require removal or reassignment of the vendor's key project staff found unacceptable to the State.

Notwithstanding anything to the contrary, the State shall have the option to terminate the contract, at its discretion, if it is dissatisfied with the vendor's replacement project staff.

H-25.6 Work Plan

Vendor shall submit a preliminary work plan in its proposal. The work plan shall include, without limitation, a detailed description of the schedule, tasks, deliverables, major milestones, task dependencies, and payment schedule. A final work plan will be due five (5) business days after contract award upon approval by Governor and Executive Council.

The vendor shall update the work plan as necessary, but no less than every two weeks to accurately reflect the status of the project, including without limitation, the schedule, tasks, deliverables, major milestones, task dependencies, and payment schedule. Any updates to the work plan shall require the written approval of the State prior to final incorporation into the contract.

Unless otherwise agreed in writing by the State, changes to the work plan shall not relieve the vendor from liability to the State for any damages resulting from the vendor's failure to perform its obligations under the contract, including without limitation, performance in accordance with the schedule.

In the event of a delay in the schedule, the vendor must immediately notify the State in writing. The written notification will identify the nature of the delay, i.e., specific actions or inactions of the vendor or State causing the problem; its estimated duration period to reconciliation; specific actions that need to be taken to correct the problem; and the expected Schedule affect the Project.

In the event the vendor requires additional time to correct deficiencies, the schedule shall not change unless previously agreed in writing by the State, except that the schedule shall automatically extend on a day-to-day basis to the extent that the delay does not result from vendor's failure to fulfill its obligations under the contract. To the extent that the State's execution of its major tasks takes longer than described in the work plan, the schedule shall automatically extend on a day-to-day basis.

Notwithstanding anything to the contrary, the State shall have the option to terminate the contract for default, at its discretion, if it is dissatisfied with the vendor's work plan or elements within the work plan.

H-25.7 Change Orders

The State may make changes or revisions at any time by written change order. Within five (5) business days of a vendor's receipt of a change order, the vendor shall advise the State, in detail, of any impact on cost (e.g., increase or decrease), the schedule, or the work plan.

A vendor may request a change within the scope of the contract by written change order, identifying any impact on cost, the schedule, or the work plan. The State shall attempt to respond to a vendor's requested change order within five (5) business days. The State, which includes the requesting agency and the Department of Information Technology, must approve all change orders in writing. The State shall be deemed to have rejected the change order if the parties are unable to reach an agreement in writing.

All change order requests from a vendor to the State, and the State acceptance of a vendor's estimate for a State requested change, will be acknowledged and responded to, either acceptance or rejection, in writing. If accepted, the change order(s) shall be subject to the contract amendment process, as determined to apply by the State.

H-25.7 Deliverables

The vendor shall provide the State with the deliverables and services in accordance with the time frames in the work plan. All deliverables shall be subject to the State's acceptance as set forth in **Section H-25.9: *Testing and Acceptance*** herein. Upon its submission of a deliverable, the vendor represents that it has performed its obligations under the contract associated with the deliverable.

By unconditionally accepting a deliverable, the State reserves the right to reject any and all deliverables in the event the State detects any deficiency in the system, in whole or in part, through completion of all acceptance testing, including but not limited to, software/system acceptance testing, and any extensions thereof.

For each denial of acceptance, the acceptance period may be extended, at the option of the State, by the corresponding time required to correct the deficiency, retest or review.

H-25.7.1 Written Deliverables Review

The State will review the written deliverables for an acceptance period of five (5) business days after receiving written certification from the vendor that the written deliverable is final, complete, and ready for review. The State will notify the vendor in writing of its acceptance or non-acceptance of a deliverable by the end of the five (5) day review period. If any deficiencies exist, the State will notify the vendor in writing of the deficiency and the vendor must correct the deficiency within five (5) business days of receiving notice from the State at no charge to the State. Upon receipt of the corrected deliverable, the State will have five (5) business days to review the corrected written deliverable and notify the vendor in writing of its acceptance or rejection thereof.

H-25.7.2 Software Deliverables Review

Described in **Section H-25.9: *Testing and Acceptance***.

H-25.7.3 Non-Software Deliverables Review

The State will review non-software deliverables to determine whether any deficiency exists and notify the vendor in writing of its acceptance or non-acceptance of the non-software deliverable. The vendor must correct the deficiencies within five (5) business days, or within the period identified in the work plan, as applicable. Following correction of the deficiency, the State will notify the vendor in writing of its acceptance or rejection of the deliverable.

H-25.8 Licenses

The State has defined the software license grant rights, terms and conditions, and has documented the evaluation criteria.

H-25.8.1 Software License Grant

The software license shall grant the State a worldwide, perpetual, irrevocable, non-exclusive, non-transferable, limited license to use the software and its associated documentation, subject to the terms of the contract.

The State may allow its agents and vendors to access and use the software, and in such event, the State shall first obtain written agreement from such agents and vendors that each shall abide by the terms and conditions set forth herein.

H-25.8.2 Software and Documentation Copies

The vendor shall provide the State with a sufficient number of hard copy versions of the software's associated documentation and one (1) electronic version in Microsoft Word and PDF format. The State shall have the right to copy the software and its associated documentation for its internal business needs. The State agrees to include copyright and proprietary notices provided to the State by the vendor on such copies.

H-25.8.3 Restrictions

Except as otherwise permitted under the contract, the State agrees not to:

a. Remove or modify any program markings or any notice of vendor's proprietary rights;
b. Make the programs or materials available in any manner to any third party for use in the third party's business operations, except as permitted herein; or
c. Cause or permit reverse engineering, disassembly or recompilation of the programs.

H-25.8.4 Title

The vendor must hold the right to allow the State to use the software or hold all title, right, and interest (including all ownership and intellectual property rights) in the software and its associated documentation.

H-25.8.5 Third Party

The vendor shall identify all third party contracts to be provided under the contract with the vendor's proposal. The terms in any such contracts must be consistent with this RFP and any resulting contract, including, but not limited to **Appendix H:** ***State of New Hampshire Terms and Conditions General Provisions Form P-37***.

H-25.9 Testing and Acceptance

The State requires that an integrated and coherent approach to complete system testing, security review and testing, deficiency correction, acceptance, and training, and that warranty services be provided to ensure a project.

In its proposal, the vendor is to include its proposed test plan methodology and any scheduling assumptions used regarding the client resource efforts required during testing. After contract award, the vendor will be required to customize its proposed test plan methodology to reflect the needs of the project and include the details of its test plan methodology in the detailed work plan (the first project deliverable). A separate test plan and set of test materials will be prepared for each software function or module.

In addition, the vendor will provide a mechanism for reporting actual test results vs. expected results and for the resolution and tracking of all errors and problems identified during test execution. The vendor will also provide training as necessary to the State staff responsible for test activities.

See **Appendix G-1** for Testing Requirements.

H-25.9.1 Remedies

If the vendor fails to correct a deficiency within the period of time allotted by the state, the vendor shall be deemed to have committed an *Event of Default*, pursuant **Appendix H *Section 8*** and ***H-25.14***, and the State shall have the right, at its option, to pursue the remedies in **Appendix H-25.14.1** as well as to return the vendor's product and receive a refund for all amounts paid to the vendor, including but not limited to, applicable license fees, within ninety (90) days of notification to the vendor of the State's refund request

Notwithstanding any provision of the contract, the State's option to terminate the contract and pursue the stated remedies will remain in effect until the vendor completes the contract to the satisfaction of the State.

H-25.9.2 System Acceptance

Upon completion of the warranty period, the State will issue a letter of final system acceptance.

H-25.10 Warranty

H-25.10.1 Warranty Period

The warranty period will initially commence upon the State issuance of a Letter of Acceptance for UAT and will continue for ninety (90) days. If within the last thirty (30) calendar days of the warranty period, the system software fails to operate as specified, the warranty period will cease, the vendor will correct the deficiency, and a thirty (30) calendar day warranty period will begin. Any further deficiencies with the software must be corrected and run fault free for thirty (30) days.

H-25.10.2 Warranties

H-25.10.2.1 System

The vendor shall warrant that the system must operate to conform to the specifications, terms, and requirements of the contract.

H-25.10.2.2 Software

The vendor shall warrant that the software is properly functioning within the system, compliant with the requirements of the contract, and will operate in accordance with the specifications. Software shall

be archived and or version controlled through the use of harvest software.

H-25.10.2.3 Non-Infringement

The vendor shall warrant that it has good title to, or the right to allow the State to use all services, equipment, and software provided under this contract, and that such services, equipment, and software ("Material") do not violate or infringe any patent, trademark, copyright, trade name or other intellectual property rights or misappropriate a trade secret of any third party.

H-25.10.2.4 Viruses; Destructive Programming

The vendor shall warrant that the software will not contain any viruses, destructive programming, or mechanisms designed to disrupt the performance of the software in accordance with the specifications.

H-25.10.2.5 Compatibility

The vendor shall warrant that all system components, including any replacement or upgraded system software components provided by the vendor to correct deficiencies or as an enhancement, shall operate with the rest of the system without loss of any functionality.

H-25.10.2.6 Professional Services

The vendor shall warrant that all services provided under the contract will be provided in a professional manner in accordance with industry standards and that Services will comply with performance standards.

H-25.10.3 Warranty Services

The vendor shall agree to maintain, repair, and correct deficiencies in the system software, including but not limited to the individual modules or functions, during the warranty period at no additional cost to the State, in accordance with the specifications and terms and requirements of the contract, including without limitation, correcting all errors, and defects and deficiencies; eliminating viruses or destructive programming; and replacing incorrect, defective or deficient software and documentation.

Warranty services shall include, without limitation, the following:

- **a.** Maintain the system software in accordance with the specifications, terms, and requirements of the contract;
- **b.** Repair or replace the system software or any portion thereof so that the system operates in accordance with the specifications, terms, and requirements of the contract;
- **c.** The vendor shall have available to the State on-call telephone assistance, with issue tracking available to the State, twenty four (24) hours per day and seven (7) days a week with an email / telephone response within two (2) hours of request, with assistance response dependent upon issue severity;
- **d.** On-site additional services within four (4) business hours of a request;
- **e.** Maintain a record of the activities related to warrant repair or maintenance activities performed for the State;
- **f.** For all warranty services calls, the vendor shall ensure the following information will be collected and maintained:
  - **1)** nature of the deficiency;
  - **2)** current status of the deficiency;
  - **3)** action plans, dates, and times;
  - **4)** expected and actual completion time;
  - **5)** deficiency resolution information;
  - **6)** identifying number i.e. work order number; and
  - **7)** issue identified by.
- **g.** The vendor must work with the State to identify and troubleshoot potentially large-scale software failures or deficiencies by collecting the following information:
  - **1)** mean time between reported deficiencies with the software;
  - **2)** diagnosis of the root cause of the problem; and
  - **3)** identification of repeat calls or repeat software problems; and
- **h.** All deficiencies found during the warranty period and all deficiencies found with the warranty releases shall be corrected by the vendor no later than five (5) business days, unless specifically extended in writing by the State, at no additional cost to the State.

If in the *Event of Default*, the vendor fails to correct the deficiency within the allotted period of time (see above), the State shall have the right, at its option: 1) declare the vendor in default, terminate the contract, in whole or in part, without penalty or liability to the State; 2) return the vendor's product and receive a refund for all amounts paid to the vendor, including but not limited to, applicable license fees within ninety (90) days of notification to the vendor of the State's intent to request a refund; 3) and to pursue its remedies available at law or in equity.

Notwithstanding any provision of the contract, the State's option to terminate the contract and pursue the remedies above will remain in effect until satisfactory completion of the full warranty period.

H-25.11 Ongoing Software Maintenance and Support Levels

The vendor shall maintain and support the system in all material respects as described in the applicable program documentation after delivery and the warranty period of ninety (90) days through the completion of the contract term.

The vendor will not be responsible for maintenance or support for software developed or modified by the State.

H-25.11.1 Maintenance Releases

The vendor shall make available to the State the latest program updates, general maintenance releases, selected functionality releases, patches, and documentation that are generally offered to its customers, at no additional cost.

H-25.11.2 Vendor Responsibility

The vendor shall be responsible for performing on-site or remote technical support in accordance with the contract documents, including without limitation the requirements, terms, and conditions contained herein.

As part of the software maintenance agreement, ongoing software maintenance and support levels, including all new software releases, shall be responded to according to the following:

a. Class A Deficiencies - The vendor shall have available to the State on-call telephone assistance, with issue tracking available to the State, eight (8) hours per day and five (5) days a week with an email / telephone response within two (2) hours of request; or the vendor shall provide support on-site or with remote diagnostic Services, within four (4) business hours of a request;

b. Class B & C Deficiencies – The State shall notify the vendor of such deficiencies during regular business hours and the vendor shall respond back within four (4) hours of notification of planned corrective action:

- The vendor shall repair or replace software, and provide maintenance of the software in accordance with the specifications, terms and requirements of the contract;
- The vendor shall maintain a record of the activities related to warranty repair or maintenance activities performed for the State;

- For all maintenance services calls, the vendor shall ensure the following information will be collected and maintained: 1) nature of the deficiency; 2) current status of the deficiency; 3) action plans, dates, and times; 4) expected and actual completion time; 5) deficiency resolution information, 6) resolved by, 7) identifying number i.e. work order number, 8) Issue identified by; and

- The vendor must work with the State to identify and troubleshoot potentially large-scale system failures or deficiencies by collecting the following information: 1) mean time between reported deficiencies with the software; 2) diagnosis of the root cause of the problem; and 3) identification of repeat calls or repeat software problems.

If the vendor fails to correct a deficiency within the allotted period of time stated above, the vendor shall be deemed to have committed an *Event of Default*, pursuant to **Appendix H Section H-25.14**, and the State shall have the right, at its option, to pursue the remedies in **Section H-25.14**, as well as to return the vendor's product and receive a refund for all amounts paid to the vendor, including but not limited to, applicable license fees, within ninety (90) days of notification to the vendor of the State's refund request

If the vendor fails to correct a deficiency within the allotted period of time stated above, the vendor shall be deemed to have committed an Event of Default, pursuant to **Appendix H Section H-25.14**, and the State shall have the right, at its option, to pursue the remedies in **Appendix H Section H-25.14**.

H-25.12 Administrative Specifications

H-25.12.1 Travel Expenses

The State will not be responsible for any travel or out of pocket expenses incurred in the performance of the services.

The vendor must assume all travel and related expenses by "fully loading" the proposed labor rates to include, but not limited to: meals, hotel/housing, airfare, car rentals, car mileage, and out of pocket expenses.

H-25.11.2 Shipping and Delivery Fee Exemption

The State will not pay for any shipping or delivery fees unless specifically itemized in the contract.

H-25.12.3 Project Workspace and Office Equipment

The State agency will work with the vendor to determine the requirements for providing all necessary workspace and office equipment, including desktop computers for the vendor's staff. If a vendor has specific requirements, they must be included in the vendor's proposal.

H-25.12.4 Work Hours

N/A

H-25.12.5 Access/Cooperation

As applicable, and reasonably necessary, and subject to the applicable State and federal laws and regulations and restrictions imposed by third parties upon the State, the State will provide the vendor with access to all program files, libraries, personal computer-based systems, software packages, network systems, security systems, and hardware as required to complete the contracted Services**.**

The State will use reasonable efforts to provide approvals, authorizations, and decisions reasonably necessary to allow the vendor to perform its obligations under the contract.

H-25.12.6 State-Owned Documents and Data

The vendor shall provide the State access to all documents, State data, materials, reports, and other work in progress relating to the contract ("State Owned Documents"). Upon expiration or termination of the contract with the State, vendor shall turn over all State-owned documents, State data, material, reports, and work in progress relating to this contract to the State at no additional cost to the State. State-owned documents must be provided in both printed and electronic format.

H-25.12.7 Intellectual Property

Title, right, and interest (including all ownership and intellectual property rights) in the software, and its associated documentation, shall remain with the vendor.

Upon completion and/or termination of the implementation of the project, the vendor shall own and hold all, title, and rights in any software modifications (custom code) developed in connection with performance of obligations under the contract, or modifications to the vendor provided software, and their associated documentation including any and all performance enhancing operational plans and the vendors' special utilities. The vendor shall license back to the State the right to produce, publish, or otherwise use such software, source code, object code, modifications, reports, and documentation developed under the contract.

In no event shall the vendor be precluded from developing for itself, or for others, materials that are competitive with, or similar to custom software, modifications developed in connection with performance of obligations under the contract. In addition, the vendor shall be free to use its general knowledge, skills, experience, and any other ideas, concepts, know-how, and techniques that are acquired or used in the course of its performance under this agreement.

H-25.12.8 IT Required Work Procedures

All work done must conform to standards and procedures established by the DoIT and the State.

H-25.12.9 Computer Use

N/A

H-25.12.10 Email Use

N/A

H-25-12.11 Internet/Intranet Use

N/A

H-25.12.12 Regulatory/Governmental Approvals

Any contract awarded under the RFP shall be contingent upon the vendor obtaining all necessary and applicable regulatory or other governmental approvals.

H-25.12.13 Force Majeure

Neither vendor nor the State shall be responsible for delays or failures in performance resulting from events beyond the control of such party and without fault or negligence of such party. Such events shall include, but not be limited to, acts of God, strikes, lock outs, riots, and acts of War, epidemics, acts of government, fire, power failures, nuclear accidents, earthquakes, and unusually severe weather.

Except in the event of the foregoing, force majeure events shall not include vendor's inability to hire or provide personnel needed for the vendor's performance under the contract.

H-25.12.14 Confidential Information

In performing its obligations under the contract, the vendor may gain access to information of the State, including confidential information. *"State Confidential Information"* shall include, but not be limited to, information exempted from public disclosure under New Hampshire RSA Chapter 91-A: Access to Public Records and Meetings (see e.g. RSA Chapter 91-A: 5 Exemptions). The vendor shall not use the State confidential information developed or obtained during the performance of, or acquired, or developed by reason of the contract, except as is directly connected to and necessary for the vendor's performance under the contract.

The vendor agrees to maintain the confidentiality of and to protect from unauthorized use, disclosure, publication, and reproduction (collectively "release"), all State confidential information of the State that becomes available to the vendor in connection with its performance under the contract, regardless of its form.

Subject to applicable federal or State laws and regulations, confidential information shall not include information which: (i) shall have otherwise become publicly available other than as a result of disclosure by the receiving party in breach hereof; (ii) was disclosed to the receiving party on a non-confidential basis from a source other than the disclosing party, which the receiving party believes is not prohibited from disclosing such information as a result of an obligation in favor of the disclosing party; (iii) is developed by the receiving party independently of, or was known by the receiving party prior to, any disclosure of such information made by the disclosing party; or (iv) is disclosed with the written consent of the disclosing party. A receiving party also may disclose confidential information to the extent required by an order of a court of competent jurisdiction.

Any disclosure of the State's information shall require prior written approval of the State. The vendor shall immediately notify the State if any request, subpoena or other legal process is served upon the vendor regarding the State's confidential information, and the vendor shall cooperate with the State in any effort it undertakes to contest the request, the subpoena or other legal process, at no additional cost to the State.

In the event of unauthorized use or disclosure of the State's confidential information, the vendor shall immediately notify the State, and the State shall immediately be entitled to pursue any remedy at law and in equity, including, but not limited to injunctive relief.

Insofar as the vendor seeks to maintain the confidentiality of its confidential or proprietary information, the vendor must clearly identify in writing the information it claims to be confidential or proprietary. The vendor acknowledges that the State is subject to the Right to Know Law, RSA Chapter 91-A. The State shall maintain the confidentiality of the identified confidential information insofar as it is consistent with applicable State or federal laws or regulations, including but not limited to, RSA Chapter 91-A. In the event the State receives a request for the information identified by the vendor as confidential, the State shall notify the vendor and specify the date the State will be releasing the requested information. At the request of the State, the vendor shall cooperate and assist the State with the collection and review of the vendor's information, at no additional expense to the State. Any effort to prohibit or enjoin the release of the information shall be the vendor's sole responsibility and at the vendor's sole expense. If the vendor fails to obtain a court order enjoining the disclosure, the State shall release the information on the date specified in the State's notice to the vendor without any State liability to the vendor.

This contract agreement, **Appendix H Section H-25.12.14: *Confidential Information*** shall survive the termination or conclusion of a contract.

H-25.12.14 Data Breach

In the event of a data breach, the vendor shall comply with provisions of NHRSA 359C:20.

H-25.13 Pricing

H-25.13.1 Activities/Deliverables/Milestones Dates and Pricing

The vendor must include, within the fixed price for IT service activities, tasks, and preparation of required deliverables, pricing for the deliverables required based on the proposed approach, and methodology and tools. A fixed price must be provided for each deliverable. Pricing worksheets are provided in **Appendix F: *Pricing Worksheets***.

H-25.13.2 Software Licensing, Maintenance, Enhancements and Support Pricing

The vendor must provide the minimum software support and services through software licensing, maintenance, enhancements, and support as detailed in **Section H-25.11: *Ongoing Software Maintenance and Support Levels***.

For software licensing, maintenance, and support costs, complete a worksheet including all costs in the table. A worksheet is provided in **Appendix F: *Pricing Worksheets***, under **Appendix F-5: *Software Licensing, Maintenance, and Support Pricing***, as **Table F-5: *Software Licensing, Maintenance, and Support Pricing Worksheet***.

H-25.13.3 Invoicing

The vendor shall submit correct invoices to the State for all amounts to be paid by the State. All invoices submitted shall be subject to the State's written approval, which shall not be unreasonably withheld. The vendor shall only submit invoices for services or deliverables as permitted by the contract. Invoices must be in a format as determined by the State and contain detailed information, including without limitation: itemization of each deliverable and identification of the deliverable for which payment is sought, and the acceptance date triggering such payment; date of delivery and/or installation; monthly maintenance charges; any other project costs or retention amounts if applicable.

H-25.13.4 Overpayments to the vendor

The vendor shall promptly, but no later than fifteen (15) business days, pay the State the full amount of any overpayment or erroneous payment upon discovery or notice from the State.

H-25.13.5 Credits

The State may apply credits due to the State, arising out of this contract, against the vendor's invoices with appropriate information attached.

H-25.13.6 Records Retention and Access Requirements

The vendor shall agree to the conditions of all applicable State and federal laws and regulations, which are incorporated herein by this reference, regarding retention and access requirements, including without limitation, retention policies consistent with the Federal Acquisition Regulations (FAR) Subpart 4.7 *Vendor Records Retention*.

The vendor and its subcontractors shall maintain books, records, documents, and other evidence of accounting procedures and practices, which properly and sufficiently reflect all direct and indirect costs, invoiced in the performance of their respective obligations under the contract. The vendor and its subcontractors shall retain all such records for three (3) years following termination of the contract, including any extensions. Records relating to any litigation matters regarding the contract shall be kept for one (1) year following the termination of all litigation, including the termination of all appeals or the expiration of the appeals period.

Upon prior notice and subject to reasonable time frames, all such records shall be subject to inspection, examination, audit and copying by personnel so authorized by the State and federal officials so authorized by law, rule, regulation or contract, as applicable. Access to these items will be provided within Merrimack County of the State of New Hampshire, unless otherwise agreed by the State. Delivery of and access to such records shall be at no cost to the State during the three (3) year period following termination of the contract and one (1) year term following litigation relating to the contract, including all appeals or the expiration of the appeal period. The vendor shall include the record retention and review requirements of this section in any of its subcontracts.

The State agrees that books, records, documents, and other evidence of accounting procedures and practices related to the vendor’s cost structure and profit factors shall be excluded from the State’s review unless the cost or any other services or deliverables provided under the contract is calculated or derived from the cost structure or profit factors.

H-25.13.7 Accounting Requirements

The vendor shall maintain an accounting system in accordance with generally accepted accounting principles. The costs applicable to the contract shall be ascertainable from the accounting system and the vendor shall maintain records pertaining to the Services and all other costs and expenditures.

H-25.14 Termination

This section H-25.14 shall survive termination or contract conclusion.

H-25.14.1 Termination for Default

Any one or more of the following acts or omissions of the vendor shall constitute an event of default hereunder ("Event of Default")

a. Failure to perform the services satisfactorily or on schedule;
b. Failure to submit any report required; and/or
c. to perform any other covenant, term or condition of the contract

Upon the occurrence of any event of default, the State may take any one or more, or all, of the following actions:

a) Unless otherwise provided in the contract, the State shall provide the vendor written notice of default and require it to be remedied within, in the absence of a greater or lesser specification of time, within thirty (30) days from the date of notice, unless otherwise indicated within by the State ("cure period"). If the vendor fails to cure the default within the cure period, the State may terminate the contract effective two (2) days after giving the vendor notice of termination, at its sole discretion, treat the contract as breached and pursue its remedies at law or in equity or both.

b) Give the vendor a written notice specifying the event of default and suspending all payments to be made under the contract and ordering that the portion of the contract price which would otherwise accrue to the vendor during the period from the date of such notice until such time as the State determines that the vendor has cured the Event of Default shall never be paid to the vendor.

c) Set off against any other obligations the State may owe to the vendor any damages the State suffers by reason of any event of default;

d) Treat the contract as breeched and pursue any of its remedies at law or in equity, or both.

e) Procure services that are the subject of the contract from another source and the vendor shall be liable for reimbursing the State for the replacement services, and all administrative costs directly related to the replacement of the contract and procuring the services from another source, such as costs of competitive bidding, mailing, advertising, applicable fees, charges or penalties, and staff time costs; all of which shall be subject to the limitations of liability set forth in the contract.

In the event of default by the State, the vendor shall provide the State with written notice of default, and the State shall cure the default within thirty (30) days.

Notwithstanding the foregoing, nothing herein contained shall be deemed to constitute a waiver of the sovereign immunity of the State, which immunity is hereby reserved to the State. This covenant shall survive termination or contract conclusion.

H-25.14.2 Termination for Convenience

The State may, at its sole discretion, terminate the contract for convenience, in whole or in part, by thirty (30) days written notice to the vendor. In the event of such termination for convenience, the State shall pay the vendor the agreed upon price, if separately stated, for deliverables for which acceptance has been given by the State. Amounts for services or deliverables provided prior to the date of termination for which no separate price is stated will be paid, in whole or in part, generally in accordance with **Appendix F: *Pricing Worksheets***.

During the thirty (30) day period, the vendor shall wind down and cease its services as quickly and efficiently as reasonably possible, without performing unnecessary services or activities and by minimizing negative effects on the State from such winding down and cessation of services.

H-25.14.3 Termination for Conflict of Interest

The State may terminate the contract by written notice if it determines that a conflict of interest exists, including but not limited to, a violation by any of the parties hereto of applicable laws regarding ethics in public acquisitions and procurement and performance of contracts.

In such case, the State shall be entitled to a pro-rated refund of any current development, support and maintenance costs. The State shall pay all other contracted payments that would have become due and payable if the vendor did not know, or reasonably did not know, of the conflict of interest.

In the event the contract is terminated as provided above pursuant to a violation by the vendor, the State shall be entitled to pursue the same remedies against the vendor as it could pursue in the event of a default of the contract by the vendor.

H-25.14.4 Termination Procedure

Upon termination of the contract, the State, in addition to any other rights provided in the contract, may require the vendor to deliver to the State any property, including without limitation, software and written deliverables, for such part of the contract as has been terminated.

After receipt of a notice of termination, and except as otherwise directed by the State, vendor shall:

- **a.** Stop work under the contract on the date, and to the extent specified, in the notice;
- **b.** Promptly, but in no event longer than thirty (30) days after termination, terminate its orders and subcontracts related to the work which has been terminated and settle all outstanding liabilities and all claims arising out of such termination of orders and subcontracts, with the approval or ratification of the State to the extent required, which approval or ratification shall be final for the purpose of this Section;
- **c.** Take such action as the State directs, or as necessary to preserve and protect the property related to the contract which is in the possession of vendor and in which State has an interest;
- **d.** Transfer title to the State and deliver in the manner, at the times, and to the extent directed by the State, any property which is required to be furnished to State and which has been accepted or requested by the State; and
- **e.** Provide written certification to the State that vendor has surrendered to the State all said property.

H-25.15 Limitation of Liability

H-25.15.1 State

Subject to applicable laws and regulations, in no event shall the State be liable for any consequential, special, indirect, incidental, punitive, or exemplary damages. Subject to applicable laws and regulations, the State's liability to the vendor shall not exceed the total contract price set forth in contract agreement, **Appendix H Section 1.8 of the *Contract Agreement – General Provisions.***

Notwithstanding the foregoing and any provision of this Contract to the contrary, in no event does the State waive its sovereign immunity or any applicable defenses or immunities.

H-25.15.2 The Vendor

Subject to applicable laws and regulations, in no event shall the vendor be liable for any consequential, special, indirect, incidental, punitive or exemplary damages and the vendor's liability to the State shall not exceed two times (2X) the total contract price set forth in the contract agreement, **Appendix H Section 1.8 of the *Contract Agreement – General Provisions***.

Notwithstanding the foregoing, the limitation of liability shall not apply to the vendor's indemnification obligations set forth in the **Appendix H *Contract Agreement -* Sections 13: *Indemnification* and confidentiality** obligations in **Appendix H 25.12.14: *Confidential Information***, and data breach obligations in **Appendix H-25.12.15 *Data Breach*** which shall be unlimited.

H-25.15.3 State's Immunity

Notwithstanding the foregoing, nothing herein contained shall be deemed to constitute a waiver of the sovereign immunity of the State, which immunity is hereby reserved to the State. This covenant shall survive termination or contract conclusion.

H.25.15.4 Survival

This contract agreement, **Section H-25.15: *Limitation of Liability*** shall survive termination or contract conclusion.

H-25.16 Change of Ownership

In the event that the vendor should change ownership for any reason whatsoever, the State shall have the option of continuing under the contract with the vendor, its successors or assigns for the full remaining term of the contract; continuing under the contract with the vendor, its successors or assigns for such period of time as determined necessary by the State; or immediately terminate the contract without liability to the vendor, its successors or assigns.

H-25.17 Assignment, Delegation and Subcontracts

The vendor shall not assign, delegate, subcontract, or otherwise transfer any of its interest, rights, or duties under the contract without the prior written consent of the State. Such consent will not be unreasonably withheld. Any attempted transfer, assignment, delegation, or other transfer made without the State's prior written consent shall be null and void and may constitute an event of default at the sole discretion of the State.

The vendor shall remain wholly responsible for performance of the entire contract regardless of whether assignees, delegates, subcontractors or other transferees ("Assigns") are used, unless otherwise agreed to in writing by the State and the assigns fully assumes in writing any and all obligations and liabilities under the contract from the effective date. In the absence of a written assumption of full obligations and liabilities of the contract, any permitted assignment, delegation, subcontract or other transfer shall neither relieve the vendor of any of its obligations under the contract nor shall it affect any remedies available to the State against the vendor that may arise from any event of default of the provisions of the contract. The State will consider the vendor to be the sole point of contact with regard to all contractual matters, including payment of any and all charges resulting from the contract.

H-25.18 Dispute Resolution

Prior to the filing of any formal proceedings with respect to a dispute (other than an action seeking injunctive relief with respect to intellectual property rights or confidential Information), the party believing itself aggrieved (the "Invoking Party") shall call for progressive management involvement in the dispute negotiation by written notice to the other party. Such notice shall be without prejudice to the invoking party's right to any other remedy permitted by this agreement.

H-25.19 Venue and Jurisdiction

Any action on the contract may only be brought in the State of New Hampshire Merrimack County Superior Court.

H-25.20 Project Holdback

The State will withhold 10 percent of the agreed deliverables pricing tendered by the vendor in this engagement until completion of the Warranty Period as defined in **Appendix H Section 25-10.1: *Warranty Period***.

H-25.21 Escrow of Code

N/A

TERMS AND DEFINITIONS

The following general contracting terms and definitions apply except as specifically noted elsewhere in this document.

Acceptance	Notice from the State that a deliverable has satisfied acceptance test or review
Acceptance Letter	An acceptance letter provides notice from the State that a deliverable has satisfied acceptance tests or review
Acceptance Period	The timeframe during which the acceptance test is performed
Acceptance Test Plan	The *Acceptance Test Plan* provided by the vendor and agreed to by the State that describes at a minimum, the specific acceptance process, criteria, and schedule for deliverables
Acceptance Test and Review	Tests performed to determine that no defects exist in the application Software or the System
Access Control	Supports the management of permissions for logging onto a computer or network
Agreement	A contract duly executed and legally binding
Appendix	Supplementary material that is collected and appended at the back of a document
Audit Trail Capture and Analysis	Supports the identification and monitoring of activities within an application or system
Best and Final Offer (BAFO)	For negotiated procurements, a vendor's final offer following the conclusion of discussions
Breach or Breach of Security	Unlawful and unauthorized acquisition of unencrypted computerized data that materially compromises the security, confidentiality or integrity of personal information maintained by a person or commercial entity
CCP	Change Control Procedures
CR	Change Request
COTS	Commercial Off-The-Shelf Software
CM	Configuration Management
Certification	The vendor’s written declaration with full supporting and written documentation (including without limitation test results as applicable) that the vendor has completed development of the deliverable and certified its readiness for applicable acceptance testing or review
Change Control	Formal process for initiating changes to the proposed solution or process once development has begun
Change Order	Formal documentation prepared for a proposed change in the Specifications
Completion Date	End date for the contract
Confidential Information	Information required to be kept confidential from unauthorized disclosure under the contract

Contract	This agreement between the State of New Hampshire and a vendor, which creates binding obligations for each party to perform as specified in the contract documents
Contract Conclusion	Refers to the conclusion of the contract, for any reason, including but not limited to, the contract completion, termination for convenience, or termination for default
Contract Documents	Documents that comprise this contract
Contract Managers	The persons identified by the State and the vendor who shall be responsible for all contractual authorization and administration of the contract. These responsibilities shall include but not be limited to processing contract documentation, obtaining executive approvals, tracking costs and payments, and representing the parties in all contract administrative activities.
Contracted vendor	The vendor whose proposal or quote was awarded the contract with the State and who is responsible for the services and deliverables of the contract.
Conversion Test	A test to ensure that a data conversion process correctly takes data from a legacy system and converts it to form that can be used by the new system.
COTS	Commercial Off the Shelf
Cure Period	The thirty (30) day period following written notification of a default within which a contracted vendor must cure the default identified.
Custom Code	Code developed by the vendor specifically for this project for the State of New Hampshire
Custom Software	Software developed by the vendor specifically for this project for the State of New Hampshire
Data	State’s records, files, forms, data and other documents or information, in either electronic or paper form, that will be used /converted by the vendor during the contract term
DBA	Database Administrator
Deficiencies/Defects	A failure, deficiency, or defect in a deliverable resulting in a deliverable, the software, or the system, not conforming to its specifications. **Class A Deficiency** – Software - Critical, does not allow system to operate, no work around, demands immediate action; written documentation - missing significant portions of information or unintelligible to State; non software - services were inadequate and require re-performance of the Service. **Class B Deficiency** – Software - important, does not stop operation and/or there is a work around and user can perform tasks; written documentation - portions of information are missing but not enough to make the document unintelligible; non software - services were deficient, require reworking, but do not require re-performance of the Service.

	Class C Deficiency – Software - minimal, cosmetic in nature, minimal effect on system, low priority and/or user can use system; written documentation - minimal changes required and of minor editing nature; non-software - services require only minor reworking and do not require re-performance of the service.
Deliverable	A deliverable is any written, software, or non-software deliverable (letter, report, manual, book, other), provided by the vendor to the State or under the terms of a contract requirement.
Department	An agency of the State
Department of Information Technology (DoIT)	The Department of Information Technology established under RSA 21-R by the legislature effective September 5, 2008.
Documentation	All information that describes the installation, operation, and use of the software, either in printed or electronic format.
Differential Item Functioning (DIF)	Analysis that provides an indication of unexpected behavior of items on a test.
Digital Signature	Guarantees the unaltered state of a file
Effective Date	The contract and all obligations of the parties hereunder shall become effective on the date the Governor and the Executive Council of the State of New Hampshire approves the contract
Encryption	Supports the transformation of data for security purposes
Enhancements	Updates, additions, modifications to, and new releases for the software, and all changes to the documentation as a result of enhancements, including, but not limited to, enhancements produced by change orders
Event of Default	Any one or more of the following acts or omissions of a vendor shall constitute an event of default hereunder ("Event of Default") a) Failure to perform the Services satisfactorily or on schedule; b) Failure to submit any report required; and/or c) Failure to perform any other covenant, term or condition of the contract
Firm Fixed Price Contract	A firm-fixed-price contract provides a price that is not subject to increase, i.e., adjustment on the basis of the vendor's cost experience in performing the contract
Fully Loaded	Rates are inclusive of all allowable expenses, including, but not limited to: meals, hotel/housing, airfare, car rentals, car mileage, and out of pocket expenses
GAAP	Generally Accepted Accounting Principles
Governor and Executive Council	The New Hampshire Governor and Executive Council
Harvest	Software to archive and/or control versions of software
Identification and	Supports obtaining information about those parties attempting to

Authentication	log on to a system or application for security purposes and the validation of those users
Implementation	The process for making the system operational for processing the data.
Implementation Plan	Sets forth the transition from development of the system to full operation, and includes without limitation, training, business and technical procedures
Information Technology (IT)	Refers to the tools and processes used for the gathering, storing, manipulating, transmitting, sharing, and sensing of information including, but not limited to, data processing, computing, information systems, telecommunications, and various audio and video technologies
Input Validation	Ensure that the values entered by users or provided by other applications meets the size, type and format expected. Protecting the application from cross site scripting, SQL injection, buffer overflow, etc.
Intrusion Detection	Supports the detection of illegal entrance into a computer system
Invoking Party	In a dispute, the party believing itself aggrieved
Key Project Staff	Personnel identified by the State and by the contracted vendor as essential to work on the project
Licensee	The State of New Hampshire
Non Exclusive Contract	A contract executed by the State that does not restrict the State from seeking alternative sources for the deliverables or services provided under the contract.
Non-Software Deliverables	Deliverables that are not software deliverables or written deliverables, e.g., meetings, help support, services, other
Notice to Proceed (NTP)	The State contract manager's written direction to the vendor to begin work on the contract on a given date and time
Open Data Formats	A data format based on an underlying open standard.
Open Source Software	Software that guarantees the user unrestricted use of the software as defined in RSA 21-R:10 and RSA 21-R:11.
Open Standards	Specifications for the encoding and transfer of computer data that is defined in RSA 21-R:10 and RSA 21-R:13.
Operating System	System is fully functional, all data has been loaded into the system, is available for use by the State in its daily operations.
Operational	Operational means that the system is operating and fully functional; all data has been loaded; the system is available for use by the State in its daily operations; and the State has issued an acceptance letter.
Order of Precedence	The order in which contract/documents control in the event of a conflict or ambiguity. A term or condition in a document controls over a conflicting or ambiguous term or condition in a document that is lower in the order of precedence
Project	The planned undertaking regarding the entire subject matter of an RFP and contract and the activities of the parties related hereto.

Project Team	The group of State employees and contracted vendor's personnel responsible for managing the processes and mechanisms required such that the Services are procured in accordance with the Work Plan on time, on budget and to the required specifications and quality
Project Management Plan	A document that describes the processes and methodology to be employed by the vendor to ensure a project.
Project Managers	The persons identified who shall function as the State's and the vendor's representative with regard to review and acceptance of contract deliverables, invoice sign off, and review and approval of *Change Requests (CR)* utilizing the *Change Control Procedures (CCP)*
Project Staff	State personnel assigned to work with the vendor on the project
Proposal	The submission from a vendor in response to the request for a proposal or statement of work.
Regression Test Plan	A plan integrated into the work plan used to ascertain whether fixes to defects have caused errors elsewhere in the application/process.
Review	The process of reviewing deliverables for acceptance
Review Period	The period set for review of a deliverable. If none is specified then the review period is five (**5**) business days.
RFP (Request for Proposal)	A Request For Proposal solicits proposals to satisfy State functional requirements by supplying data processing product and/or service resources according to specific terms and conditions
Role/Privilege Management	Supports the granting of abilities to users or groups of users of a computer, application or network
Schedule	The dates described in the Work Plan for deadlines for performance of Services and other Project events and activities under the contract
SaaS	Software as a service - Occurs where the COTS application is hosted but the State does not own the license or the code.
Service Level Agreement (SLA)	A signed agreement between the vendor and the State specifying the level of Service that is expected of, and provided by, the vendor during the term of the contract.
Services	The work or labor to be performed by the vendor on the project as described in the contract.
Software	All custom software and COTS Software provided by the vendor under the contract
Software Deliverables	COTS software and enhancements
Software License	Licenses provided to the State under this contract
Solution	The solution consists of the total solution, which includes, without limitation, software and services, addressing the requirements and terms of the specifications. The off-the-shelf software and configured software customized for the State provided by the vendor in response to this RFP.

Specifications	The written specifications that set forth the requirements which include, without limitation, this RFP, the proposal, the contract, any performance standards, documentation, applicable State and federal policies, laws and regulations, State technical standards, subsequent State-approved deliverables, and other specifications and requirements described in the contract documents. The specifications are, by this reference, made a part of the contract as though completely set forth herein.
State	Reference to the term "State" shall include applicable agencies as defined in Section 1: INTRODUCTION of this RFP.
Statement of Work (SOW)	A Statement of Work clearly defines the basic requirements and objectives of a Project. The Statement of Work also defines a high level view of the architecture, performance and design requirements, the roles and responsibilities of the State and the vendor. The SOW defines the results that the vendor remains responsible and accountable for achieving.
State's Confidential Records	State's information regardless of its form that is not subject to public disclosure under applicable state and federal laws and regulations, including but not limited to RSA Chapter 91-A
State Data	Any information contained within State systems in electronic or paper format.
State Fiscal Year (SFY)	The New Hampshire fiscal year extends from July 1st through June 30th of the following calendar year
State Project Leader	State's representative with regard to project oversight
State's Project Manager (PM)	State's representative with regard to project management and technical matters. Agency Project Managers are responsible for review and acceptance of specific contract deliverables, invoice sign off, and Review and approval of a change proposal (CP).
Subcontractors	A person, partnership, or company not in the employment of, or owned by, the vendor, which is performing services under this contract under a separate contract with or on behalf of the vendor
System	All software, specified hardware, and interfaces and extensions, integrated and functioning together in accordance with the specifications.
TBD	To Be Determined
Technical Authorization	Direction to a vendor, which fills in details, clarifies, interprets, or specifies technical requirements. It must be: (**1**) consistent with Statement of Work within Statement of Services; (**2**) not constitute a new assignment; and (**3**) not change the terms, documents of specifications of the SOW.
Test Plan	A plan, integrated in the work plan, to verify the code (new or changed) works to fulfill the requirements of the project it may consist of a timeline, a series of tests and test data, test scripts and reports for the test results as well as a tracking mechanism.

Term	The duration of the contract.
Transition Services	Services and support provided when the contracted vendor is supporting system changes.
UAT	User Acceptance Test
Unit Test	Developers create their own test data and test scenarios to verify the code they have created or changed functions properly as defined.
User Acceptance Testing	Tests done by knowledgeable business users who are familiar with the scope of the project. They create/develop test cases to confirm the system was developed according to specific user requirements. The test cases and scripts/scenarios should be mapped to business requirements outlined in the user requirements documents.
User Management	Supports the administration of computer, application and network accounts within an organization
Vendor	The contracted individual, firm, or company that will perform the duties and Specifications of the contract.
Verification	Supports the confirmation of authority to enter a computer system, application or network.
Walk Through	A step-by-step review of a specification, usability features or design before it is handed off to the technical team for development.
Warranty Period	A period of coverage during which the contracted vendor is responsible for providing a guarantee for products and services delivered as defined in the contract.
Warranty Releases	Code releases that are done during the warranty period.
Warranty Services	The services to be provided by the vendor during the warranty period.
Work Plan	The overall plan of activities for the project created in accordance with the contract. The plan and delineation of tasks, activities and events to be performed and deliverables to be produced under the project as specified in Appendix C. The work plan shall include a detailed description of the schedule, tasks/activities, deliverables, critical events, task dependencies, and the resources that would lead and/or participate on each task.
Written Deliverables	Non-software written deliverable documentation (letter, report, manual, book, other) provided by the vendor either in paper or electronic format.

Signed Addendum 1

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

RFP Questions Received by 12:00 PM on April 6, 2017
Additional Questions from the Vendor Conference

Reference	Question	Response
Page 4 of Cover Sheet, Contract Type: Not to Exceed $3M	Please clarify if the "not to exceed $3M" encompasses the entire initial contract term of 3 years.	$3M Annually
Page 35, Section C-2 Requirements Table C-2 is included as an attachment to this RFP. C2 / p. 35	It appears Table C-2 was not included as an attachment to the RFP. Please provide direction on where to find Table C-2.	Table C referred to on page 35 is the Excel Workbook attached to the RFP on the posting site. We apologize for the naming of the posted filed. It should read Table C.
Cover sheet	Please confirm that operational assessments are required for all three content areas (ELA, math, and science) for the 2017-2018 school year.	Operational assessment for all three content areas (ELA, math, and science) are required for the 2017-2018 school year.
Second page of RFP and… p. 33	**From second page of RFP:** "Vendors may bid on all five (5) components (components A, B, C, D, E), on any combination of assessment components (A, B, C, D), or any individual assessment component (A, B, C, D). For each component selected by the vendor, the proposal must address component E, and must address each component selected completely. The NH DOE will not accept bids for pieces of individual components." and… **From page 33:** **C-1 SCOPE OF WORK** "The *Scope of Work* should be applied to each Assessment Component bid on:" Q: Are the above statements (red type) indicating that if a bidder wishes to bid on all five components, for example, they	The bidder does not need to include five different sets of narratives if bidding on multiple components. Ex: If a bidder is bidding on the summative components for ELA, math and science, a single narrative could encompass the three subject areas. Ex: If a bidder is bidding on the summative and interim components for ELA, math and science, then the single narrative should ensure that any differences in the specifications of the summative and interim components be identified. This option was intended to give bidders an opportunity to present a proposal that may not be inclusive of all components. Ex: Summative option for ELA, math, science.

RFP Questions Received by 12:00 PM on April 6, 2017
Additional Questions from the Vendor Conference

	will need five different sets of narratives—each one covering most of the 45 narrative topics—and each of the five will need to apply the full “Scope of Work” to their response? Or is there a different scenario intended?	Ex: Interim option for ELA, math, science. Ex: Summative and interim option for ELA and math only. Etc.
Pricing - 4th page before page numbers begin	Regarding the “not to exceed $3M” note on “contract type:” Please confirm that this is per-year.	Confirmed. NTE of $3M is per year.
4.17 / p. 11 - Organization	Other than for the Executive Summary, is there to be any page-count limit—or guidance—for portions of the proposal, such as the 45 Narrative Topics in Section IV?	We did not want to put page limits. However we encourage clear and concise language leading to responses of reasonable length.
5.4.4 / p. 21	a. Please explain why this section talks only of the “Vendor-proposed software solution cost” getting allocated up to 24 points per bidder. Might it rather be the price of the <u>total solution</u> that somehow gets scored on the 24-point allocation? b. Given that the terms of pricing for this RFP are a “not-to-exceed $3M”…on what basis will points be awarded to each bidder?	Vendor proposed software solution cost is **inclusive of the total solution**. The State will consider both implementation costs and subsequent year license and maintenance costs, provided in Tables F-1: Activities/Deliverables/Milestones Pricing Worksheet, F-4: Software Licensing, Maintenance, and Support Pricing Worksheet and, if appropriate, F-5: Web Site Hosting, Maintenance, and Support Pricing Worksheet. The cost information required in a proposal is intended to provide a sound basis for comparing costs. Factors include but are not limited to: Price Proposal • Cost effective budget. • Sound fiscal management

RFP Questions Received by 12:00 PM on April 6, 2017
Additional Questions from the Vendor Conference

		practices that meet or exceed industry standards. Criteria for these scores will be found in but are not limited to: Proposal Section III: Responses to Requirements and Deliverables and Technical Requirements Proposal Section IV: Narrative Responses Section VII: Pricing Model Vendor Presentations Proposed Work Plan
5.4.4 / p. 21	a. Is the expectation of "not to exceed $3M" to apply to proposals bidding all five components? b. If pricing is for seven years, is inflation/COLA allowable past year 1?	a. The $3M (annual) is inclusive for all requirements (A-E) b. All vendor price quotes will be considered. We cannot be more definitive that what we have already specified in the RFP with is an NTE of $3M per year contingent upon federal and state funding and Governor and Executive Council approval.
A-1.7, A-1.8 / p. 26	Please confirm whether formative assessments—separate from summative and interim assessments—are required.	Formative assessments are not a required component for this RFP.
A-1.8 / p. 27	Please confirm, by subject area(s), whether the NHDOE is requesting an off-the-shelf product for the 2017-2018 school year, followed by a custom-developed assessment for subsequent years.	Vendor-developed solutions are acceptable. The subject areas included in this RFP are ELA, math and science.

C2 / p. 35	**From page 35:** **"C-2 REQUIREMENTS** Vendors shall complete the requirements checklist (Table C-2 General Requirements Vendor Response Checklist). Table C-2 is included as an attachment to this RFP." Q: Is there a Table C-2 that was meant to be included as an attachment to this RFP?	Table C referred to on page 35 is the Excel Workbook attached to the RFP on the posting site. We apologize for the naming of the posted filed. It should read Table C.
Topic 2.5 / p. 48	Please confirm that the first administration of new field-test items is the 2018-2019 school year—after new ELA, math, and science standards are approved.	Section A-1.8 Field tested items may appear alongside operational items in all years including Year 1 (2017-2018). Scores must be reportable for accountability purposes in all years. *Science standards will remain as is through 2022.
Topic 17 / p. 66	"The vendor shall describe how the different types of scores it is proposing, individual scale scores and subscores, will be produced and verified. The vendor must include scores produced strictly on items which are computer-scorable and scores produced based on a combination of the computer-scored and hand-scored items. The limitations in interpretation of both of these scores must be discussed." Q: Please clarify or elaborate on what is meant here by subscores.	Subscore clarification using ELA as an example: Overall scale score for ELA is then broken into subscores for each claim achievement category: The Claims for English Language Arts are: •Reading •Writing •Speaking & Listening •Research/Inquiry
Topic 19.2 / p. 68	Please confirm that one or both of these similar processes will fulfill the requirement of "an independent real-time review of the equating process."	The NH DOE may identify an independent reviewer to review the equating process, analyses, and results. The vendor(s) must support this effort by providing the consultant(s) with the

	• Third-party parallel equating • Third-party replication If not, please describe the requirements of the process in more detail.	necessary data files and other materials in a timely manner during the equating process.
Appendix F/ pp. 100-103	a. Is Appendix F: (Pricing Worksheets) available in Excel format? b. Should bidders budget for 4 years (through 2021) or 7 years (through 2024)?	An Excel file will be provided. Bidders should include a budget for the four years included on the RFP, through 2021. Bidders may include additional years or information if they choose to.
Appendix F/ pp. 100-103	a. What is the timeframe (# of years) vendors are to include in the Price category for Activities/Deliverables/ Milestones 1-35 in Table F-1? b. Where should vendors capture pricing for activities not included in 1-35 (i.e. item development, committee meetings, scoring, reporting, analysis, etc.)?	Table F-1 should function as a Year 1 implementation and payment schedule. Vendors may add additional rows to the pricing worksheet for milestones not included in Table F-1. Vendors may include additional columns or tables for milestones and other items occurring subsequent to Year 1.
Appendix F/ pp. 100-103	a. What is the timeframe (# of years) vendors are to include in the Hours X Rate for Table F-2? b. Please clarify “information is required by phase.”	Vendors should include Year 1 pricing in Table F- 2. Information is required for each implementation phase of the project.
Appendix F/ pp. 100-103	Please identify the specific years for Future Vendor Rates Worksheet Table F-3.	The information in Table F-3 is required
Pricing - General	Would a proposal for all five components, with a bottom-line annual price of more than	Bidders should tailor proposals to meet the NTE of $3M.

	$3 million, be automatically judged non-compliant and rejected? Or is any other scenario possible?	
Pricing - NTE	This RFP contains five components (A-E). Please clarify if the “not to exceed $3M” requirement is inclusive for ALL five components.	The $3M (annual) is inclusive for all requirements (A-E)
Page 42, Topic 2 Item Development, second paragraph:	While this RFP seeks an off-the-shelf solution, such a solution as well as a customized solution, may/will involve the creation of specific test items. This section dealing with Item Development applies to the development of items, whether used to tailor an off-the-shelf solution in response to this RFP or in a customized solution. Per NH RSA 193-C:3 III (d); “teachers will be involved in design and using the assessment system.” Much of the RFP seems intended to guide the custom development of a summative assessment. Considering the State is looking for an off-the-shelf solution, how does this involve New Hampshire educators in the item creation and on-going item review and release process if the State is seeking an out-of-the-box interim assessment?	NH Teachers must be involved in the on-going development of items and the item review.
Pages 39 – 95, Appendix D, Section D-1 through D-5	Are there requirements in the RFP that are intended for summative-only responses; since many of the requirements are directed at a summative assessment solution? If a	Yes there will be some tasks that can be omitted. For example: A concordance table is not necessary if only an interim solution is proposed.

	vendor chose to submit only as an interim solution, are there any requirements that can be omitted because they are summative-only?	
Page 92 – 95, Topic 45, Transition	Does the State intend to claim ownership of test items used in this solution, even if they were not developed for New Hampshire?	No the state does not intend to claim ownership of items not developed for New Hampshire.

Correction:

Page 21 Section 5.4.4
Vendor proposed software solution cost will be allocated a maximum score of twenty-four (24) points. The State will consider both implementation costs and subsequent year license and maintenance costs, provided in Tables F-1: *Activities/Deliverables/Milestones Pricing Worksheet*, F-5: *Software Licensing, Maintenance, and Support Pricing Worksheet* and, if appropriate, F-6: *Web Site Hosting, Maintenance, and Support Pricing Worksheet.*
Should read:
F-4: *Software Licensing, Maintenance, and Support Pricing Worksheet* and, if appropriate, F-5: *Web Site Hosting, Maintenance, and Support Pricing Worksheet*

Page 21 Section 5.4.4
"Vendor-proposed software solution cost"
Vendor proposed software solution cost is **inclusive of the total solution**.

Topic 1 Test Design
The summative assessments will use a common-matrix design to support a) the reporting of student-level overall **science** performance in terms of performance levels and scaled scores and b) the reporting of school- and district-level scores in a manner that reflects the depth and breadth of the academic standards.

The word **science** should be omitted.

Update:

Standards Revision and Alignment
The NH academic standards for ELA, mathematics and science are entering a revision cycle. It is therefore imperative that vendors acknowledge and include in proposals a plan for assessment revision aligned to updated standards in 2018-2019.

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

RFP Questions Received by 12:00 PM on April 6, 2017
Additional Questions from the Vendor Conference

*The State Board of Education voted on April 6, 2017 to not begin the revision process for the NH Academic Standards for Science until 2022.

*The State Board of Education will revisit the timeline for revision of the NH Academic Standards ELA and math at the May 11, 2017 Board meeting; this may alter the need for assessment re-alignment in 2018-2019.

Acknowledged by: Michael Casey, J.D. Contracts Officer
American Institutes for Research

Signed:______________________________________

Signed Addendum 2

STATE OF NEW HAMPSHIRE
Department of Education
NH Statewide Assessments: ELA, Mathematics, Science
RFP 2017-073 DOE New Hampshire Statewide Assessments

RFP Questions Received between 12:00 pm on April 6, 2017 and 12:00 pm on April 11, 2017

Reference	Question	Response
Vendor Conference	Are NECAP items available for future vendor use?	The use of items would have to be approved by the consortium states. The items are not aligned to the NH Academic Standards for Science.
Vendor Conference	Is the state looking for a formative assessment component?	A formative component was not requested in the RFP.
Vendor Conference	Given the student data privacy laws in NH, what level of longitudinal reporting is the State seeking?	Down to classroom level.
Vendor Conference	Will a list of attendees be available for viewing?	Yes. The attendee list will be posted on the NH DOE webpage. https://www.education.nh.gov/rfp/index.htm
Vendor Conference	Do you have an estimate of how many students will participate in the on-line assessment versus paper-pencil?	Unfortunately we cannot estimate these numbers for you.
Vendor Conference	Will an Excel copy of the Appendix F Worksheets be made available?	Yes. The Excel document has been posted on the NH DOE webpage. https://www.education.nh.gov/rfp/index.ht

Acknowledged by: Michael Casey, J.D. Contracts Officer
American Institutes for Research

Signed:

Appendix A:

Resumes

Table of Contents

JON COHEN, PH.D.

PRESIDENT, AIR ASSESSMENT

Education

Ph.D.	1992, Government (Methodology and American Politics) University of Texas at Austin, Austin, TX
M.A.	1988, Government (Methodology and American Politics) University of Texas at Austin, Austin, TX
B.A.	1985, Politics and Philosophy Bennington College, Bennington, VT

Present Position

Executive Vice President, AIR, and President, AIR Assessment Program, American Institutes for Research (AIR) (2012–Present)
Responsible for all aspects of AIR's K–12 Assessment program, with annual revenues of $100 million. As Chief Statistician, responsible for ensuring that AIR's statistical research contributes to AIR's mission of improving society through better social and behavioral science. This entails responsibility for developing and evaluating new statistical procedures as well as for the software to implement them.

Project Experience at AIR

Successfully directs the Assessment program, responsible for all aspects of AIR's K–12 Assessment program, with annual revenues of $100 million (2004–present), and for the statistical research of the program, as AIR's Chief Statistician (1998–present). As Director, Statistical Services for Secondary Analysis, National Assessment of Educational Progress (NAEP), leads the development, distribution, and support of statistical software designed for the analysis of NAEP data (1999–present). Serves as Director of Statistics and Psychometrics, National Assessment of Adult Literacy (1999–present), this study, which is to provide statistics and figures on adult literacy rates, utilizes a multistage sample design selected with unequal probabilities and a partially balanced, incomplete block matrix test design. Calibration of the assessment includes both traditional item response methods and a theory-driven multicomponent model developed by Dr. Cohen. Study Director, Evaluation and Review of Statistical Methods Used in NAEP (2003–2004). Statistical and Psychometric Advisor, California High School Exit Exam (CAHSEE) (2000–2002). Study Director, Models of Total Survey Error (1996–2001). Project Director, Technical and Analytical Support for the Assessment Group of National Center for Education Statistics 6(NCES) (1996–2000).

Employment History

2012–Present	Executive Vice President, AIR, and President, AIR Assessment Program, American Institutes for Research, Washington, DC

2004–2012	Senior Vice President and Director of Assessment Program, American Institutes for Research, Washington, DC
1998–Present	Chief Statistician, American Institutes for Research, Washington, DC
2003–2004	Director, Computer and Statistical Sciences Center, American Institutes for Research, Washington, DC
1998–2004	Associate Director for Research, American Institutes for Research, Washington, DC
1997–1998	Managing Associate/Senior Statistician, American Institutes for Research, Washington, DC
1988–1997	Research Associate/Research Analyst/Senior Research Analyst, American Institutes for Research, Washington, DC
1986–1988	Systems Analyst, Defense Research Technologies (DRT), Rockville, MD
1985–1988	Teaching Assistant, University of Texas at Austin, Austin, TX

STEVE KROMER, B.B.A.
SENIOR VICE PRESIDENT AND CHIEF OPERATING OFFICER, AIR ASSESSMENT

Education

B.B.A. 1985, Management Information Systems
University of Iowa, Iowa City, IA

Present Position

Senior Vice President and Chief Operating Officer, AIR Assessment, American Institutes for Research (AIR)
Responsible for building on and fulfilling AIR's goal of providing clients with customized assessments that measure student achievement against the state standards and provide meaningful score reports that can help students, parents, and educators address any area of student weakness. Manages and provides leadership to the Assessment Program with a primary focus on developing and maturing the operational policies and procedures of the business. Identifies and implements new systems and processes as needed to produce the quality results required in the high-stakes assessment industry. Communicates directly with all AIR clients.

Project Experience at AIR

Successfully manages and provides leadership to AIR's Assessment Program with a primary focus on developing and maturing the operational policies and procedures of the business as the Chief Operating Officer of Assessment (2007–present).

Employment History

2012–Present	Senior Vice President and Chief Operating Officer, AIR Assessment, American Institutes for Research, Washington, DC
2007–2012	Executive Director of Operations, American Institutes for Research, Washington, DC
1985–2006	Pearson Educational Measurement, Iowa City, IA
	Executive Vice President and General Manager (2003–2006)
	Vice President and General Manager (2000–2003)
	Director, Operations (1998–2000)
	Account Executive, Publisher Services (1994–1998)
	Manager State Assessments, Software Development (1989–1994)
	Analyst Programmer/Systems Analyst (1985–1989)

DEEANN WAGNER, M.B.A.

VICE PRESIDENT, ASSESSMENT PROGRAMS AND OPERATIONS

Education

M.B.A.	1993, Business Administration Philadelphia University, Philadelphia, PA
B.S.	1980, Business Administration (Operations Management) University of Delaware, Newark, DE

Professional Credentials and Certifications

International Organization for Standardization (ISO) Certified Lead Assessor, British Standards Institute

Six Sigma Black Belt (statistical process control to achieve total customer satisfaction in services as well as products), Six Sigma Research Institute, Motorola University

Present Position

Vice President, Assessment Programs and Operations, American Institutes for Research (AIR) (2003–Present)
Oversees the work of AIR's content, technical, score reporting, production, and operations team leaders. Manages multiple separate contracts for Ohio covering summative assessments grades 3–8 and graduation tests in five subjects, Ohio's Test of English Language Acquisition, Ohio's Alternate Assessments for Students with Disabilities, and other smaller efforts. Manages the activities of more than 80 staff and supervises the Columbus office, which runs on schedule and within budget and receives high client satisfaction. Ensures that test development staff meet deadlines. Worked to restructure almost every aspect of the development process, including committee review procedures and staffing, bias reviews, rangefinding, and item development schedules.

Manages subcontracts with AIR's operations partners and is intimately involved in test operations—from production to distribution—of regular and special versions. Personally led the redesign of the accommodations for testing in other languages, increasing the abilities of the state to test students and of AIR to deliver a high-quality product. Is involved in all new work assignments. Has developed business plans for score reporting, websites, and other new initiatives.

Project Experience at AIR

Acting Project Director/Director of Operations (2001–2003)
Maintained responsibility for managing and overseeing all aspects of the Training and Technical Assistance Center for the Comprehensive Community Mental Health Services for Children and Their Families Program. These responsibilities included working closely with the client, multiple program partners, and subcontractors to deliver relevant and timely technical assistance to 63 federally funded systems of care community sites.

Director of Operations (2000–2001)
Managed and oversaw all logistical aspects of the School District of Philadelphia's Citywide Proficiency Exams. Responsibilities included working closely with each member of the client's management team, managing multiple subcontractors and 50 internal staff to prepare and deliver district-wide assessment tools for grades K–12.

Additional responsibilities included overseeing all financial aspects of the Voluntary National Testing program of the National Assessment Governing Board.

Employment History

2003–Present	Corporate Director of Project Management, American Institutes for Research, Washington, DC
2001–2003	Acting Project Director/Director of Operations, American Institutes for Research, Washington, DC
2000–2001	Director of Operations, American Institutes for Research, Washington, DC
1992–1994	Quality Performance Specialist, Asea Brown Boveri, Richmond, VA
1990–1992	Operations Analyst/Financial Controller, Asea Brown Boveri, Richmond, VA
1985–1990	Associate Program Director, Advanced Placement Program, Educational Testing Service, Princeton, NJ
1983–1985	Assistant Program Director, College Scholarship Services, Educational Testing Service, Princeton, NJ
1980–1983	Senior Program Assistant, Test Center Management, Educational Testing Service, Princeton, NJ

HEATHER HAYES, J.D.

VICE PRESIDENT, ASSESSMENT PROGRAMS AND CLIENT SERVICES

Education

J.D.	1996, Juris Doctorate University of Iowa College of Law, Iowa City, IA
B.A.	1992, English University of Iowa, Iowa City, IA

Present Position

Vice President, Assessment Programs and Client Services, American Institutes for Research (AIR) (2010–Present)
Provides executive leadership for a program management and client services organization responsible for the effective implementation of online assessment programs and systems in California, Delaware, Florida, Hawaii, Oregon, Minnesota, Utah, and the Smarter Balanced Assessment Consortium. Monitors project activities for emerging issues and determine the best method to mitigate deviations. Develops and formalizes internal processes to improve ability to deliver and build capacity. Provides supervision and training to management staff and internal project teams. Participates in planning activities with and serves as a senior point of contact for state departments of education and consortia.

Project Experience at AIR

Senior Project Director, Oregon Assessment of Knowledge and Skills (OAKS) (2008–2010)
Managed the Oregon Assessment of Knowledge and Skills (OAKS) program, the first No Child Left Behind-approved online adaptive state assessment system. Served as the primary point of contact for the Oregon Department of Education. Created a master project plan and documented all business processes required to support the delivery of approximately 5 million tests to more than 335,000 students in grades 3–12 during a 9-month annual testing window. Led the configuration and deployment of AIR's online, distributed performance scoring system and trained Oregon educators and AIR staff to score using the system. Ensured all product specifications were met. Was accountable for the project schedule, quality of deliverables, budget performance, and scope changes.

Employment History

2010–Present	Vice President, Assessment Programs and Client Services, American Institutes for Research, Washington, DC
2008–2010	Senior Program Director, American Institutes for Research, Washington, DC
2003–2008	Senior Director, Primary Law Content Development, United States Legal Markets, LexisNexis, Charlottesville, VA
2000–2003	Project Manager, Editorial and Publishing Operations, LexisNexis, Colorado Springs, CO

1999–2000	Senior Legal Editor, Case Law Enhancement Operations, LexisNexis, Colorado Springs, CO
1997–1999	Legal Analyst, Professional Publications, LexisNexis, Charlottesville, VA
1996–1997	Tax Examiner, Internal Revenue Service, Austin, TX

Selected Publications

Heather Hayes et al, *Toward a Model Expert Witness Act: An Examination of the Use of Expert Witnesses and a Proposal for Reform*, 80 Iowa L. Rev. 1269 (1995).

Professional Licenses

Member, Iowa State Bar Association (1996).

THOMAS D. GLORFIELD, B.F.A.
PROGRAM MANAGER

Education

B.F.A. 1995, Fine Arts
Iowa State University, Ames, IA

Professional Credentials and Certifications

Project Management Professional (PMP), Project Management Institute, 2005

Present Position

Senior Program Manager, American Institutes for Research (AIR) (2012–Present)
Manages, delivers, and executes the operational components of assessment programs. Works with various internal and external teams to assist in the gathering of requirements, development, and oversight of online testing programs, test booklet and answer document production, administration manuals and other support documentation, training, and report distribution. Assists in conducting training and webinars for stakeholders in various states (Michigan, California, and Delaware). Manages subcontractors to ensure timely production and distribution of additional printed materials. Consults with external clients to determine overall program requirements and develop a plan and schedule to ensure that the program is performing in accordance with operational and financial goals.

Additional Professional Experience

Director, Program Services, Hawaii Department of Education, Hawaii State Alternate Assessment, Keystone Assessment (2010–2012)
Managed, delivered, and executed the developmental and operational components of the Hawaii State Alternate Assessment program. Served as the main liaison between the Hawaii Department of Education and Keystone Assessment as well as multiple subcontractors. Created and maintained program schedules, conducted customer status and planning meetings, and provided program documentation. Assisted in conducting content training for various Hawaii stakeholder groups both on-site and via webinars and conference calls. Managed the Keystone Hawaii Program team, which consisted of project managers.

Program Manager, New Jersey Department of Education, New Jersey State Assessments, Pearson (2004–2010)
Provided leadership in the areas of program planning, development, and execution for the New Jersey Grade Eight Proficiency Assessment (NJ GEPA), the New Jersey Alternate Proficiency Assessment (NJ APA), and the New Jersey Assessment of Skills and Knowledge Grades 3 and 4 (NJ ASK 3–4) programs. Responsible for the financial planning and development of all programs and for managing the daily budget. Coordinated and managed cross-functional teams as well as subcontractor relationships. Project budgets ranged from $3 million to $16 million annually. Managed the Pearson New Jersey Program team, which consisted of project managers and project coordinators.

Program Manager, New York State Education Department, New York State Assessments, Pearson (2008–2010)
Managed, delivered, and executed the operational components of various contracts in support of the New York State Testing Program. Responsible for the financial planning and development of assigned programs and for managing the daily budget. Coordinated and managed cross-functional teams and served as the main liaison between Pearson and subcontractors.

Project Manager, Various State Programs, ePEN, Pearson (2001–2004)
Main point of contact between state customers and the internal development team. Responsible for gathering requirements, creating design documentation, and managing scoring software upon release. Created training materials and conducted training sessions in person and via webinars. Developed a technical support call center process for domestic and international programs.

Employment History

2012–Present	Senior Program Manager, American Institutes for Research
2010–2012	Director, Program Services, Keystone Assessment
2004–2010	Program Manager, Pearson
2001–2004	Project Manager, Pearson

CYNTHIA S. BENIS, B.S.
CUSTOMER SERVICE MANAGER

Education

B.S. 1981, Marketing
The Ohio State University, Columbus, OH

Present Position

Manager, Client Service Center, American Institutes for Research (AIR) (2008–Present)
Manager, Client Service Center, American Institutes for Research (AIR) (2008–Present) Manages AIR's help desk knowledge bases, inquiry tracking systems, and telephone and electronic communications systems. Oversees full-day support across five time zones. Responsible for the recruitment, hiring, and daily activities of the AIR customer help desk staff; trains, mentors, and coaches staff on all policies and procedures, including compliance with established service-level agreements. Provides direct oversight and management, using standard help desk metrics to manage work flow. Prepares customer service quality plans, surveys, and reports and creates call monitoring standards. Evaluates department performance and communicates daily operational activity to client and staff. Provides a technology-based help desk that responds to nontechnical users of AIR's testing systems. Ensures client agreements are met.

Project Experience at AIR

Manager, Client Service Center, for the following help desk operations: Utah Statewide Computer Adaptive Assessment System, Utah Department of Education (2012–present), Pilot and Field Test Administration, Smarter Balanced Assessment Consortium and leads help desk operations on the Smarter Balanced pilot test across 20 states, including many states with no online testing experience (2011–present), Minnesota Comprehensive Assessments (MCA) and Minnesota Test of Academic Skills (MTAS), Minnesota Department of Education (2010–2014), Delaware Comprehensive Assessment System (DCAS) and Delaware Comprehensive Assessment System-Alternate 1 (DCAS-Alt1), Delaware Department of Education (2009–present), Hawaii State Assessment (HSA) and Hawaii State Alternate Assessment (HSA-Alt), Hawaii Department of Education (2009–present), New Mexico Alternate Performance Assessment (NMAPA), New Mexico Public Education Department (2008–present), South Carolina Alternate Assessment (SC-Alt), South Carolina Department of Education (2008–present), Oregon Assessment of Knowledge and Skills (OAKS), Oregon Department of Education (2008–present), the Ohio Achievement Assessments (OAA), Ohio Graduation Tests (OGT), Ohio Test of English Language Acquisition (OTELA), Ohio Alternate Assessment for Students with Significant Cognitive Disabilities (AASCD), and pilots for the Ohio Computer-Based Assessment (OCBA), Ohio Department of Education (2008–present).

Additional Professional Experience

Customer Service Team Manager, Employee Benefit Management Corporation, Dublin, OH (2004–2007)
Managed the customer service team to ensure that efficient workflows and quality initiatives were met or exceeded. Monitored calls to observe the team's technical accuracy and problem-solving skills. Compiled work volume statistics to measure productivity and allocate resources.

Maintained records of customer service requests and concerns. Populated the intranet website to assist staff with FAQs.

Customer Service Manager, CompManagement Health Systems, Dublin, OH (2001–2004)
Analyzed daily and monthly productivity reports for quality and accuracy. Improved automated call distribution abandonment rate from 20% of incoming calls to less than 1% by staffing according to peak/high-volume times. Helped accelerate average speed-of-call answer from more than one minute to less than five seconds.

MIS Requirements Researcher, Netsmart Technologies, Dublin, OH (2000–2001)
Facilitated development and implementation of a new clinical application, electronic Clinical Expert Technology (eCET), a graphical interface with the existing software. Worked with programmers and customers to ensure that designs and test plans met requirements. Duplicated and documented problems and escalated them to programmers for resolution.

Customer Support Manager, Netsmart Technologies, Dublin, OH (1998–2000)
Directed all customer support functions for user-definable software used by clients in the behavioral health industry. Supervised two team leaders, 20 applications analysts, and an administrative assistant. Interacted with the Implementation Group and Research and Development Department to resolve customer service issues.

Team Leader, MIS Support, Netsmart Technologies, Dublin, OH (1992–1998)
Supervised 15 applications analysts who resolved software and operational problems with user-definable software from clients nationwide. Developed teams to specialize in financial, clinical, and state programs to better meet client needs. Wrote information for the technical support database; modeled improved efficiency in resolving common or recurring application problems.

Applications Analyst I, MIS Support, Netsmart Technologies, Dublin, OH (1989–1992)
Provided application and technical support to customers and employees; analyzed and documented problems and reported them to R&D. Installed, configured, and tested hardware, cabling, and software; conducted end-user training; and performed system upgrades and enhancements.

Employment History

2008–Present	Customer Service Manager, American Institutes for Research, Columbus, OH
2004–2007	Customer Service Team Manager, Employment Benefit Management Corporation, Dublin, OH
2001–2004	Customer Service Manager, CompManagement Health Systems, Dublin, OH
2000–2001	MIS Requirements Researcher, Netsmart Technologies, Dublin, OH
1998–2000	Customer Support Manager, Netsmart Technologies, Dublin, OH
1992–1998	Team Leader, MIS Support, Netsmart Technologies, Dublin, OH
1989–1992	Applications Analyst I, MIS Support, Netsmart Technologies, Dublin, OH

DR. GARY PHILLIPS, PH.D.

VICE PRESIDENT AND CHIEF PSYCHOMETRICIAN

Education

Ph.D.	1983, Applied Statistics and Psychometrics University of Kentucky, Lexington, KY
M.S.	1977, School Psychology West Virginia College of Graduate Studies, Huntington, WV
B.A.	1971, Psychology and Philosophy West Virginia State College, Institute, WV

Present Position

Vice President and Chief Scientist, Psychometrics, American Institutes for Research (AIR) (2003–Present)

Provides senior-level guidance on all statistical and psychometric techniques and procedures used in assessment activities—item response theory calibrating and equating, classical statistical analysis, differential item functions, standard setting, reliability and validity studies, generalizability studies, item banking, distributional projections, and value-added studies involving hierarchical linear models and structural equation models.

Additional Professional Experience

Acting Commissioner, National Center for Education Statistics, U.S. Department of Education (1999–2002)

Also served as deputy commissioner overseeing the National Assessment of Educational Progress (NAEP), the National Adult Literacy Survey (NALS), and the Third International Mathematics and Science Study (TIMSS). Served as the architect and as the executive director of President Clinton's Voluntary National Test (VNT). Appointed to the Senior Executive Service (SES) by the Secretary of Education in 1990.

Employment History

2003–Present	Vice President and Chief Scientist, Psychometrics, , American Institutes for Research, Washington, DC
1999–2002	Acting Commissioner, National Center for Education Statistics, U.S. Department of Education
1998–1999	Deputy Commissioner, National Center for Education Statistics, U.S. Department of Education
1991–1997	Associate Commissioner, Education Assessment Division, National Center for Education Statistics, U.S. Department of Education
1986–1990	Branch Chief, Education Assessment Branch, U.S. Department of Education

1980–1986	Branch Chief, Research and Evaluation Branch, Maryland State Department of Education
1977–1983	Adjunct Instructor, West Virginia State College, West Virginia College of Graduate Studies, University of Kentucky, University of Maryland
1971–1975	Social Worker, Kanawha County Department of Welfare, Charleston, West Virginia

Selected Publications

Phillips, G. W. (Under review). *Linking NAEP to NAAL: Adult literacy using high school standards in mathematics and reading.* Washington, DC: American Institutes for Research.

Phillips, G., & Dossey, J. (2008). *Counting on the future: International benchmarks in mathematics for comparing cities and nations.* Washington, DC: American Institutes for Research.

Ferrara, S., Phillips, G., Williams, P., Leinwand, S., Mahoney, S., & Ahadi, S. (in press, 2008). Vertically articulated performance standards: An exploratory study of inferences about achievement and growth. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting.*

Phillips, G. W. (2007). *Chance favors the prepared mind: Mathematics and science indicators for comparing states and nations.* Washington, DC: American Institutes for Research.

Phillips, G. W. (2007). *Expressing international educational achievement in terms of U.S. performance standards: Linking NAEP achievement levels to TIMSS.* Washington, DC: American Institutes for Research.

Phillips, G. (1994). *Methods and issues in setting performance standards.* In A. C. Tuijnman & T. Neville Postlethwaite (Eds.), *Monitoring the Standards of Education.* Pergamon.

Phillips, G. W. (1986). *The NAEP perspective. The National Assessment of Educational Progress and the Longitudinal Studies Program: Together or apart.* Washington, DC: U.S. Department of Education and U.S. Government Printing Office.

Phillips, G. (1985). Word and number assessment inventory (review). *The ninth mental measurement yearbook* (pp. 1768–1769). Lincoln, NE: The Buros Institute of Mental Measurements.

Ahmet Turhan, Ph.D.
Lead Psychometrician

Education

Ph.D.	2006, Educational Measurement and Statistics Florida State University, Tallahassee, FL
M.A.	1997, Educational Measurement and Statistics University of Iowa, Iowa City, IA
B.A.	1991, Testing and Evaluation in Education Hacettepe University, Turkey

Present Position

Lead Psychometrician, American Institutes for Research (AIR) (2016–Present)
Manage, plan, and support form development, independent and operational field-testing for Connecticut, West Virginia, and Washington Next Generation Science Standards (NGSS). Conducted comparability study for Washington science online and paper versions. Established AIRCore scale by linking various states' data. Planned and conducted Ohio bridge study between Ohio Test of English Language Acquisition (OTELA) and Ohio English Language Proficiency Assessment (OELPA). Serve as consultant for Scholastic ReadingPRO computer adaptive test development. Carried out operational and field-test equating for Washington Measurements of Student Progress (MSP) and End-of-Course (EOC) tests.

Additional Professional Experience

Senior Research Scientist, Pearson Assessment and Instruction, Pearson Inc. (2011–2016)
Served as the psychometrician for the Partnership for Assessment of Readiness for College and Careers (PARCC). Developed and presented test construction workshop for internal content experts. Led the entire test construction process. Conducted calibration, equating, and scaling using the 2-parameter logistic (2-PL) model for multiple-choice items and generalized partial credit model (GPCM) for all score categories. Provided expertise on the parallel forms construction. Served as the lead psychometrics research scientist for the two major Florida assessment programs: The Florida Comprehensive Assessment Test (FCAT) 2.0 and Florida EOC Assessments.

Research Scientist, Pearson Assessment and Instruction, Pearson Inc. (2006–2011)
Provided psychometric, technical, and customer support for Florida and Texas contracts. Responsibilities included construction of data analysis specifications, development and implementation of sampling designs, development of Texas Assessment of Knowledge and Skills (TAKS) and Texas English Language Proficiency Assessment System (TELPAS) vertical scaling, development of procedures to confirm integrity of assessment data, interpretation and communication of research analyses results, test development, data collection and analysis, scaling, equating, and generation of technical manuals. Conducted/participated in meetings for data review, standard setting, and technical advisory committee.

Senior Psychometrician, Psychometrics, Florida Department of Education (2002–2006)
Conducted ad hoc research studies to determine trends of statewide large-scale test data. Edited FCAT technical reports and documents such as Test Construction Specifications and Calibration Specifications. Wrote SAS macros to convert hierarchical flat text files into datasets to confirm accuracy of scoring data. Consulted with other state education departments for quality assurance. Provided extensive feedback on item bank functionality and structure. Wrote control files, conducted calibrations and equating, and interpreted results for both multiple-choice and constructed-response items.

Intern and Data Analyst, Psychometrics, Florida Department of Education, (2000–2002)
Wrote SAS codes for demographic reports. Conducted data quality checks.

*Research Assistant****, Florida Department of Education (1999–2001)***
Trained other research assistants in statistical methods. Created specifications for data analysis and model runs with Mplus and Mixor. Conducted missing data analysis. Assisted Leon County School Board on large scale assessment logistics. Created practice test forms. Performed ad hoc research. Conducted extensive literature review on vertical equating for Department of Education.

*Curriculum Resource Center Assistant****, Florida State University (1998–1999)***
Managed curriculum resource center and maintained library.

Research Assistant, ***University of Iowa (1996–1997)***
Designed computerized data collection surveys and exams with Hypercard. Conducted data mining. Conducted factor analyses and structural equation modeling. Collaborated with other research assistant on administration of computerized data collection.

Employment History

2016–Present	Psychometrician, American Institutes for Research, Washington, DC
2011–2016	Senior Research Scientist, Pearson Assessment and Instruction, Pearson Inc., Pflugerville, TX
2006–2011	Research Scientist, Pearson Assessment and Instruction, Pearson Inc., Pflugerville, TX
2002–2006	Senior Psychometrician, Psychometrics, Florida Dept. of Education, Tallahassee, FL
2000–2002	Intern and Data Analyst, Psychometrics, Florida Dept. of Education, Tallahassee, FL
1999–2001	Research Assistant, Florida State University, Tallahassee, FL
1998–1999	Curriculum Resource Center Assistant, Florida State University, Tallahassee, FL
1996–1997	Research Assistant, University of Iowa, Iowa City, IA

Professional Affiliations

American Educational Research Association (AERA), National Council on Measurement in Education (NCME)

Technical Skills

IRTPRO, SAS, SPSS, Excel, R, BILOG, MULTILOG, WINSTEPS

SELINA TOLOSA, M.B.A.

VICE PRESIDENT, ASSESSMENT TECHNOLOGY SERVICES AND SOLUTIONS

Education

M.B.A.	Strategic Management, Marketing and Public Policy Wharton School of Business, University of Pennsylvania, Philadelphia, PA
B.S.	Business Management (cum laude) Ateneo de Manila University, Quezon City, Philippines

Present Position

Vice President, Assessment Technology Services and Solutions, American Institutes for Research (AIR) (2007–Present)
Leads, manages, and supervises a software development team of 80 personnel for the Assessment Division. The team includes software engineers, web graphic designers, testers, project managers, system engineers, data analysts, and technical writers. Ensures all assessment software can effectively provide test development, online and paper-based services, and scoring and reporting of large-scale high-stakes assessments. These systems include the following: a web-based, secure Test Delivery System that administers fixed or computer-adaptive tests and scores them dynamically (supports Windows, Macs, and Linux platforms); machine-scorable constructed-response questions; a quality management system that automatically reviews each test before it is reported; a web-based learning management system that allows teachers and students to organize and use a wide range of curricular materials, including identifying learning prerequisites; web-based reports that provide immediate results at every level, track student progress, and tie testing back to learning; user registration system and web services that regularly extract student data from district and school systems; and an item banking system that enables test development and customized reviews and workflows.

Project Experience at AIR

Directs the CSSC and manages the Information Technology Administration (2007–present). Successfully managed software development projects as the Senior Project Manager and Quality Assurance Lead for Computer and Statistical Sciences Center (2005–2007).

Employment History

2007–Present	Vice President, Assessment Technology Services and Solutions, American Institutes for Research, Washington, DC
2005–2007	Senior Assessment Management Specialist, Computer and Statistical Sciences Center (CSSC), American Institutes for Research, Washington, DC
2003–2004	Director of Information Services, New American Schools, Alexandria, VA
2003–2004	Financial Officer, New American Schools, Alexandria, VA
2002–2003	Senior Associate, New American Schools, Alexandria, VA
2001	Summer Associate, Ball Foundation, Glen Ellyn, IL
1999–2000	Part-Time Professor, Ateneo de Manila University, Quezon City, Philippines

Professional Affiliations

Co-Chair, International Admissions Committee, Wharton School of Business, University of Pennsylvania (2001–2002)

Executive Editor, eBusiness Review, Wharton School of Business, University of Pennsylvania (2001–2002)

Technical Skills

Visual Basic, Ruby, Standard Query Language, Java, C

SONJA HUBBARD, M.S.
TECHNOLOGY CONSULTANT

Education

M.S.	2012, Information Systems Technology George Washington University, Washington, DC
B.A.	2002, Media Studies Pomona College, Claremont, CA

Professional Credentials and Certification

CIO Certificate in Federal Executive Competencies (2012)

Honors and Awards

Technical Leadership Award, AED, 2008
Innovation Award, AED, 2007
Staff Advisory Council Employee of the Year Award, AMIDEAST, 2004

Present Position

Software Project Manager, Student Registration, American Institutes for Research (AIR) (2013–Present)
Develops requirements, budgets, and schedules for software development projects. Has day-to-day responsibility for ensuring that projects are completed on time and within budget, and that all deliverables are of the highest quality. Facilitates requirements meetings and status meetings with clients and project team. Establishes milestones, identifies potential issues, and monitors adherence to project scope, requirements and design documents, schedules, and to the Computer Science and Statistical Center (CSSC) software development process. Maintains client relations and communications including status reporting and informal conversations.

Project Experience at AIR

Test Information Distribution Engine (TIDE) (2013–Present)
Serves as the primary point of contact for four clients as well as overseeing special development projects that affect multiple clients. Coordinates service requests across multiple CSSC teams. Provides product management support for open-source single-sign-on implementation. Researches escalated help desk cases by analyzing data, code, and application logs. Applies a user-centered approach to new-feature design decisions. Develops and manages the TIDE internal research and development schedule. Identifies and implements process improvements related to issue tracking and knowledge management.

Additional Professional Experience

Technical Project Manager, Sonjara, Inc. (2010–2013)
Managed several multi-year web application development projects including SeaPerch, STEM2Stern, the International Tax Policy Forum, and the IRS Tax Forums. Clients included the

Office of Naval Research, USAID, PricewaterhouseCoopers, and the Society for Naval Architects and Marine Engineers. Created and managed work plans, risk assessments, and budgets. Acted as primary point of contact for clients, developed understanding of clients' business processes and challenges, proposed and obtained approval for follow-on work. Provided guidance on best practices to project managers and developers. Implemented improvements in project management across the company, formalized and documented procedures, and configured enterprise-level tools to support project management. Analyzed and documented project requirements, including workflows, data models, business rules, and user experience considerations. Deployed code to production environments. Planned, managed, and performed software testing. Implemented change requests and bug fixes using PHP, SQL, HTML, and CSS. Provided superior technical support to clients via phone, email, and on-site visits. Advocated for users and implemented user experience improvements based on user research and usability analysis. Wrote technical documentation and user manuals and trained clients on using web-based content management systems and collaboration tools.

Technical Manager, AED (2005–2010)
Oversaw development of the J2EE-based Global Learning Portal. Managed teams of U.S.-based and offshore designers, developers, testers, and system administrators. Hired and supervised junior technical staff. Gathered and documented requirements for new features and change requests. Tested and deployed functional and user interface enhancements. Established and managed web project plans and budgets. Designed and delivered training to staff and end users on online content management tools and techniques. Traveled to Africa and the Middle East for requirements gathering and technical consulting and training activities with field-based teams. Identified and implemented process improvements to streamline enhancements in web content and functionality. Projects included the Afghanistan Higher Education Portal and the Palestine Youth Portal.

Program Officer, AMIDEAST (2004–2005)
Oversaw web content for multiple sections of the corporate website. Served as liaison between program departments and the IT department. Coordinated operations and marketing for educational testing programs at 13 field offices in the Middle East and North Africa.

Senior Program Assistant, AMIDEAST (2002–2004)
Coordinated design and content for web and print materials. Updated web content, standardized existing HTML and CSS code to comply with organization standards.

Webmaster, Freelance (1999–2002)
Designed, developed, and maintained web sites for several university departments, conferences, news media organizations. Prepared print materials for online delivery and created original online content, graphics, and photographs.

Employment History

2013–Present	Software Project Manager, American Institutes for Research, Washington, DC
2010–2013	Technical Project Manager, Sonjara, Inc., Falls Church, VA
2005–2010	Technical Manager, AED, Washington, DC
2004-2005	Program Officer, AMIDEAST, Washington, DC
2002-2004	Senior Program Assistant, AMIDEAST, Washington, DC
1999-2002	Webmaster, Freelance, Claremont, CA

SABAPATHY KARUNANITHI, M.S., M.B.A.
SOFTWARE PROJECT MANAGER, STUDENT REGISTRATION

Education

M.B.A.	2013, Management, Information Systems and Data Analytics University of Maryland, Robert H. Smith School of Business, College Park, MD
M.S.	2007, Electrical Engineering University of Bridgeport, Bridgeport, CT
B.S.	2005, Electronics and Communication Engineering Anna University, India

Professional Credentials and Certifications

SAS Certified Base Programmer, 2008
SAS Certified Advanced Programmer, 2008

Present Position

Technical Project Manager, American Institutes for Research (AIR)
Manages the Test Information Distribution Engine (TIDE), the user and student registration system. Manages and coordinates the collaboration of data analysts, software developers, database administrators, network engineers, and quality assurance professionals to achieve on-time project deployment, monitor project performance, and continuously and successfully meet client expectations.

Project Experience at AIR

Supports student registration systems for assessment projects in Delaware, Hawaii, Minnesota, Ohio, Oregon, Utah, and the Smarter Balanced Assessment Consortium (2013–present).

Senior Statistical Data Analyst, American Institutes for Research (AIR) (2008–2012)
Worked on data analysis components of the value-added model projects based on generalized linear regression models. Wrote several generic and very efficient SAS macros to put the data files through various data cleaning and manipulation steps and score reporting procedures. Has extensive experience working on large databases, various SAS/SQL procedures, and efficient data analysis and techniques using SAS, SQL, and R statistical packages.

Employment History

2013–Present	Technical Project Manager, American Institutes for Research, Washington, DC
2008–2012	Senior Statistical Data Analyst, American Institutes for Research, Washington, DC
2006	Data Analyst, RGS Technologies, Singapore
2005	Data Analyst, Gramar Infotech, Chennai, India

Eyal Moses, Ph.D.

Senior Application Developer, Student Registration

Education

Ph.D.	2002, Applied Mathematics and Engineering Tel Aviv University, Israel
M.S.	1997, Engineering Tel Aviv University, Israel
B.S.	1994, Engineering Technion Institute of Technology, Israel

Present Position

Senior Software Engineer, American Institutes for Research (AIR) (2013–Present)
Develop methods for optimized data retrieval using SQL Server Database. Develop algorithms for efficient data storage and fast queries. Perform research on methods for fast data retrieval in the Online Reporting System (ORS) team. Design and implement scalable multi-server architecture for statistical computation. This architecture is utilized by the ORS to compute in a fraction of a second statistical measures such as proficiency and average for large sets of millions of students.

Additional Professional Experience

Algorithms Developer Lead, Elbit Systems Ltd. (2005–2012)
Conducted research in expert systems, neural networks, Bayesian networks, decision trees, clustering, time series and regression for prediction models, unsupervised anomaly detection, portfolio analysis, behavior segmentation, document classification, social-networks analysis, and modus-operandi pattern discovery. Developed algorithms for resource utilization optimization for a satellite mission planning system. Designed and implemented an easy-to-use, scalable, Web-based application for data mining and statistical analysis using SQL Server 2012 (SSAS, SSIS), C#.NET, ASAP.NET, PMML models, DMX, and ADO.NET.

Algorithms Developer Lead, Lenslet Labs, Ltd. (2000–2005)
Worked as a technical project manager for an optical processor designed for large-scale linear transformation (vector matrix multiplication) for various applications including bioinformatics sequencing and wireless communication.

Principal Engineer, IDF Technological Unit (1994–2000)
Developed multiple methods of discovering hidden patterns in extremely large-volume data for the purpose of identifying potential threats.

Employment History

2013–Present	Senior Software Engineer, American Institutes for Research, Washington, DC
2005–2012	Algorithms Developer Lead, Elbit Systems Ltd., McLean, VA

2000–2005 Algorithms Developer Lead, Lenslet Labs, Ltd., Israel

1994–2000 Principal Engineer, IDF Technological Unit, Israel

Publications

Moses, E., Weitman, G., Glaser, A. (2010). Method and apparatus for distributing assignments. *U.S.Patent No*. 7788199. http://www.freepatentsonline.com/7788199.html

IL141314: Optimization Methods of Targets Derivation for Satellite Acquisition

141/04996: Optical Multiplier with Coherent Feedback

Goren, A., Sariel, A., Levit, S., Asaf, Y., Liberman, S., Sender, B., Tzelnick, T., Hefetz, Y., Moses, E., Machal, V. (2004). Vector-matrix multiplication. *U.S. Patent No*. 20040243657. http://www.freepatentsonline.com/y2004/0243657.html

Gordon, T., & Moses, E. (2012). Devices for Simultaneous Egress System from High-Rise Buildings. *U.S. Patent No.* 8151940 B2. http://www.google.com/patents/US8151940

Standard Editing: ASTM E-2513-07, Standard Specification for Multi-Story Building External Evacuation Platform Rescue Systems

Ryvkin M., Fuchs M., Lipperman, F. & Moses E. (2004). Non-homogenized approach to the analysis of periodic system: topology optimization. *Continuum Models and Discrete Systems* (pp.129–134). Kluver Academic Publishers.

Moses, E., Fuchs, M. B., & Ryvkin, M. (2002). Topological design of modular structures under arbitrary loading. *Structural and Multidisciplinary Optimization*, *24*, 407–417.

Moses, E., Ryvkin, M., & Fuchs, M. (2001). A FE methodology for the static analysis of infinite periodic structures under general loading. *Computational Mechanics*, *27*, 369–377.

Fuchs, M. B. & Moses, E. (2000). Optimal topologies with transmissible loads. *Multidisciplinary Optimization*, *19*, 263–273.

Technical Skills

More than 10 years of industrial experience in various applications for designing, modeling, analyzing, implementing, and testing complex systems in defense, law enforcement, remote sensing, and optical computing.

Ability to define and develop methods in data mining, text analytics, relational database modeling, ETL process, and machine-learning for high-volume datasets.

Well-versed in developing and optimizing high-performance distributed systems.

Computer Languages: C/C++ (Windows/Unit), C#.NET, SQL, MPI.

Computer Frameworks: Matlab, Simulink, Microstrategy, SQL Server, SSAS/SSIS, Oracle Data-Mining, Rapid-Miner, IBM-SPSS, IBM-Cplex.

SIRISHA NAGABHAIRAVA, M.S.
SOFTWARE PROJECT MANAGER, ITEM BANK

Education

M.S.	Computer Science, 2004 Fairleigh Dickinson University, Teaneck, NJ
B.S.	Computer Science, 2002 Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

Present Position

Technical Director, American Institutes for Research (AIR) (2016–Present)
Responsible for developing and tracking requirements and schedules for software projects, including: authoring requirement gathering and analysis, facilitating requirements meetings through reviews and approvals, supporting peer review sessions, and supporting status meetings with external software vendors. Develops and writes software specifications to support work performed by an external software vendor. Regularly communicates with software engineering teams and external software vendor on status and progress regarding design, reliability, and maintenance. Leads integration analysis of requirements that impact several systems. Tracks progress toward milestones, anticipates issues, and monitors adherence to project scope, requirements and design documents, schedules, and to the software development process. Through daily monitoring of issues and new work, identifies and escalates complex issues to senior management and monitors project progress and milestones. Performs data mining based on ad hoc data requests from the development team.

Additional Professional Experience

Senior Manager, New product Development, ATPCO (2010–2016)
Led the end-to-end product strategy and development for multiple highly complex ATPCO products that resulted in increased revenue and profit, widened customer-base, and improved customer satisfaction. Launched an improved Airline Fares and Airline Rules offering using cutting edge technologies that transformed the product into a revenue generator by broadening its appeal to the target market and increasing customer spending. Led a large cross-functional and matrixed team to launch the largest product improvement in five years. Worked closely with senior leaders to craft a three-year strategy for the product, then led multiple teams across the business to create and launch the new offering. The product improvement exceeded growth and revenue targets by 150%. Led the exploration of several innovative and cutting-edge products and services that helped maintain company's status as market leader. Adept at facilitating discussions between companies with opposing business drivers to unify and simplify airline industry processes. Led a series of strategic initiatives for the customer segment that resulted in new market entries, new product offerings, product improvement, and revenue lift. Managed, guided, coached, and motivated multiple partner teams to deliver results above expectations on various metrics. Translated and interpreted the business strategies, and influenced shape and direction of technical team's strategy. As part of the leadership committee of a massive employee engagement exercise, helped craft the donation strategy, communications, and success metrics to deliver 99% employee participation. Served as key member of management team responsible for negotiating the

first multi-year outsourcing agreement for technology services in the airline industry with an annual value of over $50 million.

Technical Project Manager, New product Development, Real Networks (2005–2007)
Delivered integrated mobile entertainment solution called Music On Demand worth $40 million in revenue for wireless carriers (chiefly Verizon). Led full lifecycle projects from problem scoping to delivery, managing clients and coaching analysts and consultants. Conducted in-depth market assessment, competitor analysis, and customer research and segmentation to drive business. Performed complex product design activities requiring extensive research and analysis. Managed a team of product development professionals responsible for delivery of development projects and ongoing support of existing products and customers. Created a four-year strategy for international external partners that reinforced the brand, improved service expectations, and increased customer satisfaction. Participated in planning and projection efforts that shaped the company’s future research and development in the mobile telecommunications field.

Software Architect, Iron Mountain (2003–2005)
Led a team of software developers to design and develop software code using Java JEE technologies to extend the features of the product that enabled end-to-end content creation and other state-of-art system features. Developed a web service on technology escrow services Researched infrastructure requirements for the product, including target environment, performance criteria, and competitive issues. Designed the functional requirements from customer meetings and mapped use cases into design specifications using UML. Installed and deployed the application in a horizontal clustered environment for load-balancing and fail-over capabilities. Created a three-year expansion plan involving product and market expansion for a business unit to vastly improve competitive standing. Completed projects led to corporate-level acquisition deals and expansion of existing products into unexplored markets.

Employment History

2016–Present	Technical Director, American Institutes for Research, Washington, DC
2010–2016	Senior Manager, New Product Development, ATPCO, Dulles, VA
2005–2007	Technical Project Manager, New Product Development, Real Networks, Reston, VA
2003–2005	Software Architect, Iron Mountain, Sterling, VA

Technical Skills

J2EE, Asana, Trello, Basecamp, Microsoft Project, OneNote, Evernote, Azendoo, JIRA, big data, JEE, XML, Visio, XML, JBoss, WAS, Apache, Web services, Oracle, DB2, HTML, cloud computing, Windows and Linux, networks, mobile computing (iPhone and Android)

TEDDY XIONG, M.S.

SENIOR APPLICATION DEVELOPER, ITEM BANK

Education

M.S. 2001, Computer Science Engineering
The Johns Hopkins University, Baltimore, MD

Present Position

Lead Software Engineer, American Institutes for Research (AIR) (2012–Present)
Manages ITS developer team and serves as technical lead. Responsible for leading all web-based application developments through all stages of SDLC for the Item Tracking System.

Project Experience at AIR

Senior Software Engineer, Item Tracking System (2005–2012)
Supports AIR and Department of Education users in building and maintaining configurable item banks for multiple state assessment projects. Coordinates with other software teams to make sure content generated from ITS is handled properly in the Test Delivery System, Quality Assurance System, and Reporting System. Designs major ITS upgrades and enhances ITS by improving processes and software for content development and test publication.

Additional Professional Experience

Application Developer, Computer Aid, Inc. (2003–2005)
Worked on site at the International Monetary Fund as Application Developer. Was responsible for building financial application system for IMF economists: planning, design, code, test, and maintenance. Was also responsible for maintenance of some existing systems. The tools mainly used were .NET, VB, VBA, SQL, ADO, Excel, and Access. Provided both technical expertise and high-quality customer service.

Software Engineer, Esymmetrix, Inc. (2001–2003)
Responsibilities included software design, programming, testing, debugging, and database administration for an end-to-end business system supporting all operations of a parts distribution company.

Specific duties included the following: Analyzed, planned, designed, and created software supporting order entry, picking/packing, purchasing, inventory management, and accounting. Used Microsoft Visio for OOP design, applying knowledge of object-oriented systems, UML, design pattern, and refactoring.

Used Visual Basic 6.0 and Visual SourceSafe to develop a three-tier architecture application with a flexible MDI-based user interface. Used ADO for database access along with design and development of the data storage layer that is represented primarily in stored procedures. Participated in COM+ development including creation of several custom-built ActiveX controls and advanced use of various third-party ActiveX controls for customizable toolbars, sophisticated grids, etc.; created many DLL and EXE projects.

Used ASP, XML/XSL, VBScript, JavaScript, and HTML to import and export information to create online sales orders, receipts, invoices, etc. Used Microsoft SQL for database management; this system includes several hundred tables and stored procedures.

Tutor, Center for Talented Youth at The Johns Hopkins University (2000–2001)
Taught C programming, algebra, and advanced calculus via web technologies such as white board at the Center for Talented Youth, which provides online community education for talented youth throughout the world.

Employment History

2012–Present	Lead Software Engineer, Item Tracking System, American Institutes for Research (AIR), Washington, DC
2005–2012	Senior Software Engineer, Item Tracking System, American Institutes for Research (AIR), Washington, DC
2003–2005	Application Developer, Computer Aid, Inc., Alexandria, VA
2001–2003	Software Engineer, Esymmetrix, Inc., Dayton, MD
2000–2001	Tutor, Center for Talented Youth at The Johns Hopkins University

Professional Certifications

Microsoft Certified Database Administrator (MCDBA)
Microsoft Certified System Engineer (MCSE)
Oracle9i Database Administrator Certified Associate (Oracle OCA)

Skills

.NET, VB, VBA, C/C++, JAVA, JavaScript, Jquery, Ajax, MVC, SQL, ADO, ASP, XML, XP/UNIX/LINUX, object-oriented systems, database systems

S. CHRISTIAN REDMOND, M.S., PMP

SOFTWARE PROJECT MANAGER, TEST DELIVERY

Education

M.S.	2012, Software Engineering Management Carnegie Mellon University, Pittsburgh, PA
B.A.	1992, Psychology University of Maryland, College Park, MD
B.A.	1991, English University of Maryland, College Park, MD

Professional Certifications

PMP, Project Management Institute (2008)

SEI Certificate, Software Process Improvement Management

Additional training from the Software Engineering Institute (SEI): Introduction to CMMI Version 1.2, Documenting Software Architectures, Software Architecture Design and Analysis

Present Position

Software Project Manager, American Institutes for Research (AIR) (2012–Present)
Manages web projects across different stages of the development life cycle. Responsible for configuring and delivering Test Delivery System (TDS) in 20 states including for Smarter Balanced assessments. Serves as the primary liaison to clients and the rest of the project team. Responsible for the requirements document and the final test plan, working with the developers, quality assurance team, clients, and senior program manager to ensure that the requirements are clear, accurate, and sufficient and that the test plan adequately addresses all requirements.

Employment History

2012–Present	Software Project Manager, American Institutes for Research, Washington, DC
2010–2012	Project Manager, Latitude/Fig Leaf Software, Vienna, VA
2011–2012	Team Lead, Graduate School Project, Pitt Ohio, Pittsburgh, PA
2010	Consultant, Federal Reserve Board, Washington, DC
2009–2010	Consultant, National Education Association, Rockville, MD
2005–2009	Developer and Technical Administrator, Capitol Information Group, Falls Church, VA
2003–2005	Developer, L&E Meridian, Springfield, VA
2003–2005	Consultant, Adams Hussy and Associates, Arlington, VA
2003	Database and CRM Administrator, Northern Virginia Family Service, Oakton, VA
1999–2003	Database Administrator/Developer, Adams Hussy and Associates, Arlington, VA

Technical Skills

Methodologies: Object-Oriented Design, Agile, CMMI, PMP. **Applications:** Visual Studio, Visio, Microsoft Project, Base Camp, Jira, Crystal Reports. **Programming:** .NET, PHP, Python. **Databases:** Microsoft SQL Server, Oracle, MySQL.

ALAN REEVE, M.S.
SENIOR APPLICATION DEVELOPER, TEST DELIVERY

Education

M.S.	1992, Computer Science Northern Illinois University, DeKalb, IL
B.S.	1991, Computer Science, Theoretical Emphasis Northern Illinois University, DeKalb, IL

Present Position

Lead Software Engineer I, American Institutes for Research (AIR) (2015–Present)
Responsible for maintaining AIR's test delivery system (TDS) and enhancing the customer experience. Maintain several of the existing item types such as the rich text editor. Have developed and maintain many of the other assistive technologies used on AIR's TDS platform including reading mode and the text-to-speech module.

Project Experience at AIR

Senior Software Engineer II, AIR (2014–2015)
Developed several test item types including Gap Match and Graphic Gap Match items according to the question and test interoperability (QTI) specification.

Professional Experience

Owner, REEVEsoft, Inc. (2002–2013)
Started and operated REEVEsoft, Inc., which provided a web-based content management system designed for real estate offices and agents. Managed interfacing to 25+ multiple listing services to ensure that nightly property listing data, photographs, and virtual tours were available for searching on client operated, public-facing websites. Also solely responsible for maintaining the CompuAgent product allowing real estate professionals to edit their websites in real time, as well as providing the front-facing engine that customers would access. Created and monitored the company's server infrastructure and principally used ASP.NET, C#, JavaScript, SQL Server, IIS and Mail Enable to develop and host the company's product offerings.

Temporary Instructor, Elgin Community College (2013–2013)
Taught 20 credit hours of college level classes in the Fall 2013 semester. Responsible for developing course work, lecturing, assisting students with hands on programming projects, creating quizzes/tests, and grading. Taught CIS I and II (C++) and C#/.NET programming.

Senior Software Engineer III, Radiant Systems (2000–2002)
Worked on the QSR team on their KPS (Kitchen Production System) software written in C++ on a proprietary Windows CE based device. Lead the Culvers/Cousins project and was directly involved in the Disney World/McDonalds project which was the highest volume fast food site in North America at the time.

Employment History

2015–Present	Lead Software Engineer I, American Institutes for Research, Washington, DC
2014–2015	Senior Software Engineer II, American Institutes for Research, Washington, DC
2013–2013	Temporary Instructor, Elgin Community College, Elgin, IL
2002–2013	Owner, REEVEsoft, Inc., Marietta, GA and Geneva, IL
2000–2002	Senior Software Engineer III, Radiant Systems, Alpharetta, GA

Professional Affiliations

Association for Computing Machinery

Technical Skills

JavaScript, .NET, C#, and T-SQL, ASP.NET, SQL Server, IIS and Mail Enable, CIS I and II (C++) and C#/.NET

SCOTT WICKETT, B.S., PMP
SOFTWARE PROJECT MANAGER, ANALYSIS SYSTEMS

Education

B.S.	1987, Engineering/Computer Science University of Florida, Gainesville, FL

Professional Credentials and Certifications

PMP Certification, Project Management Institute, 2016

Present Position

AIR Senior Technical Project Manager, American Institutes for Research (AIR) (2016–Present)
Responsible for the client release of the Analysis reporting component of the Assessment product, involving translating customer requirements into software changes and database configuration updates while ensuring a quality product. Current clients responsible for support include Connecticut, New Hampshire, and Vermont.

Responsible for creating successful research and development releases that are delivered on time and within budget. Experience in applying best practices in project management, including those used to initiate, plan, control, monitor, and report on all phases of project activity. Involved in all phases of the software development cycle, developing project plans, requirements documents, and project budgets.

Additional Professional Experience

Product Release Manager/Senior Technical Project Manager, Xerox Services—State & Local Solutions Division (2015–2016)
Senior technical release manager for the Open Payments Fare product line. Oversaw all project processes and deliverables including obtaining customer acceptance, live activation, and initial rollout to the general public for the South Eastern Pennsylvania Transit Authority (SEPTA). Interacted with offshore development and test groups to create a quality product. Web development included C#, ASP.NET, JavaScript, J2EE, XML, SQL server, jQuery, and Rest APIs. Separation of future development into phases, DevOps efforts, and cloud migration. Implemented Product Change Control Boards for greater communication and control of development, test, and production environments. Product release consultant for the Tolling product line (consisting of the EZ-Pass and TXDOT member organizations) to consolidate current software and feature offerings into a single releasable product baseline.

Software Engineering Manager, Xerox Services—Transportation Management Solutions Division (2012–2015)
Release manager of Intelligent Transportation Systems (ITS) software in support of the IVU-3100 embedded software, which was deployed on a fleet of 20,000 buses, light rail, and support vehicles, reducing four versions into one master baseline. Manager of IVU-3100 software department consisting of 12 employees. Restructured the Configuration Control Boards (CCBs) for greater efficiency. Instrumental in structuring a quality management system and obtaining

ISO 9001 certification. Participated in the Capability Maturity Model Integration (CMMI) effort as a subject matter expert in integration. A key member of the management team to shift product focus from continuous maintenance to development that will support new and existing customers.

Management Consultant, OpalStaff (2011–2012)
Software management consultant responsible for transforming the division's software development life cycle processes to increase the profitability of underperforming product lines. Lead software engineer driving the re-productizing of the IVU-3100 embedded vehicle software. Release manager responsible for the planning, configuration, and testing of vehicle software. Interacted with the CCB to define the release contents.

Engineering Manager, ACS (Orbital Sciences Corp)—Transportation Management Systems Division (1998–2010)
Product development manager for Orbital's Transportation Management Systems (TMS). Software product manager for the OrbCAD/UNIX Control Center's Computer Aided Dispatch/Automated Vehicle Locator (CAD/AVL) system. Manager of the engineering/CM and documentation and training departments. Built and configured the CM department structure, standard operating procedures, and work instructions for the design and implementation of division-wide processes and standards for product control in coordination with ISO 9001 and SEI/CMMI standards. Implemented and managed CMMI Maturity Level 3 processes.

Employment History

2016–Present	Senior Technical Project Manager, AIR, Washington, DC
2015–2016	Product Release Manager/ Senior Technical Project Manager, Xerox State & Local Solutions, Germantown, MD
2012–2015	Software Engineering Manager, Xerox Transportation Management Solutions, Columbia, MD
2011–2012	Management Consultant, OpalStaff, Columbia, MD
1998–2010	Engineering Manager, ACS (previously Orbital Sciences Corporation) Transportation Management Systems, Columbia, MD
1997–1998	President and Chief Architect, SWAC Technologies, Frederick, MD
1995–1997	Product Manager/Member of Technical Staff, COMSAT Corporation, Clarksburg, MD
1995–1995	Technical Lead, SAIC Ideas Group
1988–1994	Product Manager, IBM Corporation

Technical Skills

Internet Technology: XML, Java, C#, .net, REST, SOAP, HTML. **Operating Systems and Server Administration and Networking:** Windows OS, UNIX, Linux, Mainframe. **Software Engineering and Tools:** C++, C, Perl, Agile Object-Oriented Development, Visual Studio, Oracle, PowerBuilder, SQL, Siebel, WebSphere, JIRA, Subversion, Rational ClearCase, Rational ClearQuest, TeamCity, Serena, Jenkins tool automation, VISIO, Visual Source Safe.

ADAM MCLAUGHLIN, B.A.

SENIOR APPLICATION DEVELOPER, ANALYSIS SYSTEMS

Education

B.A.	1999, Computer Science, Concentration in Mathematics University of Delaware, Newark, DE

Present Position

Lead Software Engineer II, Computer and Statistical Sciences Center (CSSC), Assessment Program, American Institutes for Research (AIR) (2010–Present)
Responsible for the design, development, and enhancement of software systems for the Assessment program, which offers a range of testing services, including online and paper-based testing, scoring, and reporting. Designed and built support for receiving paper-pencil tests into the online post-delivery systems in order to take full advantage of automated scoring, validation routines, and online reporting. This included an automated issue reporting and resolution processing system to address incorrectly bubbled or incomplete test books. Works for the Analysis group, which handles several of the post-delivery systems in the online testing suite, including the test scoring engine, paper-pencil test processing systems, reporting systems, the rubric editing and validation system, and the Database of Record, which is the central repository for all test data. Developed a new adaptive item selection algorithm designed to improve accuracy and to support additional features. The Analysis group is also responsible for the adaptive item selection algorithm. Work requires a high degree of proficiency with such diverse technologies as C#/.Net Framework, SQL Server/ T-SQL, XML/XSLT/XSD, web services, and Windows services. Redesigned the Database of Record to reduce operational support costs, meet more demanding reporting requirements, and improve throughput. Developed reporting queries for the Learning Point Navigator system and integrated them into the site.

Project Experience at AIR

Supports online systems to support assessment projects in Delaware (2010–present), Hawaii (2010–present), Minnesota (2010–2014), New Mexico (2010–present), Ohio (2010–present), Oregon (2010–present), South Carolina (2010–present), Utah (2012–present), the Smarter Balanced Assessment Consortium (2012–present), and California (2013–present).

Employment History

2010–Present	Lead Software Engineer, CSSC, American Institutes for Research, Washington, DC
2009–2010	Software Engineer, Dell Services Federal Government, Fairfax, VA
2007–2008	Senior Systems Analyst, Blue Cross Blue Shield Association, Chicago, IL
2007	Senior Software Developer, Emmis Interactive, Chicago, IL
2006–2007	Senior Software Developer, Heidrick and Struggles, Chicago, IL
2001–2006	Technical Lead, Retirement Practice Systems Group (RPSG), Watson Wyatt Worldwide, Arlington, VA

2001 Senior Software Developer, Advanced Management Technology, Inc. (AMTI), Arlington, VA

2000–2001 Software Engineer/Developer, Contract with Advanced Management Technology, Inc. (AMTI), Arlington, VA, Triad Management Systems, Inc., Rockville, MD

1999–2000 Software Engineer, ExpressTRAK Baseline Requirements, Analysis, & Design (RAD)/ Software Engineering Solutions (SES), Verizon Communications, Arlington, VA

Technical Skills

Programming Languages/APIs/ Technologies: .Net Framework using C#/VB.Net. ASP.Net, WinForms, ADO.Net, SQL/T-SQL, LINQ to SQL, XML, XSLT, JavaScript, CSS, and D/HTML, ASP, VBScript, WSH, ADO, Visual Basic 6 (COM/+, EXE), Perl/CGI, Lingo (Director, Shockwave), Java. **Training:** ASP.Net Development, OOA/D, Structured System Analysis and Design, JAD, Walkthroughs and Inspections, Requirements Management, OO Development Using VisualWorks Smalltalk. **Applications/Components/Tools:** Visual Studio.Net 2010, SQL Server DBMS, SQL Server Reporting Services (SSRS), SQL Server Integration Services (SSIS), XML Spy, NUnit, TFS, Telerik ASP.Net AJAX controls, Developer Express WinForms components, Data Dynamics Active Reports for .Net, Microsoft Visio, Microsoft Visual Basic 6, Microsoft Project, Rational Robot, Seagate Crystal Reports, Adobe PhotoShop, TortoiseCVS, VSS, PVCS Tracker, FogBugz.

KUSHAL BUDHWAR, M.S.

SOFTWARE PROJECT MANAGER, ONLINE REPORTING

Education

M.S.	2014, Information Systems & Management Heinz College/Carnegie Mellon University, Pittsburgh, PA
B.E.	2007, Computer Science & Engineering R.V College of Engineering/University, Bangalore, Karnataka

Present Position

Senior Technical Project Manager, American Institutes for Research (AIR) (2016–Present)
Manages a team of software engineers, data analysts, database engineers, and deployment engineers working on the online score reporting system. Responsible for managing multiple projects throughout the entire software development process with departments of education in multiple states.

Additional Professional Experience

Business Systems Analyst, SanDisk (Product Management and Operations Strategy) 2014–2016)
Integration Manager, Multiple Third Party Integration Projects, Corporate IT (2014–2016): Managed Hiperos-3PM integration project to streamline the global supplier management process. Led engineering, operations, and CRM teams to implement the new product introduction (NPI) interface. Enhanced the B2B interface, resulting in its seamless integration with the interfaces of 20 contract manufacturers. Responsible for ongoing support of the interface. Developed turnkey interface to enable cost monitoring of turnkey items in real time, resulting in annual savings of $0.5M and 80 man-days.

Internal Operations Execution and Support Applications Manager, Supply Chain Systems (2015–2016): Single point of contact for critical and complicated Agile PLM-SAP interface that handled a daily volume of over ~1,000 master data exchanges between systems. Worked closely with on-site and offshore vendor teams to ensure production issues were resolved within the desired SLA. Involved in planning, prioritizing, and delivering of major system enhancements that needed a quick turnaround. Partnered with the internal operations team to design and develop mission critical reports for planners and buyers.

Approval Automation Project Product Manager, ERP Systems (2015): Managed the entire life cycle of the info record/quota/contract approval project that provides higher visibility to procurement managers on the price changes in the system. Innovated a solution to handle the entire approval/rejection process entirely via mobile/smart devices. Led global rollout (APAC and United States), change management, and cutover activities.

Fusion-IO Merger and Acquisition Integration Techno-Functional Analyst, Procurement (2014): Analyzed the project requirement and led the execution to integrate Fusion-IO into the SanDisk ERP system. Collaborated with logistics team to setup four new plants (one captive and three contract manufacturers) in the system. Coordinated with SAP-SMEs and the Fusion-IO team to perform $300M worth of inventory conversion and migration. Developed and launched innovative analytical reporting tools to help M&A strategy team in migration activities. The system integration project went live within three months and all work was completed in-house.

Senior Software Engineer, Accenture (2007–2013): Client Services Expert, India Domestic Business Consulting (IDB) (2011–2013): Managed setting up of –Technical Centre of Excellence for domestic delivery that resulted in yearly savings of $200K. Implemented logistics and human capital management projects for three clients simultaneously, thereby optimizing resource utilization. Received performance award for stretching beyond expectations. Integrated data feeds from Ariba (a strategic sourcing solution) with SAP R3 directly, thereby eliminating the need of middleware, resulting in savings of $500K. Designed business intelligence and analytical reports for clients, thereby assisting senior management in the decision-making process of a utilities portfolio valued at approximately $5M. Created a reusable asset tool for IDB, reducing duplication of effort and build time of subsequent projects. Collaborated with McKinsey over two strategic IT initiatives and incorporated best practices for a power company

Luanda, Angola Technical SME/Lead, Angola LNG (ALNG) (2010–2011): Handled dual role as a business and technical consultant for the implementation of procurement processes at ALNG. Led the development team on all eight implementation projects of ALNG (energy startup). Contributed to request for proposals, business blue prints, fit gap analysis, and gap estimations. Proposed, designed, and developed travel management workflow that resulted in automation of 1,000+ trips and 6,000+ expense reports that were generated every day at ALNG. Collaborated with senior management to create prototypes for pre- sales activities that resulted in a proposal conversion rate of 90% for Accenture Angola. Managed training, hiring, mentoring, and skill-building of junior consultants (24 local recruits) and clients (800 end users). Facilitated setting up and expansion of Accenture Angola's technology growth practice.

India Application Developer, British Petroleum (BP) (2007–2010): Managed all development activities related to BP integrated supply and trading business (high-priority unit of BP). Conceptualized roadmap for future application architecture and eliminated redundant applications. Automated ethanol business of BP and received prestigious ASE Achievers award for the business impact. Mapped BP loyalty card program into SAP, resulting in saving 90,000 man-hours of work every year. Upgraded the interface architecture, which was the backbone of almost 3,000 interfaces. Coordinated with cross-functional teams in the United States and the United Kingdom to maintain smooth build and release management processes

Employment History

2016–Present	Senior Technical Project Manager, American Institutes for Research, Washington, DC
2014–2016	Business Systems Analyst, SanDisk, Milpitas, CA
2007–2013	Senior Software Engineer, Accenture, Global, Los Angeles, CA

Technical Skills

Web Development Skills: HTML, CSS, JavaScript, jQuery, WordPress, PHP, MySQL, APIs. **Data Analysis Skills:** SQL, R programming, Tableau, Minitab; **Enterprise Applications Skills:** .NET Applications, SAP; **Project Management Skills:** Project Portfolio Management, Agile Methodology, Quality Centre, Change. Management, Object Oriented Analysis and Design, Service Now, SharePoint, Spreadsheet Modeling and Decision Analysis, Financial and Risk Analysis.

DEEPAK SHARMA, M.S.
SENIOR APPLICATION DEVELOPER, ONLINE REPORTING

Education

M.S.	2001, Computer Science University of Louisiana, Lafayette, LA
B.E.	1999, Computer Science & Engineering Marathwada Institute of Technology, Aurangabad, India

Present Position

Senior Software Engineer, American Institutes for Research (AIR) (2011–Present)
Responsible for design, development and maintenance of the Interactive Online Score Reporting modules. Leads the operational team that performs complex production deployments and provides production-level support. Also designs and develops software and processes that facilitate staging and production activities.

Project Experience at AIR

***Online Score Reports* -** Secure web-based application and associated services developed for AIR Assessment clients to provide online student score reports and data files for authorized state and local education personnel. Supervises design/development activities and leads the operational team for currently supported clients including - Arizona, California, Connecticut, Delaware, Hawaii, Idaho, Iowa, Montana, New Hampshire, North Dakota, Ohio, Oregon, South Dakota, Utah, Vermont, Virgin Islands, Washington, West Virginia, Wyoming

Additional Professional Experience

Senior Software Engineer, Datum Software, Atlanta, GA (2007–2011)
Designed and developed web-based applications, multi-threaded windows services and SOA based web services for clients including Network Communications Inc., State of Georgia, Macy's Systems and Technology and FiServ.

Employment History

2011–Present	Senior Software Engineer, American Institutes for Research, Washington, DC
2007–2011	Senior Software Engineer, Datum Software, Atlanta, GA
2005–2007	Systems Analyst, C & C Technologies, Lafayette, LA
2002–2005	Software Developer, Center for Innovative Learning and Assessment Technologies, University of Louisiana, Lafayette, LA

Technical Skills

Frameworks/Technologies: .NET, ASP.NET, AJAX, MVC, Entity Framework, Unity, WPF, WCF, Web Services, JQuery, Google Web Toolkit, XML, HTML, CSS. **Languages:** C#, Java, JavaScript, VB.NET, VB, C++, C, LISP, Scheme, Pascal, FORTRAN, MATLAB. **Development Tools:** Microsoft Visual Studio 200x, MS-SQL Server Management Studio, SQL Server BI Studio, MyEclipse; **Version Control:** Mercurial, CVS, Subversion, Team Foundation Server, Visual Source Safe. **Change/Issue Management:** JIRA, FogBugz, Rational ClearQuest, Bugzilla, TFS, TestTrack Pro, FootPrints; **Databases:** MS-SQL Server 200x, Oracle, PostGreSQL, MySQL, MS-Access. **Platforms:** Windows, UNIX (LINUX, Solaris), Mac OS X, MSDOS

DORIAN SOFIAJ, M.S.
SOFTWARE QUALITY ASSURANCE LEAD

Education

M.S.	2001, Computer Science American University, Washington, DC
M.S. Certificate	2001, Applied Physics American University, Washington, DC
B.S.	1994, Mechanical Engineering University of Tirana, Tirana, Albania

Present Position

Senior Quality Control Specialist, Manager of Software Testing, CSSC, American Institutes for Research (AIR) (2009–Present)
Leads and manages a 12-member quality assurance team through efficient and effective testing practices, which ensure that every system deployed meets specifications, functions, provides a satisfying user experience, and accurately captures or generates data. Also responsible for ensuring that assessment systems and statistical products delivered to clients are properly tested within the expected timeframes. Supports the efforts toward improving quality assurance and testing in all aspects of AIR's Computer and Statistical Sciences Center.

Employment History

2009–Present	Senior Quality Control Specialist, Manager of Software Testing, CSSC, American Institutes for Research, Washington, DC
2005–2009	Senior Technical Risk Specialist, Test Automation Team Lead, Fannie Mae, Washington, DC
2004–2005	Senior Technical Risk Specialist, Backend Test Automation Team Lead, Fannie Mae, Washington, DC
2002–2004	IT Analyst, Backend Lead Tester/Automation Engineer, Fannie Mae, Washington, DC
2001–2002	IT Analyst, Backend Tester, Fannie Mae, Washington, DC

Technical Skills

Programming Language: C#, Java. **Database Languages:** SQL, PL/SQL, TSQL. **Scripting:** Perl, UNIX scripting. Scripting Languages: TSL (WinRunner), VBScript (QTP)

FARA TAPSCOTT, B.S.

SENIOR NETWORK ENGINEER

Education

B.S.	2014, Computer and Information Science University of Maryland University College, Adelphi, MD

Present Position

Network Engineering Lead, American Institutes for Research (AIR) (2012–Present)
Responsible for the overall network and system engineering for the Assessment Division's web hosting environments. Additional responsibilities include system and network analysis, performance tuning, and application support for the Computing and Statistical Sciences Center group and its team of software developers, statisticians, and quality assurance specialists.

Project Experience at AIR

Network Engineer, American Institutes for Research (AIR) (2008–Present)
Serves as the primary system engineer for the initial and subsequent deployments of the deployment and system management of the web-based test delivery system (TDS) platform. Responsible for building hardware specifications, system configuration, and performance tuning to support the transactional and bandwidth requirements that the senior computer scientists provide.

Network Specialist, American Institutes for Research (AIR) (2004–Present)
As the network specialist, responsibilities include the overall network and system engineering for the Assessment Division's web hosting environments.

Employment History

2004–Present	Network Specialist, American Institutes for Research, Washington, DC
2002–2004	Network Analyst II, Infopro Inc., Reston, VA
1999–2002	System Engineer, Legg Mason, Wood Walker, Baltimore, MD
1998–1999	Field Technician, US Communications, Annapolis, MD

Professional Affiliations

Usenix—System Administrator Special Interest Group (SAGE)
The League of Professional System Administrators (LOPSA)

Technical Skills

Software: Windows Server and UNIX/Linux System administration; Microsoft SQL and MySQL database administration; VB, PHP, Perl, C, C++, and C# programming; ASP.NET web development environments; TCP/IP network administration and troubleshooting; and network/system design and documentation.

BANI DHEER, J.D.

VICE PRESIDENT, ASSESSMENT COMMUNICATIONS AND REPORTING

Education

J.D.	1995, Juris Doctorate University of British Columbia, Vancouver, BC
B.A.	1990, International Relations University of British Columbia, Vancouver, BC

Present Position

Vice President, Assessment Communications and Reporting, American Institutes for Research (AIR) (2004–Present)
Oversees all AIR Assessment activities related to the design, production, deployment, and delivery of paper and online score reports for the Delaware, Hawai‘i, Oregon, Ohio, New Mexico, and South Carolina departments of education. Builds state-of-the-art paper and online reporting capabilities for the AIR Assessment program. Leverages professional expertise in communications and market research to design, test, and refine data use and score reporting efforts with Department of Education officials to produce and distribute test score reports with evidence of impact on instruction. Scales AIR’s paper test score reporting capabilities from a single district pilot in 2004 to seven large-scale assessment programs in 2008. Integrated the paper reporting function with online reporting capabilities in one-and-a-half years. Manages operations involving designing and delivering more than 30 million pages of full-color score reports with variable text and graphics intended to drive changes in instruction for large-scale assessment programs.

Project Experience at AIR

Hawaii Department of Education Reports for Grades 3–8 and 10 (2006–Present)
Oversees the design, development, production, and delivery of family, class, school, and complex area reports for Hawaii’s public schools. Oversees deployments of the Online Reporting System to ensure close integration between paper and electronic reporting.

Minnesota Department of Education Reports for Grades 3–8 and High School and Monthly Reports for the GRAD Retests (2011–2014)
Oversaw the design and development of new paper student reports for four Minnesota assessment programs including the monthly GRAD retests. Oversaw all deployments of AIR’s Online Reporting System for Minnesota to provide real-time scores for the adaptive online tests and online scores for fixed form paper and online tests.

Delaware Department of Education Reports for Grades 2–10 (2009)
Oversaw all deployments of the online reporting system for the Department of Education as well as the delivery of all paper reports.

Ohio Graduation Test (2006–2009)
Oversaw production and delivery of reports for high school students taking the Ohio Graduation Test three times per year. Oversaw development of an electronic reporting system to ensure close integration between paper and electronic reporting.

Ohio Score Reports for Grade 3–8 Assessment System (2004–2008)
Oversaw the production and delivery of just under one million reports per year statewide (including a family report, a class math report, a class reading report, a school report, and a district report) for grades 3–8. Ensured timely delivery of all reports statewide within a six-week window from final data delivery to delivery of report shipments in the district.

Ohio Score Reports for Alternate Assessment for Students with Disabilities (2004–2008)
Oversaw production of 20,000 family-level and school-level reports annually for students with disabilities. These reports translate student performance on alternate assessments into meaningful recommendations for both families and educators.

South Carolina Reports for End-of-Course High School Exit Exam (2006–2007)
Oversaw the design, development, production, and delivery of family-level and class-level reports targeting students who do not pass the end-of-year high school Mathematics exit exam. These reports offer high school teachers and parents tailored recommendations to help remediate failing students by identifying individual student strengths and weaknesses.

Chicago Public School Reports on Learning Environment (2006–2007)
Oversaw the design, development, production, and delivery of school-level reports for all of Chicago's public schools reporting on the Conditions for Learning indiactors.

Director, International Benchmarking for State Assessments (2009–Present)
Establishes and manages AIR's relationship with the Programme for International Student Achievement (PISA) of the Organization for Economic Cooperation and Development (OECD). Negotiates license agreement with the OECD to embed PISA items in all AIR state assessment programs. Oversees statistical team to develop PISA-equivalent reporting scales for large-scale state assessments. Oversees efforts working across states to benchmark state performance standards to international standards.

Employment History

2004–Present	Vice President, Assessment Communications and Reporting, American Institutes for Research, Washington, DC
2002–2004	Senior Researcher, The Mellman Group, Washington, DC
2001–2004	Principal, Cook & Dheer, Vancouver, BC
2000–2001	Research Associate, The Strategic Counsel, Toronto, ON
1997–2000	Researcher/Consultant, d-Code, Inc., Toronto, ON
1995–2000	Managing Editor and Co-Owner, Sidestream Publications, BC

MARGARET WON, M.A.
LEAD, SCORE REPORTING

Education

M.A.	2012, International Economics and Finance Brandeis International Business School, Waltham, MA
B.A.	2011, Economics Wellesley College, Wellesley, MA

Honors and Awards

KOWIN DC Fellowship, Korean Women's International Network, Washington DC Chapter, 2012

Present Position

Project Manager, American Institutes for Research (AIR) (2015–Present)
Collaborates with the project team and developers to ensure on-time delivery of the online reporting system to multiple clients. Creates and maintains a schedule, a specifications document, issues logs, and UAT documents for each delivery. Ensures quality control and resolves issues while keeping to the scope of work. Serves as the key point of contact for all e-mail communication regarding score reporting for all clients. Manages a team of employees to ensure quality control.

Additional Professional Experience

Data Analyst, Department of Veterans Affairs, VBMS PMO, LongView International Technology Solutions (2013-2015)
Analyzed and interpreted real-time application performance data for the Veterans Benefits Management System tracking the number of concurrent users and transaction response times among other key performance indicators in weekly statistical and troubleshooting report deliverables using Microsoft Excel, Word, and PowerPoint and Tableau. Identified and recommended analytics opportunities and new modeling using Excel and Tableau to understand the performance of VBMS and the change in user behavior. Performed gap analysis according to the Project Management Accountability System requirements to determine which documents need to be updated and/or created for each milestone review. Created and distributed daily IT updates to all stakeholders in the project, including the VA Deputy Under Secretary for Benefits.

Employment History

2015–Present	Project Manager, American Institutes for Research, Washington, DC
2013–2015	Statistical/Communications Analyst, LongView International Technology Solutions, Reston, VA

Technical Skills

Data Analysis and Tools: R, Tableau, SQL.

DAMON HARTZLER, B.A.

LEAD, PERFORMANCE SCORING

Education

B.A.	1995, English and History The Ohio State University, Columbus, OH

Present Position

Performance Scoring Specialist, American Institutes for Research (AIR) (2013–Present)
Supervises workflow and temporary employees for the Smarter Balanced Assessment Consortium (SBAC) platform review. Supervises workflow and temporary employees for scoring statewide assessments. Works with state Departments of Education to ensure scoring align with state standards.

Additional Professional Experience

Scoring Specialist, Maryland Department of Education (2007–2009)
Oversaw scoring for Pearson Education projects, including grades 3–8 reading, grade 4 mathematics, and the alternate assessments. Oversaw Measurement Incorporated projects, including high school government and high school English. Developed projects for the Bridge Plan for academic validation.

Scoring Director, Measurement Incorporated (1999–2006; 2009–2013)
Worked with state Departments of Education to develop scoring materials that ensured scoring of a particular assessment adhered to state guidelines. Managed the day-to-day scoring of the assessments for grades 3–8 and high school English language arts (ELA) as well as social studies and government assessments for New Jersey and Maryland.

Employment History

2015–Present	Performance Scoring Specialist, American Institutes for Research, Columbus, OH
2009–2013	Scoring Director, Measurement Incorporated, Durham, NC
2007–2009	Scoring Specialist, Maryland Department of Education, Baltimore, MD
1999–2006	Scoring Director, Measurement Incorporated, Greensboro, NC

JUNE ZACK, M.A.

VICE PRESIDENT, ASSESSMENT CONTENT AND TEST DEVELOPMENT

Education

M.A. 1980, Educational Psychology
(coursework included psychometric theory, educational psychology, reading instruction, cognitive psychology, psycholinguistics, and neuropsychology)
New York University, New York, NY

B.A. 1968, Psychology
(coursework included psychology, speech pathology, audiology, and educational psychology)
Douglass College, Rutgers, the State University of New Jersey

Present Position

Vice President, Assessment Content and Test Development, American Institutes for Research (AIR) (2000–Present)
Responsible for overseeing the development of reading, writing, mathematics, science, social studies, and special education items for all AIR state and national assessment contracts including Ohio's K–2 Diagnostic Assessments and grades 3–8 Ohio Achievement Tests in reading, writing, and mathematics, science and social studies.

Project Experience at AIR

Principal Research Analyst, Manager of Test Development, Ohio K–8 Assessment, State Assessment Projects (2000–Present)
Responsible for directing the item writing and the assembly of diagnostic and achievement assessments for grades K–5. The components of this program include screening and diagnostic assessments in reading, writing, and mathematics components for grades K–2; screening and diagnostic assessments in writing, science, and social studies for grade 3; screening and diagnostic assessments for science and social studies for grade 4; screening and diagnostic assessments in writing for grade 5; achievement tests in reading and mathematics in grades 3–5; achievement tests in writing for grade 4; and achievement tests in science and social studies in grade 5.

Item Development Manager, South Carolina Alternate Assessment and New Mexico Alternate Assessment
Managed the development of the blueprints, item specifications, tasks and items, and test forms. Led standard setting for the South Carolina Alternate Assessment.

Principal Research Analyst, Manager of Test Development, Philadelphia K–4 Assessment Program
Responsible for directing the item writing and assembly of the diagnostic and achievement assessments for grades K–4. The components of this program included diagnostic assessments in literacy and mathematics for grades K–2 and achievement tests in reading and mathematics for grade 4.

Principal Research Analyst, Manager of Reading Development, NAEP Foreign Language Assessment
Responsible for directing the item writing and assembly of the NAEP assessment for high school students taking Spanish language courses. The components of this program included reading, listening, and conversational assessments.

Principal Research Analyst, Manager of Reading Test Development, Voluntary National Test Grade
Responsible for directing the passage and item development for the reading assessment.

Employment History

2000–Present	Vice President, Assessment Content and Test Development, American Institutes for Research, Washington, DC
1988	Evaluator of Reading Programs, New York City Board of Education, New York, NY
1987–2000	Director, ETS Atlanta Test Development Staff
1987–2000	Team Member, ETS Computer-Delivered Tests Training Modules
1985–1988	Adjunct Instructor, New York University, New York, NY
1985–1988	Adjunct Instructor, Lehman College CUNY, New York, NY
1969–1988	Reading and Language Arts Teacher in the Waldwick Public School System, Waldwick, NJ

JACOB WILKES, M.A.
ENGLISH LANGUAGE ARTS CONTENT LEAD

Education

M.A.	2013, Learning Sciences Northwestern University, Evanston, IL
M.A.	2009, English Brigham Young University, Provo, UT
B.A.	2006, English Brigham Young University, Provo, UT

Professional Credentials and Certifications

Teacher Certification, ELA 8–12, Texas State Board of Education, 2009
Teacher Certification, Reading 8–12, Texas State Board of Education, 2009
Teacher Certification, English as a Second Language, Texas State Board of Education, 2009

Present Position

Test Developer II, American Institutes for Research (AIR) (2013–Present)
Authors ELA assessment items across multiple grades; coordinates outside vendor development of items and reading passages; creates standards-based assessment items that incorporate technology and media; reviews and revises variety of item types aligned to the full range of state content standards; selects passage for test development appropriate to interest, grade, and standard alignment.

Project Experience at AIR

ELA Lead—Summative Assessment (2013–Present)
Directs authorship and review of secondary ELA items for state-wide assessments; coordinates item development with state office, including training on item development for outside item writers, item and passage review, and authoring of a range of single and multi-part assessment items.

ELA Developer, SBAC Performance Tasks (2013–2014)
Coordinated large-scale item development for multi-state assessment program; reviewed and authored Performance Tasks (multi-part activities that included authoring and excerpting reading passages, authoring short-response items and exemplars, and writing long-response essay questions); reviewed and developed teacher-led classroom activities used to introduce Performance Tasks.

ELA Lead—Formative Assessment (2013–2014)
Directed development of several hundred K–12 ELA formative assessment items; identified and developed standards-based reading passages; authored and reviewed classroom activities and lesson plans; designed system to track item and passage development from conception to client deployment.

Additional Professional Experience

National Online Community Leader—English as a Second Language, Teach for America (2011–2013)
Provided online instructional support to over 800 English beginning teachers across the United States; led team of mentor teachers that developed lessons, curriculum, and teaching strategies for English and English as a Second Language teachers; drafted best-practice standards for working with English Language Learners. District Curriculum Committee, English Language Arts, Houston ISD (2012–2013) Nominated to district-wide curriculum steering committee based on teaching excellence and outstanding student performance; coordinated development of curriculum materials; authored lessons and activities for high school English teachers.

English Teacher, Westbury High School (2009–2013)
Consistently led multiple grades of students to significant gains in English proficiency; authored activities and assessment items later adopted department-wide; lead school-wide training on best practices in teaching English Language Learners; developed program for implementing multimedia into classroom; jointly developed training video (with central district office) used for professional development of English as a Second Language teachers; taught high school English, reading, and English as Second Language classes.

Content Writer, Performance Management Training Program, Central HR—Church of Jesus Christ of Latter-day Saints (2008–2009)
Co-authored e-course on performance management (best practices leadership and management training program for mid-level managers) distributed to 3,000 managers; coordinated course multimedia development including photo shoots, sound recording, and online multimedia; met or exceeded all deadlines; authored each course 50% faster than similar courses in training series.

Writing Instructor, Freshman Composition, Brigham Young University (2006–2008)
Taught multiple sections of first-year university writing course; developed syllabus, course materials (including writing assignments, rubrics, tests, quizzes, and multimedia-based classroom activities); selected as one of six instructors to teach special computer-intensive writing and rhetoric course.

English as a Foreign Language Teacher, Korea Pusan Mission (2001–2003)
Taught and supervised multiple community-based English classes; developed English curriculum for beginning to advanced-level students; served as city-wide supervisor, overseeing multiple teachers and coordinating regular training of volunteers.

Employment History

2013–Present	Senior Software Engineer, American Institutes for Research, Washington, DC
2011–2013	National Online Community Leader, Teach for America, Home Office
2009–2013	English Teacher, Westbury High School, Houston, TX
2008–2009	Content Writer, Central HR—Church of Jesus Christ of Latter-day Saints, Salt Lake City, UT
2006–2008	Writing Instructor, Brigham Young University, Provo, UT
2001–2003	EFL Teacher, Korea Pusan Mission, Pusan, South Korea

ALYSA KARTEE, M.A.
MATHEMATICS CONTENT LEAD

Education

M.A. 2012, Mathematics
Stony Brook University, Stony Brook, NY

B.A. 2010, Mathematics (cum laude)
St. Joseph's College, Patchogue, NY

Professional Credentials and Certifications

Initial Certification of Mathematics Secondary Education, New York (2010)

Present Position

Mathematics Test Developer I, American Institutes for Research (AIR) (2014–Present)
Writes and reviews math test items aligned to Common Core and state-specific content standards for a variety of grade levels and projects. Builds and reviews online and paper fixed forms. Leads various client and educator committee meetings.

Project Experience at AIR

Florida Grade Band Lead, Florida Department of Education, Florida Standards Assessments (2014–Present)
Translates the Florida Standards Assessments blueprint into item development plans for grades 3 through 5. Manages item development for 200 elementary items and monitors item pool of 450 items. Performs content reviews and prepares batches of items prior to delivering to client. Successfully meets client deadlines with a 98% item acceptance rate. Leads client and educator committee meetings in item review, rubric validation, test construction, and standard setting. Works with client to create and revise elementary test item specifications and practice tests. Works closely with other Florida grade band leads and content leads to maintain consistency across grades.

Grade Level Lead, Internal AIR Core Project (2014–2015)
Created an item development plan of 425 items aligned to the blueprint and Common Core State Standards for grade 7. Managed item writing and review assignments for an internal team of 20 content specialists and three outside vendor companies. Monitored item pool to ensure consistency with item development plan and content standards.

Test Developer, Smarter Balanced Assessment Consortium (2014)
Assisted with the configurations of the field test and practice test for mathematics and ELA.

Additional Professional Experience

New York State Assessment Scorer, Optimum Solutions Corporation (2013)
Followed strict state guidelines to score New York State mathematics exams for grades 3 through 8.

Summer School Math Teacher, Eastern Suffolk BOCE S (2011–2012)
Prepared lesson plans and taught students in grades 7 through 12. Administered, scored, and prepared students for the New York State Regents Exams.

Substitute Math Teacher, Sachem School District (2010–2012)
Followed lesson plans of absent teachers for grades 6 through 12 in various subjects and provided additional assistance to students as necessary.

Employment History

2014–Present	Mathematics Test Developer I, American Institutes for Research, Washington, DC
2013	New York State Assessment Scorer, Optimum Solutions Corporation, Central Islip, NY
2011–2012	Summer School Math Teacher, Eastern Suffolk BOCES, Patchogue, NY
2010–2012	Substitute Math Teacher, Sachem School District, Holtsville, NY

RACHEL A. AAZZERAH, M.S.
SCIENCE CONTENT LEAD

Education

M.S. 2003, Science and Mathematics Education/Biochemistry
Oregon State University, Corvallis, OR

B.S. 1999, Chemistry
Oregon State University, Corvallis, OR

Professional Credentials and Certifications

Oregon Professional Teaching License (Biology, Chemistry, Integrated Science, Mathematics, Middle School Science and Physics), Oregon Teacher Standards and Practices Commission, 2000–2021

Present Position

Senior Manager, Test Development, American Institutes for Research (AIR) (2016–Present)
Oversees a team of Next Generation Science Standards item writers and test developers in developing science items aligned to the K–12 Framework for Science Education (NGSS adapted states). Conducts senior content, style, and appropriateness reviews of item clusters and item batches. Leads content staff meetings, client calls, and face-to-face meetings. Interacts with project management staff and the client to address and resolve all test development issues. Trains the item and test developers on project specific processes and assists/supervises the content leads with training clients on AIR-specific processes. Assists the content leads in resolving disputes. Oversees stakeholder content review meetings and data review meetings for Science. Leads and/or oversees rubric validation meetings for constructed-response items and machine-scored graphic response (grid) items. Builds, revises, and reviews Science test forms and blueprints. Responds to the needs of project management staff, other AIR teams, and the client, as necessary.

Additional Professional Experience

Science and Social Sciences Assessment Specialist, Oregon Assessment of Knowledge and Skills (OAKS) and Smarter Balanced, Oregon Department of Education (2011–2016)
Managed the item development for the Oregon Assessment of Knowledge and Skills (OAKS) in Science and Social Sciences, and managed the item development for Smarter Balanced Assessments in Mathematics (subcontract). Led meetings and training including: item writer training for teachers; content review committee meetings; range-finding and standard setting meetings; bias and sensitivity review meetings; accessibility review meetings; and rubric validation meetings. Conducted User Acceptance Testing on all testing platforms and operating systems, and provided professional development for Oregon Educators on both the Next Generation Science Standards and Smarter Balanced Assessments.

Science and Math Item Content Reviewer, Oregon Assessment of Knowledge and Skills (OAKS), Oregon Department of Education (2006–2011)
Reviewed math and science items that were aligned with the Oregon Content Standards for grades 3–11 for content, bias and sensitivity, and accessibility. Reviewed and scored Scientific Inquiry and Math Problem Solving tasks for grades 5, 8, and 11.

Science and Math Item Developer, Oregon Assessment of Knowledge and Skills (OAKS), Oregon Department of Education (1999–2006)
Developed math and science items aligned with Oregon Content Standards for grades 3–10. Developed Scientific Inquiry classroom based performance tasks and scoring rubrics. Scored and gave score justifications on student Scientific Inquiry and Math Problem Solving anchor papers.

Employment History

2016–Present Senior Manager, Test Development, American Institutes for Research, Washington, DC

2011–2016 Science and Social Sciences Assessment Specialist, Oregon Department of Education, Salem, OR

2006–2011 Science and Math Item Content Reviewer, Oregon Department of Education, Salem, OR

1999–2006 Science and Math Item Developer, Oregon Department of Education, Salem, OR

Professional Affiliations

American Chemical Society (ACS)
Council of Chief School Officers, Science (SCASS)
Council of State Science Supervisors (CSSS)
National Association of Biology Teachers (NABT)
National Association of Science Teachers (NSTA)
Oregon Council of Teachers of Mathematics (OCTM)
Oregon Science Teachers Association (OSTA)

Publications

Bishop, C.V., Aazzerah, R.A., Quennoz, L.M., Hennebold, J.D. and Stouffer, R.L. (2014*) Effects of steroid ablation and progestin replacement on the transcriptome of the primate corpus luteum during simulated early pregnancy. Mol. Hum. Reprod, 20*(3), 222–34.

Professional Presentations

King, K., Romero, N., Parton, C. and Aazzerah, R. (2016, June). *Science Assessment Item Collaborative (SAIC), Next Generation Science Standards (NGSS) Assessment Framework, and Item Cluster Prototypes: New Tools to Support NGSS Large-Scale Assessment Development*. Presented at meeting of National Conference on Student Assessment, Philadelphia, PA.

KARA PREZOCKI, B.A.
SENIOR PRODUCTION MANAGER

Education

B.A.	1996, Psychology Wittenberg University, Springfield, OH

Professional Credentials and Certifications

Studio Art Courses, Cincinnati Academy of Design, 2000
Continuing Education Course in Editing/Proofreading, University of Virginia, 2002

Present Position

Senior Production Manager, American Institutes for Research (AIR) (2008–Present)
Oversees the production of graphics and test materials for all AIR Assessment contracts, including Florida Standards Assessments, MAAC (Smarter Balanced: Washington, Idaho, West Virginia), Arizona (AzMERIT), Ohio (OGT, OCBA, OTELA), Hawaii (HSA, EOC, HSA-Alt), Delaware (DCAS, DCAS – Alt1), and Utah computer-based assessments. Manages the process and quality standards for test item artwork, test books, and ancillaries for online and print test formats. Oversees proposal production and editing efforts. Hires and supervises graphic designers, desktop publishers, and production editors, ensuring that staff meet deadlines, adhere to state style guides, and follow quality control procedures as related to production.

Project Experience at AIR

Test Production Specialist (2005–2008)
Managed and trained production and graphic artists on styles, processes, procedures, and quality control measures. Coordinated with editorial, item development, and psychometric teams to meet schedules. Worked closely with clients and print subcontractors for South Carolina EOCEP and HSAP programs. Responsibilities included the production of print test products, including test books, manuals, and brochures for South Carolina HSAP, South Carolina EOCEP, Ohio Achievement Assessments, Ohio Graduation Tests, Hawaii HSA, and South Carolina Alternate Assessments.

Additional Professional Experience

Production Team Lead, Production Coordinator, Production Assistant, Bernan Associates (2001–2005)
Designed government information books and covers and coordinated with print vendors, authors, and warehouse distribution staff. Approved bluelines for print and proofread and edited text. Hired and managed production support staff. Works included joint ventures with the Library of Congress and the World Trade Organization.

Lead Pre-School Teacher, Discovery World Child Development Center (1997–2001)
Responsibilities included the creation and implementation of weekly lesson plans and art activities suitable for the early education of children 3 to 4 years of age.

Employment History

2008–Present Senior Production Manager, American Institutes for Research, Washington, DC

2005–2008 Test Production Specialist, American Institutes for Research, Washington, DC

2001–2005 Production Team Lead, Production Coordinator, Production Assistant, Bernan Associates, Lanham, MD

1997–2001 Lead Pre-School Teacher, Discovery World Child Development Center, Cincinnati, OH

Technical Skills

Software: Adobe Professional Suite (Illustrator, InDesign, Photoshop, Distiller), QuarkXPress, Microsoft Office Suite, Microsoft Project, SmartCVS

Appendix C: Sample Risk Register

Sample Risk Register Format

Risk #	Date Identified	Paper or Online	Status	Description	Rating	Impact	Trigger Event	Owner	Agreed Response	Response Plan
1	10/15/1c	Online	Open	Example: Request change of test windows – open date move back 1 week.	H	H	Request from field	AIR/FDOE	Yes	Schedule will be updated – notifications to the field will be sent.
2	5/1/1c	Online	Open	Example: ITS – new version of web preview delay	H	H	UAT	AIR	Yes	Contingency is to use previous version until new version is ready.

Appendix D:
Project Schedule

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
1			New Hampshire Statewide Assessments Administration Schedule - Year 1 (2017 - 2018)		Mon 7/3/17		
2			**Test Windows**	**235 days**	**Mon 8/7/17**	**Fri 6/29/18**	
3			**Interim Test Windows**	**144 days**	**Tue 11/21/17**	**Fri 6/8/18**	
4			Interim Test Window (ELA and Math)	144 days	Tue 11/21/17	Fri 6/8/18	
5			Interim Test Window (Science)	144 days	Tue 11/21/17	Fri 6/8/18	4FF
6			**Summative Test Windows**	**63 days**	**Tue 3/13/18**	**Thu 6/7/18**	
7			Summative Test Window (ELA and Math)	63 days	Tue 3/13/18	Thu 6/7/18	
8			Summative Test Window (Science)	63 days	Tue 3/13/18	Thu 6/7/18	7FF
9			**Training and Practice Tests**	**235 days**	**Mon 8/7/17**	**Fri 6/29/18**	
10			Training Tests	235 days	Mon 8/7/17	Fri 6/29/18	
11			Practice Tests	235 days	Mon 8/7/17	Fri 6/29/18	10FF
12			**In-Person Meetings**	**261 days**	**Mon 7/10/17**	**Mon 7/9/18**	
13			Project Kickoff Meeting	2 days	Mon 7/10/17	Tue 7/11/17	
14			**TAC Meeting #1**	**2 days**	**Tue 2/20/18**	**Wed 2/21/18**	
15			Create Draft Agenda	1 day	Wed 1/10/18	Wed 1/10/18	26SS-35 days
16			Post Draft Agenda for Review	0 days	Thu 1/11/18	Thu 1/11/18	15
17			Review Draft Agenda	5 days	Thu 1/11/18	Wed 1/17/18	16
18			Apply Changes/Edits	2 days	Thu 1/18/18	Fri 1/19/18	17
19			Approve Draft Agenda	2 days	Mon 1/22/18	Tue 1/23/18	18
20			Create Meeting Materials	10 days	Wed 1/24/18	Tue 2/6/18	19
21			Post Meeting Materials for Review	0 days	Wed 2/7/18	Wed 2/7/18	20
22			Review Meeting Materials	5 days	Wed 2/7/18	Tue 2/13/18	21
23			Apply Changes/Edits	2 days	Wed 2/14/18	Thu 2/15/18	22
24			Approve Meeting Materials	5 days	Fri 2/16/18	Thu 2/22/18	23
25			Send TAC Meeting Materials	0 days	Fri 2/23/18	Fri 2/23/18	24

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
26			Conduct TAC Meeting	2 days	Wed 2/28/18	Thu 3/1/18	
27			Conduct TAC Meeting Debrief	1 day	Fri 3/2/18	Fri 3/2/18	26
28			Post TAC Meeting Action Items	0 days	Mon 3/5/18	Mon 3/5/18	27
29			Post TAC Meeting Notes	0 days	Mon 3/5/18	Mon 3/5/18	28
30			**TAC Meeting #2**	**2 days**	**Tue 6/26/18**	**Wed 6/27/18**	
31			Create Draft Agenda	1 day	Tue 5/8/18	Tue 5/8/18	42SS-35 days
32			Post Draft Agenda for Review	0 days	Wed 5/9/18	Wed 5/9/18	31
33			Review Draft Agenda	5 days	Wed 5/9/18	Tue 5/15/18	32
34			Apply Changes/Edits	2 days	Wed 5/16/18	Thu 5/17/18	33
35			Approve Draft Agenda	2 days	Fri 5/18/18	Mon 5/21/18	34
36			Create Meeting Materials	10 days	Tue 5/22/18	Mon 6/4/18	35
37			Post Meeting Materials for Review	0 days	Tue 6/5/18	Tue 6/5/18	36
38			Review Meeting Materials	5 days	Tue 6/5/18	Mon 6/11/18	37
39			Apply Changes/Edits	2 days	Tue 6/12/18	Wed 6/13/18	38
40			Approve Meeting Materials	5 days	Thu 6/14/18	Wed 6/20/18	39
41			Send TAC Meeting Materials	0 days	Thu 6/21/18	Thu 6/21/18	40
42			Conduct TAC Meeting	2 days	Tue 6/26/18	Wed 6/27/18	
43			Conduct TAC Meeting Debrief	1 day	Thu 6/28/18	Thu 6/28/18	42
44			Post TAC Meeting Action Items	0 days	Fri 6/29/18	Fri 6/29/18	43
45			Post TAC Meeting Notes	0 days	Fri 6/29/18	Fri 6/29/18	44
46			Exit Meeting (if applicable)	1 day	Mon 7/9/18	Mon 7/9/18	
47			**Weekly Status Meeting**	**246 days**	**Wed 7/19/17**	**Wed 6/27/18**	
98			**Program Management Documents/Plans**	**260 days**	**Mon 7/3/17**	**Sun 7/1/18**	
99			**Monthly Project Status Report**	**239 days**	**Tue 8/1/17**	**Sun 7/1/18**	
112			Work Plan	10 days	Mon 7/3/17	Fri 7/14/17	
113			Infrastructure Plan	10 days	Mon 7/3/17	Fri 7/14/17	

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
114			Security Plan	10 days	Mon 7/3/17	Fri 7/14/17	
115			Communications/Change Management Plan	10 days	Mon 7/3/17	Fri 7/14/17	
116			Requirements Traceability Matrix	10 days	Mon 7/3/17	Fri 7/14/17	
117			Software Configuration Plan	10 days	Mon 7/3/17	Fri 7/14/17	
118			Systems Interface Plan	10 days	Mon 7/3/17	Fri 7/14/17	
119			Testing Plan	10 days	Mon 7/3/17	Fri 7/14/17	
120			Data Conversion Plan/Design	10 days	Mon 7/3/17	Fri 7/14/17	
121			Deployment/Roll-Out Plan	10 days	Mon 7/3/17	Fri 7/14/17	
122			Training Plan	10 days	Mon 7/3/17	Fri 7/14/17	
123			End-User Support Plan	10 days	Mon 7/3/17	Fri 7/14/17	
124			Business Continuinty Plan	10 days	Mon 7/3/17	Fri 7/14/17	
125			**Help Desk Set-Up**	**194 days**	**Tue 10/3/17**	**Fri 6/29/18**	
126			Creation of Help Desk Management Plan	1 day	Tue 10/3/17	Tue 10/3/17	
127			Help Desk Training	2 days	Thu 10/5/17	Fri 10/6/17	
128			Help Desk Open	189 days	Tue 10/10/17	Fri 6/29/18	4SS-5 days
129			**Science (NGSS) Item Development**	**144 days**	**Mon 7/17/17**	**Thu 2/1/18**	
130			**Item Specifications**	**10 days**	**Mon 7/17/17**	**Fri 7/28/17**	
131			Review AIR drafted NGSS Item Specifications	10 days	Mon 7/17/17	Fri 7/28/17	
132			State provides feedback on NGSS Item specifications	10 days	Mon 7/31/17	Fri 8/11/17	131
133			**Review of AIR Core NGSS Developed Clusters and Items**	**35 days**	**Mon 7/17/17**	**Fri 9/1/17**	
134			AIR selects developed AIR Core clusters and send to states for review	10 days	Mon 7/17/17	Fri 7/28/17	
135			State selects clusters to be brought to Item Review Meeting	10 days	Mon 7/31/17	Fri 8/11/17	134
136			AIR prepares materials for Item Review Meeting and Bias/Sensitivity Meeting	5 days	Mon 8/14/17	Fri 8/18/17	135
137			Item Review Meeting	3 days	Mon 8/28/17	Wed 8/30/17	

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
138			Bias and Sensitivity Meeting	2 days	Thu 8/31/17	Fri 9/1/17	137
139			**Development and Review of NH NGSS Clusters and Items**	**108 days**	**Tue 9/5/17**	**Thu 2/1/18**	
140			**Item Specifications Lockdown**	**10 days**	**Tue 9/5/17**	**Mon 9/18/17**	
141			Review feedback and update specifications from August Item Review Meeting	9 days	Tue 9/5/17	Fri 9/15/17	138
142			Finalize Item Specifications	1 day	Mon 9/18/17	Mon 9/18/17	141
143			**Test Blueprint**	**24 days**	**Tue 9/5/17**	**Fri 10/6/17**	
144			Review Draft NGSS Test Blueprint	9 days	Tue 9/5/17	Fri 9/15/17	
145			AIR makes edits based on client feedback	10 days	Mon 9/18/17	Fri 9/29/17	144
146			Prepare blueprints for TAC presentation	5 days	Mon 10/2/17	Fri 10/6/17	145
147			**Item Development**	**104 days**	**Mon 9/11/17**	**Thu 2/1/18**	
148			**Batch 1**	**25 days**	**Mon 9/11/17**	**Fri 10/13/17**	
149			AIR drafts NGSS clusters and standalone items	10 days	Mon 9/11/17	Fri 9/22/17	
150			AIR sends Batch 1 items to state for review and feedback	1 day	Fri 9/22/17	Fri 9/22/17	149FF
151			State reviews and provides feedback	10 days	Mon 9/25/17	Fri 10/6/17	150
152			AIR makes edits based on client feedback	5 days	Mon 10/9/17	Fri 10/13/17	151
153			**Batch 2**	**25 days**	**Mon 9/25/17**	**Fri 10/27/17**	
154			AIR drafts NGSS clusters and standalone items	10 days	Mon 9/25/17	Fri 10/6/17	150
155			AIR sends Batch 1 items to state for review and feedback	1 day	Fri 10/6/17	Fri 10/6/17	154FF
156			State reviews and provides feedback	10 days	Mon 10/9/17	Fri 10/20/17	155
157			AIR makes edits based on client feedback	5 days	Mon 10/23/17	Fri 10/27/17	156
158			**Batch 3**	**25 days**	**Mon 10/9/17**	**Fri 11/10/17**	
159			AIR drafts NGSS clusters and standalone items	10 days	Mon 10/9/17	Fri 10/20/17	155

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
160			AIR sends Batch 1 items to state for review and feedback	1 day	Fri 10/20/17	Fri 10/20/17	159FF
161			State reviews and provides feedback	10 days	Mon 10/23/17	Fri 11/3/17	160
162			AIR makes edits based on client feedback	5 days	Mon 11/6/17	Fri 11/10/17	161
163			**Item Review/Bias and Sensitivity Meeting**	**23 days**	**Mon 11/13/17**	**Wed 12/13/17**	
164			Prepare materials for meeting	15 days	Mon 11/13/17	Fri 12/1/17	162
165			Item Review Meeting	3 days	Mon 12/4/17	Wed 12/6/17	
166			Bias and Sensitivity Meeting	2 days	Thu 12/7/17	Fri 12/8/17	165
167			Resolve Committee feedback from meetings	5 days	Mon 12/11/17	Fri 12/15/17	166
168			**Content Lockdown**	**34 days**	**Mon 12/18/17**	**Thu 2/1/18**	
169			AIR makes edits to items based on feedback	10 days	Mon 12/18/17	Fri 12/29/17	167
170			State reviews and approves the updated item	15 days	Wed 12/20/17	Tue 1/9/18	169SS+2 days
171			All NGSS clusters and items are reviewed and finalized	2 days	Wed 1/10/18	Thu 1/11/18	170
172			NGSS clusters and items are tagged for TTS	10 days	Fri 1/12/18	Thu 1/25/18	171
173			Pilot items are web approved and locked down	5 days	Fri 1/26/18	Thu 2/1/18	172
174			**Test Form Development (ELA and Math)**	**105 days**	**Mon 8/7/17**	**Fri 12/29/17**	
175			**Online Adaptive Test**	**26 days**	**Mon 8/7/17**	**Mon 9/11/17**	
176			AIR drafts adaptive algorithm design	10 days	Tue 8/8/17	Mon 8/21/17	
177			NH DOE reviews algorithm design	5 days	Tue 8/22/17	Mon 8/28/17	176
178			AIR updates algorithm design based on NH DOE feedback	5 days	Tue 8/29/17	Mon 9/4/17	177
179			NH DOE reviews/approves algorithm design	5 days	Tue 9/5/17	Mon 9/11/17	178
180			**Fixed form (Paper)**	**80 days**	**Mon 9/11/17**	**Fri 12/29/17**	
181			AIR selects items for paper versions	20 days	Mon 9/11/17	Fri 10/6/17	
182			NH DOE reviews item selection and sends feedback to AIR	10 days	Mon 10/9/17	Fri 10/20/17	181
183			AIR makes updates to item selection and sends to NH DOE for review/approval	5 days	Mon 10/23/17	Fri 10/27/17	182

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
184			NH DOE approves items for fixed form	5 days	Mon 10/30/17	Fri 11/3/17	183
185			AIR creates paper booklets and scannable answer documents	20 days	Mon 11/6/17	Fri 12/1/17	184
186			NH DOE reviews paper test booklets and answer documents	20 days	Mon 12/4/17	Fri 12/29/17	185
187			**Test Form Development (Science)**	**287 days**	**Wed 4/19/17**	**Thu 5/24/18**	
188			**Online Adaptive Test**	**26 days**	**Wed 4/19/17**	**Wed 5/24/17**	
189			AIR drafts adaptive algorithm design	10 days	Tue 8/8/17	Mon 8/21/17	176FF
190			NH DOE reviews algorithm design	5 days	Tue 8/22/17	Mon 8/28/17	189
191			AIR updates algorithm design based on NH DOE feedback	5 days	Tue 8/29/17	Mon 9/4/17	190
192			NH DOE reviews/approves algorithm design	5 days	Tue 9/5/17	Mon 9/11/17	191
193			**Fixed form (Paper)**	**80 days**	**Fri 2/2/18**	**Thu 5/24/18**	
194			AIR selects items for paper versions	20 days	Fri 2/2/18	Thu 3/1/18	173
195			NH DOE reviews item selection and sends feedback to AIR	10 days	Fri 3/2/18	Thu 3/15/18	194
196			AIR makes updates to item selection and sends to NH DOE for review/approval	5 days	Fri 3/16/18	Thu 3/22/18	195
197			NH DOE approves items for fixed form	5 days	Fri 3/23/18	Thu 3/29/18	196
198			AIR creates paper booklets and scannable answer documents	20 days	Fri 3/30/18	Thu 4/26/18	197
199			NH DOE reviews paper test booklets and answer documents	20 days	Fri 4/27/18	Thu 5/24/18	198
200			**Online Systems**	**192 days**	**Mon 7/3/17**	**Tue 3/27/18**	
201			**Portal**	**10 days**	**Tue 8/8/17**	**Mon 8/21/17**	
202			AIR and NH DOE review and finalize portal specifications	1 day	Tue 8/8/17	Tue 8/8/17	
203			Client UAT	4 days	Wed 8/9/17	Mon 8/14/17	202
204			Edits to portal based on UAT feedback	4 days	Tue 8/15/17	Fri 8/18/17	203
205			Portal GO LIVE	1 day	Mon 8/21/17	Mon 8/21/17	204
206			**TIDE**	**92 days**	**Mon 7/3/17**	**Tue 11/7/17**	

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
207			Develop Specs	20 days	Mon 7/3/17	Fri 7/28/17	
208			Specs Finalized	1 day	Mon 7/31/17	Mon 7/31/17	207
209			Develop/Configure TIDE	30 days	Tue 8/1/17	Mon 9/11/17	208
210			Test	10 days	Tue 9/12/17	Mon 9/25/17	209
211			Prepare Client UAT Document and set up client UAT environment	3 days	Tue 9/26/17	Thu 9/28/17	210
212			Internal UAT	5 days	Fri 9/29/17	Thu 10/5/17	211
213			Client UAT #1	5 days	Fri 10/6/17	Thu 10/12/17	212
214			Updates to TIDE based on UAT feedback	5 days	Fri 10/13/17	Thu 10/19/17	213
215			Client UAT #2	2 days	Fri 10/20/17	Mon 10/23/17	214
216			Client provides updated institution file	1 day	Tue 10/24/17	Tue 10/24/17	215
217			Client provides student file	1 day	Wed 10/25/17	Wed 10/25/17	216
218			TIDE downtime	9 days	Thu 10/26/17	Tue 11/7/17	217
219			Deploy TIDE to production/TIDE GO LIVE	1 day	Tue 11/7/17	Tue 11/7/17	4SS-10 days,218SS
220			**Secure Browser**	**51 days**	**Mon 7/3/17**	**Mon 9/11/17**	
221			Update the Secure Browsers	40 days	Mon 7/3/17	Fri 8/25/17	
222			Internal UAT	5 days	Mon 8/28/17	Fri 9/1/17	221
223			Client UAT	5 days	Mon 9/4/17	Fri 9/8/17	222
224			Secure Browser GO LIVE	1 day	Mon 9/11/17	Mon 9/11/17	223
225			**TDS**	**150 days**	**Wed 8/16/17**	**Tue 3/13/18**	
226			**Interim Assessments (ELA, Math, and Science)**	**70 days**	**Wed 8/16/17**	**Tue 11/21/17**	
227			Develop Specs	19 days	Wed 8/16/17	Mon 9/11/17	
228			Specs Finalized	1 day	Mon 9/11/17	Mon 9/11/17	227FF
229			Load Configs	20 days	Tue 9/12/17	Mon 10/9/17	228
230			Internal UAT	9 days	Tue 10/10/17	Fri 10/20/17	229
231			Client UAT #1	5 days	Mon 10/23/17	Fri 10/27/17	230

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task		External Milestone		Manual Summary Rollup	
Split		Inactive Task		Manual Summary	
Milestone		Inactive Milestone		Start-only	
Summary		Inactive Summary		Finish-only	
Project Summary		Manual Task		Deadline	
External Tasks		Duration-only		Progress	

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
232			Updates to TDS based on UAT Feedback	5 days	Mon 10/30/17	Fri 11/3/17	231
233			Client UAT #2	5 days	Mon 11/6/17	Fri 11/10/17	232
234			TDS Downtime - preparation for deployment	6 days	Mon 11/13/17	Mon 11/20/17	233
235			Interim Assessments GO LIVE	1 day	Tue 11/21/17	Tue 11/21/17	4SS,234
236			**Summative Assessments (ELA, Math, and Science)**	**92 days**	**Mon 11/6/17**	**Tue 3/13/18**	
237			Develop Specs	18 days	Mon 11/6/17	Wed 11/29/17	
238			Specs Finalized	1 day	Thu 11/30/17	Thu 11/30/17	237
239			Load Configs	12 days	Mon 1/15/18	Tue 1/30/18	
240			Internal UAT	10 days	Wed 1/31/18	Tue 2/13/18	239
241			Client UAT #1	5 days	Wed 2/14/18	Tue 2/20/18	240
242			Updates to TDS based on UAT Feedback	5 days	Wed 2/21/18	Tue 2/27/18	241
243			Client UAT #2	5 days	Wed 2/28/18	Tue 3/6/18	242
244			TDS Downtime - preparation for deployment	4 days	Wed 3/7/18	Mon 3/12/18	243
245			Summative Assessments GO LIVE	1 day	Tue 3/13/18	Tue 3/13/18	107SS
246			**THSS**	**47 days**	**Mon 9/18/17**	**Tue 11/21/17**	
247			Develop THSS Specs	5 days	Mon 9/18/17	Fri 9/22/17	
248			Specs Finalized	1 day	Mon 9/25/17	Mon 9/25/17	247
249			Develop/Configure THSS	10 days	Tue 9/26/17	Mon 10/9/17	248
250			Internal UAT	9 days	Tue 10/10/17	Fri 10/20/17	249
251			Client UAT #1	5 days	Mon 10/23/17	Fri 10/27/17	250
252			Downtime - fix any UAT issues	5 days	Mon 10/30/17	Fri 11/3/17	251
253			Client UAT #2	5 days	Mon 11/6/17	Fri 11/10/17	252
254			Downtime - preparation for deployment	6 days	Mon 11/13/17	Mon 11/20/17	253
255			Interim Assessments GO LIVE	1 day	Tue 11/21/17	Tue 11/21/17	4SS
256			**ORS**	**172 days**	**Mon 7/31/17**	**Tue 3/27/18**	
257			**Interim Assessments (Participation and Score Reports**	**82 days**	**Mon 7/31/17**	**Tue 11/21/17**	

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
258			Develop Specs	15 days	Mon 7/31/17	Fri 8/18/17	
259			Finalize Specs	1 day	Mon 8/21/17	Mon 8/21/17	258
260			Develop/Configure ORS	39 days	Tue 8/22/17	Fri 10/13/17	259
261			Internal UAT	9 days	Mon 10/16/17	Thu 10/26/17	260
262			Client UAT #1	5 days	Fri 10/27/17	Thu 11/2/17	261
263			Make updates to ORS based on client feedback	5 days	Fri 11/3/17	Thu 11/9/17	262
264			Client UAT #2	5 days	Fri 11/10/17	Thu 11/16/17	263
265			ORS Downtime	2 days	Fri 11/17/17	Mon 11/20/17	264
266			Deploy Participaton Reports and Score Reports for Interim	1 day	Tue 11/21/17	Tue 11/21/17	4SS
267			**Summative Assessments**	**92 days**	**Mon 11/20/17**	**Tue 3/27/18**	
268			Develop Specs	15 days	Mon 11/20/17	Fri 12/8/17	
269			Specs Finalized	1 day	Mon 12/11/17	Mon 12/11/17	268
270			Develop/Configure ORS	39 days	Tue 12/12/17	Fri 2/2/18	269
271			Internal UAT	10 days	Mon 2/5/18	Fri 2/16/18	270
272			Client UAT #1	5 days	Mon 2/19/18	Fri 2/23/18	271
273			Updates to ORS based on UAT Feedback	5 days	Mon 2/26/18	Fri 3/2/18	272
274			Client UAT #2	5 days	Mon 3/5/18	Fri 3/9/18	273
275			ORS Summative GO LIVE (Participation Reports)	1 day	Tue 3/13/18	Tue 3/13/18	7SS
276			Score Reports GO LIVE	1 day	Tue 3/27/18	Tue 3/27/18	275SS+10 days
277			**Trainings**	**172 days**	**Mon 8/7/17**	**Tue 4/3/18**	
278			**Regional Training (Assessment Administration Training)**	**55 days**	**Mon 8/7/17**	**Fri 10/20/17**	
279			Reserve locations	19 days	Mon 8/7/17	Thu 8/31/17	
280			Planning meeting with NH DOE	1 day	Mon 8/28/17	Mon 8/28/17	
281			Build registration site	4 days	Tue 8/29/17	Fri 9/1/17	280
282			Registration	30 days	Mon 9/4/17	Fri 10/13/17	281

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task		External Milestone		Manual Summary Rollup	
Split		Inactive Task		Manual Summary	
Milestone		Inactive Milestone		Start-only	
Summary		Inactive Summary		Finish-only	
Project Summary		Manual Task		Deadline	
External Tasks		Duration-only		Progress	

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
283			Create outline	1 day	Mon 9/4/17	Mon 9/4/17	
284			Outline approved by NH DOE	1 day	Tue 9/5/17	Tue 9/5/17	283
285			Create PPT	4 days	Wed 9/6/17	Mon 9/11/17	284
286			NH DOE reviews 1st draft	5 days	Tue 9/12/17	Mon 9/18/17	285
287			Incorporate NH DOE comments	2 days	Tue 9/19/17	Wed 9/20/17	286
288			NH DOE reviews 2nd draft	5 days	Thu 9/21/17	Wed 9/27/17	287
289			PPT finalized	1 day	Thu 9/28/17	Thu 9/28/17	288
290			Face-to-Face Trainings	4 days	Tue 10/17/17	Fri 10/20/17	
291			**Webinars**	**126 days**	**Tue 10/10/17**	**Tue 4/3/18**	
292			Technology Director Training Webinar	1 day	Tue 10/10/17	Tue 10/10/17	
293			District Administrator Webinar	1 day	Tue 10/17/17	Tue 10/17/17	
294			Test Administrator Webinar	1 day	Tue 10/24/17	Tue 10/24/17	
295			Systems Webinar	1 day	Tue 10/31/17	Tue 10/31/17	
296			Online Reporting System Webinar #1 (Assessment Results Training - Interim Assessments)	1 day	Tue 11/28/17	Tue 11/28/17	266FF+5 days
297			Online Reporting System Webinar #2 (Assessment Results Training - Summative Assessments)	1 day	Tue 4/3/18	Tue 4/3/18	276FF+5 days
298			**Training Materials**	**26 days**	**Mon 8/14/17**	**Mon 9/18/17**	
299			**End of Life Support Document**	**16 days**	**Mon 8/14/17**	**Mon 9/4/17**	
300			AIR drafts EOL document	5 days	Mon 8/14/17	Fri 8/18/17	
301			NH DOE reviews/provides feedback	5 days	Mon 8/21/17	Fri 8/25/17	300
302			AIR updates EOL document based on NH DOE feedback	5 days	Mon 8/28/17	Fri 9/1/17	301
303			EOL document posted to portal	1 day	Mon 9/4/17	Mon 9/4/17	302
304			**System Requirements Document**	**16 days**	**Mon 8/14/17**	**Mon 9/4/17**	
305			AIR drafts Systems Requirements document	5 days	Mon 8/14/17	Fri 8/18/17	
306			NH DOE reviews/provides feedback	5 days	Mon 8/21/17	Fri 8/25/17	305

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task		External Milestone		Manual Summary Rollup	
Split		Inactive Task		Manual Summary	
Milestone		Inactive Milestone		Start-only	
Summary		Inactive Summary		Finish-only	
Project Summary		Manual Task		Deadline	
External Tasks		Duration-only		Progress	

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
307			AIR updates Systems Requirements document based on NH DOE feedback	5 days	Mon 8/28/17	Fri 9/1/17	306
308			AIR posts System Requirements document to portal	1 day	Mon 9/4/17	Mon 9/4/17	307
309			**Technical Specifications Document**	**16 days**	**Mon 8/28/17**	**Mon 9/18/17**	
310			AIR drafts Tech Specs document	5 days	Mon 8/28/17	Fri 9/1/17	
311			NH DOE reviews/provides feedback	5 days	Mon 9/4/17	Fri 9/8/17	310
312			AIR updates Tech Specs document based on NH DOE feedback	5 days	Mon 9/11/17	Fri 9/15/17	311
313			AIR posts Tech Specs document to portal	1 day	Mon 9/18/17	Mon 9/18/17	312
314			**Secure Browser Installation**	**16 days**	**Mon 8/21/17**	**Mon 9/11/17**	
315			AIR drafts SB Installation document	5 days	Mon 8/21/17	Fri 8/25/17	
316			NH DOE reviews/provides feedback	5 days	Mon 8/28/17	Fri 9/1/17	315
317			AIR updates SB Installation document based on NH DOE feedback	5 days	Mon 9/4/17	Fri 9/8/17	316
318			AIR posts SB Installation document to portal	1 day	Mon 9/11/17	Mon 9/11/17	317
319			**User Guides and Manuals**	**31 days**	**Mon 10/2/17**	**Mon 11/13/17**	
320			**Test Coordinator Manual (TCM)**	**26 days**	**Mon 10/2/17**	**Mon 11/6/17**	
321			AIR drafts TCM	10 days	Mon 10/2/17	Fri 10/13/17	
322			NH DOE reviews/provides feedback	10 days	Mon 10/16/17	Fri 10/27/17	321
323			AIR updates TCM based on NH DOE feedback	5 days	Mon 10/30/17	Fri 11/3/17	322
324			AIR post TCM to portal	1 day	Mon 11/6/17	Mon 11/6/17	323
325			**Test Administration Manual (TAM)**	**26 days**	**Mon 10/2/17**	**Mon 11/6/17**	
326			AIR drafts TAM	10 days	Mon 10/2/17	Fri 10/13/17	
327			NH DOE reviews/provides feedback	10 days	Mon 10/16/17	Fri 10/27/17	326
328			AIR updates TAM based on NH DOE feedback	5 days	Mon 10/30/17	Fri 11/3/17	327
329			AIR post TAM to portal	1 day	Mon 11/6/17	Mon 11/6/17	328

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
330			**TIDE User Guide**	**26 days**	**Mon 10/2/17**	**Mon 11/6/17**	
331			AIR drafts TIDE UG	10 days	Mon 10/2/17	Fri 10/13/17	
332			NH DOE reviews/provides feedback	10 days	Mon 10/16/17	Fri 10/27/17	331
333			AIR updates TIDE UG based on NH DOE feedback	5 days	Mon 10/30/17	Fri 11/3/17	332
334			AIR post TIDE UG to portal	1 day	Mon 11/6/17	Mon 11/6/17	333
335			**TA User Guide**	**26 days**	**Mon 10/9/17**	**Mon 11/13/17**	
336			AIR drafts TA UG	10 days	Mon 10/9/17	Fri 10/20/17	
337			NH DOE reviews/provides feedback	10 days	Mon 10/23/17	Fri 11/3/17	336
338			AIR updates TA UG based on NH DOE feedback	5 days	Mon 11/6/17	Fri 11/10/17	337
339			AIR post TA UG to portal	1 day	Mon 11/13/17	Mon 11/13/17	338
340			**ORS User Guide**	**26 days**	**Mon 10/9/17**	**Mon 11/13/17**	
341			AIR drafts ORS UG	10 days	Mon 10/9/17	Fri 10/20/17	
342			NH DOE reviews/provides feedback	10 days	Mon 10/23/17	Fri 11/3/17	341
343			AIR updates ORS UG based on NH DOE feedback	5 days	Mon 11/6/17	Fri 11/10/17	342
344			AIR post ORS UG to portal	1 day	Mon 11/13/17	Mon 11/13/17	343
345			**THSS User Guide**	**26 days**	**Mon 10/9/17**	**Mon 11/13/17**	
346			AIR drafts THSS UG	10 days	Mon 10/9/17	Fri 10/20/17	
347			NH DOE reviews/provides feedback	10 days	Mon 10/23/17	Fri 11/3/17	346
348			AIR updates THSS UG based on NH DOE feedback	5 days	Mon 11/6/17	Fri 11/10/17	347
349			AIR post THSS UG to portal	1 day	Mon 11/13/17	Mon 11/13/17	348
350			**Data Files**	**140 days**	**Mon 12/18/17**	**Fri 6/29/18**	
351			Scoring Specfications Finalized	1 day	Mon 12/18/17	Mon 12/18/17	
352			Data File Layout Finalized	1 day	Mon 12/18/17	Mon 12/18/17	
353			Analysis team deploys test scoring engine and configuration to production	1 day	Mon 1/15/18	Mon 1/15/18	
354			Tech team verifies scoring	20 days	Tue 1/16/18	Mon 2/12/18	353

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task		External Milestone		Manual Summary Rollup	
Split		Inactive Task		Manual Summary	
Milestone		Inactive Milestone		Start-only	
Summary		Inactive Summary		Finish-only	
Project Summary		Manual Task		Deadline	
External Tasks		Duration-only		Progress	

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
355			Analysis team builds DFG machine layout	20 days	Tue 1/16/18	Mon 2/12/18	353
356			AIR generates preliminary file #1	1 day	Tue 4/10/18	Tue 4/10/18	
357			AIR generates preliminary file #2	1 day	Tue 5/8/18	Tue 5/8/18	356SS+20 days
358			Deliver Final Data File	1 day	Fri 6/29/18	Fri 6/29/18	
359			**Standard Setting**	**276 days**	**Mon 12/18/17**	**Mon 1/7/19**	
360			**Standard Setting Plan**	**44 days**	**Mon 1/8/18**	**Thu 3/8/18**	
361			Create Standard Setting Plan	30 days	Mon 12/18/17	Fri 1/26/18	
362			NH DOE Review Plan	5 days	Mon 1/29/18	Fri 2/2/18	361
363			Apply Changes/Edits	2 days	Mon 2/5/18	Tue 2/6/18	362
364			NH DOE Review Revised Plan	3 days	Wed 2/7/18	Fri 2/9/18	363
365			Apply Changes/Edits	2 days	Mon 2/12/18	Tue 2/13/18	364
366			NH DOE Review and Approval of Standard Setting Plan	2 days	Wed 2/14/18	Thu 2/15/18	365
367			Standard Setting Plan Approved	0 days	Fri 2/16/18	Fri 2/16/18	366
368			**Standard Setting Logistics**	**147 days**	**Mon 12/18/17**	**Tue 7/10/18**	
369			**Recruit Participants**	**60 days**	**Mon 12/18/17**	**Fri 3/9/18**	
370			Identify Participants	5 days	Fri 2/16/18	Thu 2/22/18	367
371			Recruit Participants	40 days	Fri 2/23/18	Thu 4/19/18	370
372			NH DOE Review Participant List	5 days	Fri 4/20/18	Thu 4/26/18	371
373			Apply Changes/Modifications	5 days	Fri 4/27/18	Thu 5/3/18	372
374			NH DOE Review and Approve Participant List	5 days	Fri 5/4/18	Thu 5/10/18	373
375			Register Participants	15 days	Fri 5/11/18	Thu 5/31/18	374
376			**Standard Setting Training Materials and Handouts**	**35 days**	**Mon 12/18/17**	**Fri 2/2/18**	
377			Prepare Standard Setting Training Materials and Handouts	20 days	Mon 6/25/18	Fri 7/20/18	392SS-40 days
378			NH DOE Review Standard Setting Training Materials and Handouts	5 days	Mon 7/23/18	Fri 7/27/18	377
379			Apply Changes/Edits	3 days	Mon 7/30/18	Wed 8/1/18	378

Project: NH Statewide Assessmen
Date: Fri 4/21/17

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

ID	Indicators	Task Mode	Task Name	Duration	Start	Finish	Predecessors
380			NH DOE Review Revised Materials	3 days	Thu 8/2/18	Mon 8/6/18	379
381			Apply Changes/Edits	2 days	Tue 8/7/18	Wed 8/8/18	380
382			NH DOE Review and Approve Standard Setting Training Materials and Handouts	2 days	Thu 8/9/18	Fri 8/10/18	381
383			Standard Setting Training Materials and Handouts Approved	0 days	Mon 8/13/18	Mon 8/13/18	382
384			**Data Analysis**	**67 days**	**Mon 12/18/17**	**Tue 3/20/18**	
385			ELA Machine-Scored Items Available	0 days	Thu 6/7/18	Thu 6/7/18	7FF
386			Math Machine-Scored Items Available	0 days	Thu 6/7/18	Thu 6/7/18	7FF
387			ELA Hand-Scored Items Available	0 days	Thu 6/21/18	Thu 6/21/18	7FF+10 days
388			Math Hand-Scored Items Available	0 days	Thu 6/21/18	Thu 6/21/18	7FF+10 days
389			Science Machine-Scored Items Available	1 day?	Thu 6/21/18	Thu 6/21/18	7FF+10 days
390			Science Hand-Scored Items Available	1 day?	Thu 6/21/18	Thu 6/21/18	7FF+10 days
391			Analyze Data	20 days	Fri 6/22/18	Thu 7/19/18	390
392			Conduct Standard Setting	10 days	Mon 8/20/18	Fri 8/31/18	
393			**Standard Setting Report**	**28 days**	**Mon 12/18/17**	**Wed 1/24/18**	
394			Prepare Standard Setting Report	15 days	Mon 8/27/18	Fri 9/14/18	392FF+10 days
395			NH DOE Review of Standard Setting Report	5 days	Mon 9/17/18	Fri 9/21/18	394
396			Apply Changes/Edits	5 days	Mon 9/24/18	Fri 9/28/18	395
397			NH DOE Approves Standard Setting Report	3 days	Mon 10/1/18	Wed 10/3/18	396

Task	External Milestone	Manual Summary Rollup
Split	Inactive Task	Manual Summary
Milestone	Inactive Milestone	Start-only
Summary	Inactive Summary	Finish-only
Project Summary	Manual Task	Deadline
External Tasks	Duration-only	Progress

Appendix E:
Test Information Distribution Engine (TIDE)
User Acceptance Testing (UAT) Document

New Hampshire Smarter Balanced Assessment South Dakota Smarter Balanced Assessment

Test Information and Distribution Engine

User Acceptance Testing

2016–2017

Published September 1, 2016

Prepared by the American Institutes for Research®

Table of Contents

Section I. Introduction

The purpose of this document is to ensure that the Test Information Distribution Engine (TIDE) works according to New Hampshire Department of Education's (NHDOE) expectations. This document can be followed conduct user acceptance testing (UAT) of the application. Testers may conduct any additional testing as needed.

Assumptions

- For the purposes of this UAT, AIR will use demo student data on a secure site.
- The test cases provided in this document are a subset of the ones used by AIR.
- AIR will continue to test the system during the course of this UAT.
- AIR will try to deploy any fixes to the UAT site during off-business hours.
- Only the user credentials appearing in Table 1 should be used to access the UAT website.

UAT Timetable

Milestone	Date
AIR provides UAT site to NHDOE	September 15, 2016
UAT Period – Round 1	September 15–19, 2016
UAT Period – Round 2	September 20–21, 2016
Go Live for TIDE	October 6, 2016

UAT Workflow Assumptions

- NHDOE may provide feedback on the testing progress and a list of any issues.
- AIR will try to deploy fixes for minor issues ASAP or will wait for the UAT cycle to be completed before deploying fixes.
- AIR will communicate with NHDOE before any UAT code deployment to prevent interruptions to the UAT process. AIR will also provide a list of bugs that are being fixed as part of that release.

UAT Environment

URL

The UAT site is available at the following URL:
https://tide.uat.airast.org/apps/NEAC/TIDE/Login/

Users

Table 1 lists usernames and passwords that have been added to schools in district `Demo District 1`.

Table 1. User Roles Available for UAT

Role	Description	Username	Password
STATE	State administrator	NH-STATE1@demo.user	UAT@bc$56
DC	District test coordinator	NH-DC1@demo.user	UAT@bc$56
SC	School test coordinators	NH-SC1@demo.user	UAT@bc$56
TE	Teacher	NH-TE1@demo.user	UAT@bc$56
TA	Test Administrator	NH-TA1@demo.user	UAT@bc$56

Known Issues

As of this time, the following issues exist in the UAT environment:

- TBD

Section II. Test Cases

This section describes test cases that are guidelines for overall UAT.

Verify Global Features

This test is available to all user roles.

Action	Expected Result	Pass/Fail
Log in to TIDE with a valid user account.	The following appears in the banner. • Role: User Role • Administration In addition, there are links for resources, help, settings, help desk contact info, and logout.	
Click **Contact Help Desk** in the footer.	A window appears showing Help Desk contact information.	
In the banner, click **Help**.	A new browser tab or window opens showing the first page of the help file.	
In the banner, click **Manage Account > Change Role**.	The Change Role modal window appears. Verify that the values in the Role, State, District, and School drop-down lists are the only ones available for your user role. Test this feature when logged in as all of the users listed in Table 1. This verifies that users at different levels of the hierarchy have access to all and only the corresponding schools and districts.	
In the banner, click **Manage Account > My Contact**.	The My Contact Information modal window appears. Modify fields as appropriate. Close the window and then reopen, modified values appear.	

Test Cases Pertaining to Users

The tests in this section pertain to tasks under the **Users** accordion and the **Settings** drop-down list.

View and Export Users

This test is available to the STATE, DC, and SC roles.

Action	Expected Result	Pass/Fail
Open the **Users** accordion and click **View/Edit/Export Users**.	The **View/Edit/Export Users** page appears.	
Verify the search fields that appear on the **View/Edit/Export Users** page.	The following fields appear: • Role • E-mail Address • First Name • Last Name • Phone • TA Certified • Search (button)	
Select a role.	The State, District, and School drop-down lists appear, depending on your user role and the role you selected. The options in those drop-down lists match your user role as well. For example, if you are a district-level user, the options in the district drop-down list are related to you, and the options in the school drop-down list are related to your districts.	
Verify that you can search for users: 1. Select from the State, District, and School drop-down lists. 2. Enter an existing first name, last name, e-mail address, or telephone number. 3. Click **Search**.	The Search Results table populates with at least one search result.	

Action	Expected Result	Pass/Fail
Verify that the Search Results table populates correctly.	The listing includes the following: • Export button • Delete button • Print button • Number of Users (# count) • Table with the following columns (the actual columns displayed may be different depending on your user role and the role you selected): o Checkbox o Edit button o Role o District o E-mail address o First Name o Last Name o Phone	
Verify that all results can be exported. Mark the checkbox in the table header, click , and click **PDF**, **Excel**, or **CSV**.	A dialog box appears prompting to open or save the file of search results.	
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	
Verify that selected results can be exported: 1. Click checkboxes for some users. 2. Click and click **PDF**, **Excel**, or **CSV**.	A dialog box appears prompting the user to open or save the file of search results.	
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	

Add Users

This test is available to the STATE, DC, and SC roles.

Action	Expected Result	Pass/Fail
Open the **Users** accordion and click **Add Users**.	The Add Users page appears.	
Fill out the form and click **Save**.	A confirmation message appears.	
Open the **Users** accordion, click **View/Edit/Export Users**, and search for the new user.	The new user appears in the retrieval.	

Modify Users

This test is available to the STATE, DC, and SC roles. Ideally, complete this test after completing the Add Users test.

Action	Expected Result	Pass/Fail
Open the **Users** accordion and click **View/Edit/Export Users**.	The View/Edit/Export Users page appears.	
Retrieve an existing user.	The user's record appears in the View/Edit/Export Users page.	
Click corresponding to the user.	The Edit User page displays the user's details.	
Modify one of the fields and click **Save**.	A confirmation message appears.	
Click **Continue**.	The user appears in the table with the modified value.	

Delete Users

This test is available to the STATE, DC, and SC roles. Ideally, complete this test after completing the Add Users test.

Action	Expected Result	Pass/Fail
Open the **Users** accordion and click **View/Edit/Export Users**.	The View/Edit/Export Users page appears.	
Retrieve an existing user.	The user's record appears in the View/Edit/Export Users page.	
Mark the checkbox corresponding to the user.	The Delete button above the table becomes active.	

Action	Expected Result	Pass/Fail
Click .	A pop-up window appears asking for confirmation that you want to delete the selected user.	
Click **Yes**.	The user is deleted from TIDE (and all related systems). Verify the deletion by searching for the deleted user. The deleted user does not appear in the search results.	

Upload Users

This test is available to the STATE, DC, and SC roles.

Action	Expected Result	Pass/Fail
Open the **Users** accordion and click **Upload Users**.	The Upload Users page appears.	
Click **Download Templates** and select **Excel** or **CSV**.	The browser downloads or opens the file.	
Review the headings in the downloaded file.	The headings match those in the specs or in the user guide.	
If downloaded an Excel file and if the Excel file has validations, display those validations by clicking with the mouse.	The validations match the valid values listed in the specs or in the user guide.	
Compose the upload file.	N/A	
Click **Browse** and navigate to the upload file.	N/A	
Click **Next**.	A preview page appears listing the values.	
Click **Next**.	A validation report appears.	
Click **Continue with Upload**.	A confirmation report appears indicating records that were committed.	
Using the test View and Export Users, retrieve one of the users included in the upload file.	Uploaded record appears in the retrieval.	

Test Cases Pertaining to Student Information

The tests in this section pertain to tasks under the **Students** and **Print Test Tickets** accordions.

View, Edit, and Export Student Records

This test is available to all user roles.

Action	Expected Result	Pass/Fail
Open the **Test Settings and Tools** accordion and click **View/Edit/Export Test Settings and Tools**.	The View/Edit/Export Test Settings and Tools page appears.	
Verify the search fields that are displayed on the View/Edit/Export Test Settings and Tools page.	The following fields appear (list varies depending on your user role): • State • District • School • Last Name • First Name • SSID • Grade Level When Assessed	
Verify that you can search for students: 1. Select a state, district, and school as available for your user role. 2. Add additional search criteria as appropriate. 3. Click **Search**.	The Search Results table populates with records matching the search criteria (if any).	

Action	Expected Result	Pass/Fail
Verify that the Search Results table populates correctly.	The page includes the following: • Export button • Delete button • Print button • Number of students • Table with the following columns: o Checkbox o button o District o School IRN o Last Name o First Name o Middle Name o Birth Date (MMDDYYYY) o SSID o Accommodations (as appropriate)	
Verify that you can export all students. 1. Click the checkbox in the table header. 2. Click and click **Excel** or **CSV**.	A dialog box appears prompting to open or save the file of search results.	
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	
Verify that selected results can be exported: 1. Click checkboxes for some students. 2. Click and click **Excel** or **CSV**.	A dialog box appears prompting to open or save the file of search results.	
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	
Click for a student.	The student's record appears in the Edit Student form.	
Modify a few fields and click **Save**.	You return to the View/Edit/Export Students page. If field is visible in the listing, its updated value appears.	
Click for the student you just changed.	The updated settings appear in the Edit Student form.	

Print Test Tickets and PreID Labels

This test is available to all user roles.

Action	Expected Result	Pass/Fail
Open the **Students** accordion and click **View/Edit/Export Students**.	The View/Edit/Export Students page appears.	
Retrieve students.	The students' records appear in the View/Edit/Export Students page.	
Select a few students, click , and click **Test Tickets**.	Pop-up appears displaying a model of labels on a page.	
Click one of the dimensions, such as 3×2.	Model's display changes depending on selection. For example, if selected 3×2, the labels are positioned 3 down and 2 across.	
Click **Print**.	Browser downloads a PDF.	
Open the downloaded file.	File is a test ticket file with labels in the position of the selected model. For example, if selected 3×2, the labels are positioned 3 down and 2 across.	
Close the model.	N/A	
Select at least five students, click , and click **PreID labels**.	Pop-up displays a model of labels on a page.	
Click the number of the starting position, such as 2.	Model's display changes depending on selection.	
Click **Print**.	Browser downloads a PDF.	
Open the downloaded file.	File is a PreID file that includes the tests for the selected students. The first label on the first page is in the selected starting position; first label on subsequent pages is in position 1.	
Close the model.	N/A	

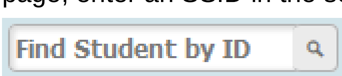
Upload Students' Test Settings

This test is available to the STATE, DC, and SC roles.

Action	Expected Result	Pass/Fail
Open the **Test Settings and Tools** accordion and click **Upload Test Settings and Tools**.	The Upload Test Settings and Tools page appears.	
Click **Download Templates** and select **Excel** or **CSV**.	The browser downloads or opens the file.	
Review the headings in the downloaded file.	The headings match those in the specs or in the user guide.	
If downloaded an Excel file and if the Excel file has validations, display those validations by clicking with the mouse.	The validations match the valid values listed in the specs or in the user guide.	
Compose the upload file.	N/A	
Click **Browse** and navigate to the upload file.	N/A	
Click **Next**.	A preview page appears listing the values.	
Click **Next**.	A validation report appears.	
Click **Continue with Upload**.	A confirmation report appears indicating records that were committed.	
Using the test View, Edit, and Export Student Records retrieve one of the test settings included in the upload file.	Uploaded record appears in the retrieval.	

Search for Student by SSID

This test is available to the STATE roles. Before starting this test, use the View, Edit, and Export Student Records test to obtain an existing SSID.

Action	Expected Result	Pass/Fail
In the upper-right corner of the home page, enter an SSID in the search field. 	N/A	
Click .	The Edit Student page displays the student's demographics and test settings.	

Test Cases Pertaining to Orders

The tests in this section pertain to tasks under the **Paper Orders** and **Track Return Packages** accordions.

Creating Additional Orders

This test is available to the DC role.

<table>
<tr><th>Action</th><th>Expected Result</th><th>Pass/Fail</th></tr>
<tr><td>Open the Paper Ordering accordion and click Additional Orders.</td><td>The Additional Orders page appears.</td><td></td></tr>
<tr><td>Mark the District radio button and click Search.</td><td>The following sections appear:
Shipping Address
Comments
Table of items included in the order. The table is sorted by grade and has the following columns:
Material Description
Quantity You Will Receive
Quantity Approved
Quantity Pending Approval
Additional Quantity</td><td></td></tr>
<tr><td>In the Additional Quantity column, enter a quantity for an item and click Save Orders.</td><td>A prompt appears for entering a comment.</td><td></td></tr>
<tr><td>Enter a comment and click Submit.</td><td>The Quantity Ordered field returns to zero.
The Quantity Pending Approval field increases by the amount entered.</td><td></td></tr>
</table>

Review Order History, Tracking, and Reports

This test is available to the DC role.

Action	Expected Result	Pass/Fail
Open the **Paper Ordering** accordion and click **Order History**.	The Order History page appears.	
Mark the **District** radio button and click **Summary**.	A list of orders for the district appears. The listing includes the following: • Order Number • Order Type • Submitted By • Order Status • Submitted Date • Tracking • Reports	
In the Order column, click an order number.	The Order Details page appears. The sections included are Order Information, Shipping Information, and Order Quantity. (Other sections are also available.)	
Click **Close**.	The Order Details page closes.	
On the Order History page, click .	The Order Details page appears, this time with the Track Shipments section expanded.	
Click a tracking number.	The tracking information appears in a new browser window or tab, sourced from the shipping company's web site.	
Click **Close**.	The Order Details page closes.	
On the Order History page, click .	The Order Details page appears, this time with the Track Documents section expanded.	
Click a PDF link.	Browser offers to download the PDF. The PDF's contents match the hint in the link (packing list, security checklist, pallet manifest).	
Click **Close**.	The Order Details page closes.	

Review Order Quantities

This test is available to the DC role.

Action	Expected Result	Pass/Fail
Open the **Paper Ordering** accordion and click **View Order Quantity Reports**.	The View Order Quantity Reports page appears.	
Mark the **District** radio button, from the Search Order By drop-down list mark the **Initial Order** and **Additional Order** checkboxes, and then click **Search**.	A list of orders for the district appears. The listing includes the following: • District Expected Shipment • District Awaiting Approval These columns appear for the total level, district level, and school level.	

Review Order Quantity by Material Type

This test is available to the STATE role.

Action	Expected Result	Pass/Fail
Open the **Paper Ordering** accordion and click **Quantity By Material Type**.	The Quantity By Material Type page appears.	
1. From **Institution Type** drop-down list, select District. 2. Mark the **On-Time** and **Additional** checkboxes. 3. From the **Material** drop-down list, select one of the materials. 4. Click **Search**.	A list of orders for the selected material appears.	
Verify that selected results can be exported. Click  and select a file format.	The browser downloads a file containing the order listing.	

Review State-Level Order Quantity

This test is available to the STATE role.

Action	Expected Result	Pass/Fail
Open the **Paper Ordering** accordion and click **View State Quantity Reports**.	The State Quantity page appears.	
1. Mark the **Initial Order** and **Additional** checkboxes. 2. From the **Test Administration** drop-down list, select a test administration. 3. From the **District** drop-down list, select a district. 4. Click **Search**.	A list of orders for the selected test administration and district appears.	

Test Cases Pertaining to Rosters

The tests in this section pertain to tasks under the **Rosters** and **Print Test Tickets** accordions.

Add Rosters

This test is available to the STATE, DC, SC, and TE roles.

Action	Expected Result	Pass/Fail
Open the **Rosters** accordion and click **Add Roster**.	The Add Roster page appears.	
Search for students by making selections for district, school, grade, and other criteria. Click **Search.**	A list of students matching the search criteria appears. These students have the following characteristics: • They are enrolled in the selected school ID-grade combination. • They match the other selection fields, such as test settings.	
Add a roster name and select a teacher.	Available teachers in the Teacher Name drop-down list are those in the selected school.	
Select at least one student in the Available Students List and click **Add Selected**.	Students' names disappear from the Available Students list and appear in the Students in Roster list.	
Click **Save**.	Confirmation message appears.	

View and Edit Rosters

This test is available to the STATE, DC, SC, and TE roles. Ideally, complete the Add Rosters test before performing this test.

Action	Expected Result	Pass/Fail
Open the **Rosters** accordion and click **View/Edit/Export Roster**.	The View/Edit/Export Roster page appears.	
Select a district and school as available for your user role and click **Search**. (Recommend selecting the district and school used in the Add Rosters test.)	The Search Results table populates with records matching the search criteria.	

Verify that the Search Results table populates correctly.	The page includes the following:
	• Delete button
	• Print button
	• Number of rosters
	• Table with the following columns:
	  o Checkbox
	  o Edit button
	  o Roster Name
	  o Subject
	  o Grades in Roster
	  o Number of Students	

Retrieve an existing roster.	The roster's record appears in the View/Edit/Export Roster page.	
Click corresponding to the roster.	The Edit Roster page displays the roster's details.	
Search for students by making selections for district, school, grade, and other criteria. Click **Search**.	A list of students matching the search criteria appears. These students have the following characteristics: • They are enrolled in the selected school ID-grade combination. • They match the other selection fields, such as test settings.	
Select at least one student in the **Available Students** list and click **Add Selected**.	Students names disappear from the Available Students list, appear in the Students in Roster list.	
Select at least one student in the **Students in Roster** list and click **Remove Selected**.	Students names disappear from the Students in Roster list, appear in the Available Students list.	
Click **Save**.	Confirmation message appears.	

Print Test Tickets and Test Settings for Students in a Roster

This test is available to the STATE, DC, SC, and TE roles. Ideally, complete the Add Rosters test before performing this test.

Action	Expected Result	Pass/Fail
Open the **Print Test Tickets** accordion and click **View/Edit/Export Roster**.	The View/Edit/Export Roster page appears.	
Select a district and school as available for your user role and click **Search**. (Recommend selecting the district and school used in the Add Rosters test.)	The Search Results table populates with records matching the search criteria.	
Select a few rosters, click , and click **Test Tickets**.	Pop-up displays a model of labels on a page.	
Click one of the dimensions, such as 3×2.	Model's display changes depending on selection. For example, if selected 3×2, the labels are positioned 3 down and 2 across.	
Click **Print**.	Browser downloads a PDF.	
Open the downloaded file.	File is a test ticket file with labels in the position of the selected model. For example, if selected 3×2, the labels are positioned 3 down and 2 across. File contains test tickets for the selected rosters.	
On the View/Edit/Export Roster page, select a few rosters, click , and click **Student Settings and Tools**.	Report appears listing students and associated test settings and tools.	
Click **Print**.	Browser downloads a PDF.	
Open the downloaded file.	File matches that in report.	

Delete Rosters

This test is available to the STATE, DC, SC, and TE roles. Ideally, complete the Add Rosters test before performing this test.

Action	Expected Result	Pass/Fail
Open the **Rosters** accordion and click **View/Edit/Export Roster**.	The View/Edit/Export Roster page appears.	
Retrieve an existing roster.	The roster's record appears in the View/Edit/Export Roster page.	
Mark the checkboxes next to the roster and click .	A confirmation message appears.	

Action	Expected Result	Pass/Fail
Click **OK** in the confirmation message.	The roster listing appears. The deleted roster does not appear in the listing.	

Upload Roster

This test is available to the STATE, DC, SC, and TE roles.

Action	Expected Result	Pass/Fail
Open the **Rosters** accordion and click **Upload Rosters**.	The Upload Roster page appears.	
Click **Download Templates** and select **Excel** or **CSV**.	The browser downloads or opens the file.	
Review the headings in the downloaded file.	The headings match those in the specs or in the user guide.	
If downloaded an Excel file and if the Excel file has validations, display those validations by clicking with the mouse.	The validations match the valid values listed in the specs or in the user guide.	
Compose the upload file.	N/A	
Click **Browse** and navigate to the upload file.	N/A	
Click **Next**.	A preview page appears listing the values.	
Click **Next**.	A validation report appears.	
Click **Continue with Upload**.	A confirmation report appears indicating records that were committed.	
Using the test View and Edit Rosters, retrieve one of the rosters included in the upload file.	Uploaded record appears in the retrieval.	

Test Cases Pertaining to Appeals

The tests in this section pertain to tasks under the **Appeals** accordion.

Create Appeals

This test is available to the STATE, DC, and SC roles. Before performing this test, acquire a Result ID, Session ID, or SSID for which you can create an appeal. In addition, use the admin tool to identify an appeal type with a reason that requires additional comments.

Action	Expected Result	Pass/Fail
Open the **Appeals** accordion and click **Create Requests**.	The Create Requests page appears.	
Observe the available appeal types.	The available appeal types match those in the specifications.	
Select an appeal type with a reason that does not require additional comments.	N/A	
From Search Student By drop-down list, select a value.	The appropriate field appears. For example, if selecting `Result ID` from the drop-down list, a field for entering the Result ID appears.	
Enter a target for the Result ID, Session ID, or SSID. Click **Search**.	The matching test Result IDs appear.	
Mark checkboxes for one or more of the retrieved results and click **Create**.	An prompt appears for entering a reason for the appeal.	
Enter a reason for the appeal.	N/A	
Click **Submit**.	The prompt closes; appeal appears in a list of Result IDs.	
Enter a different target for the Result ID, Session ID, or SSID. Click **Search Student Results**.	The matching test Result IDs appear.	
For one of the retrieved results, click **Create**.	An **Appeal Reason** layer appears.	
From the **Comments** drop-down list, select a reason that requires additional comments.	A prompt for entering comments appears.	
Enter the comments and an external reference ID.	N/A	
Click **Submit**.	Appeal Reason layer closes; appeal appears in a list of Result IDs.	

View, Export, and Print Appeals

This test is available to the STATE, DC, and SC roles. Ideally, complete the Create Appeals test before performing this test.

Action	Expected Result	Pass/Fail
Open the **Appeals** accordion and click **View/Approve/Export Requests**.	The View/Approve/Export Requests page appears.	
Select an appeal type and other selection criteria. For Request Status, mark **Approved** and **Denied**.	N/A	
Click **Search**.	The matching appeals appear. The Status column displays only Processed and Pending Approval.	
Click the balloon for an appeal that has status as "processed."	Comments regarding the approver's action appears.	

Upload Appeals

This test is available to the STATE, DC, and SC roles.

Action	Expected Result	Pass/Fail
Open the **Appeals** accordion and click **Upload Requests**.	The Upload Requests page appears.	
Click **Download Templates** and select **Excel** or **CSV**.	The browser downloads or opens the file.	
Review the headings in the downloaded file.	The headings match those in the specs or in the user guide.	
If downloaded an Excel file and if the Excel file has validations, display those validations by clicking with the mouse.	The validations match the valid values listed in the specs or in the user guide.	
Compose the upload file.	N/A	
Click **Browse** and navigate to the upload file.	N/A	
Click **Next**.	A preview page appears listing the values.	
Click **Next**.	A validation report appears.	
Click **Continue with Upload**.	A confirmation report appears indicating records that were committed.	
Using the test View, Export, and Print Appeals, retrieve one of the appeals included in the upload file.	Uploaded record appears in the retrieval.	

Retract Appeals

This test is available to the STATE, DC, and SC roles. Before starting this test, ensure that there are some appeals with status Pending Approval available for your user role.

Action	Expected Result	Pass/Fail
Open the **Appeals** accordion and click **View/Approve/Export Requests**.	The View/Approve/Export Requests page appears.	
Select an appeal type and other selection criteria. For Request Status, mark **Pending**.	N/A	
Click **Search**.	The matching appeals appear. The Status column displays only Pending Approval.	
Mark the checkbox for an appeal.	The Process button becomes enabled.	
Click **Process** and then click **Retract**.	A prompt for entering comments appears.	
Enter a prompt and click **Submit**.	Prompts closes. Returning to the list of appeals, the retracted appeal does not appear.	

Approve or Deny Appeals

This test is available to the STATE role. Before starting this test, ensure that there are some appeals with status Pending Approval available for your user role.

Action	Expected Result	Pass/Fail
Open the **Appeals** accordion and click **View/Approve/Export Requests**.	The View/Approve/Export Requests page appears.	
Select an appeal type and other selection criteria. For Request Status, mark **Pending Approval**.	N/A	
Click **Search**.	The matching appeals appear. The Status column displays only Pending Approval.	
Mark the checkbox for an appeal.	The Process button becomes enabled.	
Click **Process** and then click **Approve** or **Reject**.	A prompt for entering comments appears.	
Enter a comment and click **Submit**.	Prompt closes. Returning to the list of appeals, the approved or rejected appeal does not appear.	

Test Cases Pertaining to Monitoring Test Progress

The tests in this section pertain to tasks under the **Monitoring Test Progress** accordion.

View, Export, and Print Students Not Yet Tested

This test is available to all user roles.

Action	Expected Result	Pass/Fail
Open the **Monitoring Test Progress** accordion and click **Plan and Manage Testing**.	The Plan and Manage Testing page appears.	
Enter search criteria. In particular, in the Get Specific section, make selections for students who have not completed the first test opportunity.	N/A	
Click **Generate Report**.	The matching students appear. The listed students have not yet started the selected test.	
Verify that results can be exported. Click and click **PDF**, **Excel**, or **CSV**.	A dialog box appears prompting to open or save the file of search results.	
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	
Click .	Browser downloads a PDF. Contents of file match that in the listing.	

View, Export, and Print Test Completion Rates

This test is available to the all user roles.

Action	Expected Result	Pass/Fail
Open the **Monitoring Test Progress** accordion and click **Test Completion Rates**.	The Test Completion Rates page appears.	
Enter search criteria.	N/A	

Action	Expected Result	Pass/Fail
Click **Generate Report**.	The matching tests appear. The following columns are available: • Date • Test Name • Opportunity • Total Student • Total Student Started • Total Student Completed • Percent Started • Percent Completed	
Verify that results can be exported. Click  and click **PDF**, **Excel**, or **CSV**.	A dialog box appears prompting to open or save the file of search results.	
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	
Click .	Browser downloads a PDF. Contents of file match that in the listing.	

View, Export, and Print State Participation Counts

This test is available to the STATE role.

Action	Expected Result	Pass/Fail
Open the **Monitoring Test Progress** accordion and click **State Participation Counts**.	The State Participation Counts page appears.	
Enter search criteria.	N/A	
Click **Generate Report**.	The matching test subjects appear. The following columns are available: • Date • Window • Instrument • Subject • Daily Started Count • Daily Completed Count • Cumulative Started Count • Cumulative Completed Count	
Verify that results can be exported. Click  and click **PDF**, **Excel**, or **CSV**.	A dialog box appears prompting to open or save the file of search results.	

Action	Expected Result	Pass/Fail
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	
Click .	Browser downloads a PDF. Contents of file match that in the listing.	

View, Export, and Print State Test Status Codes

This test is available to the DC and SC roles.

Action	Expected Result	Pass/Fail
Open the **Monitoring Test Progress** accordion and click **Test Status Code Report**.	The Test Status Code Report page appears.	
Enter search criteria.	N/A	
Click **Generate Report**.	The matching students appear. The following columns are available: • Student Name • SSID • Test Name • Test Status • Date Started • Special Code • Assigned School ID • Assigned School Name	
Verify that results can be exported. Click  and click **PDF**, **Excel**, or **CSV**.	A dialog box appears prompting to open or save the file of search results.	
Open the exported file and compare the data with the data in the UI.	The downloaded file includes the same data as in the Search Results table.	
Click .	Browser downloads a PDF. Contents of file match that in the listing.	

Test Cases Pertaining to Data Cleanup

The test in this section pertains to tasks under the **Data Cleanup** accordion.

Add Non-Participation Codes

This test is available to the STATE and DC roles. Before starting this test, ensure that there are students in TIDE who have not yet completed a test.

Action	Expected Result	Pass/Fail
Open the **Data Cleanup** accordion and click **Non-Participation Codes**.	The Non-Participation Codes page appears.	
Under *Search for Non-Participation Codes to Edit*, make selections to retrieve a student who has not yet started or submitted a test.	N/A	
Click **Search**.	Student records matching the search criteria appear.	
Click for a student.	The Edit Non-Participation Codes page appears. The subjects appearing in the *Special Codes* section are those for which the student has not yet started or submitted a test.	
From the drop-down lists, select a special code and then click **Save**.	Returning to the listing in the Non-Participation Codes page, the student no longer appears.	

Test Cases Pertaining to Testing Resources

The tests in this section pertain to tasks available under the **Resources** tab in the banner.

Download Voice Pack

This test is available to all user rules.

Action	Expected Result	Pass/Fail
Open the **Resources** drop-down list and click **Voice Pack**.	The Voice Pack page appears.	
Click **Download** for a voice pack.	The browser downloads the installation file.	
Click **Installing the NeoSpeech™ Voices Packs**.	The browser downloads the installation instructions.	

Download THSS Materials

This test is available to all roles.

Action	Expected Result	Pass/Fail
Open the **Resources** drop-down list and click **Teacher Hand-Scoring System**.	The Teacher Hand-Scoring System page appears.	
Click a link for one of the scoring materials.	The browser downloads the corresponding file.	

Appendix F:
Errata Notice Template

New Hampshire Statewide Assessments

[Document with Error]

Errata Notice [Template]

This errata notice template serves as a demonstrative document for review by the Department in accordance with the requirements in RFP 2017-073, DOE New Hampshire Statewide Assessments, Topic 39 Support Center. AIR would be pleased to customize this errata notice to the specifications of the Department during the initial kickoff meeting.

[Insert description of the error here.]

[Additional detail about the error may appear here, as appropriate.]

[Insert timeline for error resolution here.]

[Insert any required actions that need to be taken by district, school, and/or test administrators here.]

The updated version of the [document title] is available on the New Hampshire Statewide Assessment Portal, here:

http://nh.portal.airast.org/

If you have any questions, please call the New Hampshire Help Desk at 1-844-202-7584 or e-mail NHHelpDesk@air.org.

Appendix G:
AzMERIT Annual Technical Report

AMERICAN INSTITUTES FOR RESEARCH®

Annual Technical Report

Arizona Statewide Assessment in English Language Arts and Math

2015-2016 School Year

February 24, 2017

ARIZONA STATEWIDE ASSESSMENTS

ARIZONA'S MEASUREMENT OF EDUCATIONAL READINESS TO INFORM TEACHING (AzMERIT)

ELA GRADES 3-11

MATH GRADES 3-8, ALGEBRA I, GEOMETRY, AND ALGEBRA II

2015-2016 ANNUAL TECHNICAL REPORT

FEBRUARY 24, 2017

Prepared by American Institutes for Research (AIR) in collaboration with the Arizona Department of Education

TABLE OF CONTENTS

APPENDICES

1. INTRODUCTION: THE VALIDITY OF AZMERIT TEST SCORE INTERPRETATIONS

1.1 OVERVIEW

The purpose of this technical report is to document the evidence supporting the claims made for how AzMERIT test scores may be interpreted. Evidence for the validity of test score interpretations is central to claims that AzMERIT test scores can be used to evaluate the effectiveness with which Arizona districts and schools teach students the Arizona College and Career Ready Standards and whether individual students have achieved those standards by the end of each school year. Thus, the report begins with a review of validity evidence evaluated to date. Evidence for the validity of test score interpretations is expected to accrue over time, so that this section will be expanded as further evidence is gained.

Chapter 2 of the report describes the design and development of the AzMERIT assessment system, including the Arizona College and Career Ready standards which define the content domain to be assessed by AzMERIT, the development of test specifications, including blueprints, that ensure the breadth and depth of the content domain is adequately sampled by the assessments, as well as well as test development procedures that ensure alignment of test forms with the blueprint specifications.

Chapters 3 and 4 provide summaries of the AzMERIT test administrations. Chapter 3 presents results of the fall 2015 administration of the high school end-of-course (EOC) assessments, and Chapter 4 presents results of the spring 2016 administration of the full AzMERIT assessment system, including end of year assessments in ELA and math for grades 3-8, and high school EOC tests in ELA and math. These chapters provide summaries of the test taking student population and their performance on the assessments. In addition, these chapters describe administration specific evidence for the reliability of the AzMERIT assessments, including internal consistency reliability, standard errors of measurement, and the reliability of performance level classifications.

The remaining chapters document technical details of the test development, administration, scoring, and reporting activities. Chapter 5 describes the item development process and especially the sequence of reviews that each item must pass through before being eligible for AzMERIT test administration. This chapter also describes the procedures for constructing test forms from items successfully passing through the review process.

Chapter 6 documents the test administration procedures, including eligibility of participation in the AzMERIT assessments, testing conditions, including accessibility tools and accommodations, systems security for assessments administered online, as well as test security procedures for all test administrations.

A description of the score reporting system and the interpretation of test scores is provided in Chapter 7. Chapter 8 describes the procedures that the Arizona Department of Education used to identify and adopt performance standards for AzMERIT assessments, and Chapter 9 describes the procedures used to scale and equate the AzMERIT assessments for scoring and reporting.

Chapter 10 describes the procedures for scoring constructed-response items, both machine- and hand-scored items, and provides summary rater agreement results. Finally, Chapter 11 provides an overview of the quality assurance processes described throughout that are used to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1.2 VALIDITY EVIDENCE

Validity refers to the degree to which test score interpretations are supported by evidence, and speaks directly to the legitimate uses of test scores. Establishing the validity of test score interpretations is thus the most fundamental component of test design and evaluation. *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the Standards describe the range of evidence that may be brought to bear to support the validity of test score interpretations.

The kinds of evidence required to support the validity of test score interpretations depend centrally on the claims made for how test scores may be interpreted. Moreover, the standards make explicit that validity is not an attribute of tests, but rather of test score interpretations. Some test score interpretations may be supported by validity evidence, while others are not. Thus, the test itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

Central to evaluating the validity of test score interpretations is determining whether the test measures the intended construct. Such an evaluation in turn requires a clear definition of the measurement construct. For Arizona's AzMERIT, the definition of the measurement construct is provided by the Arizona College and Career Ready Standards (ACCRS).

In 2010, Arizona adopted new academic content standards in English language arts (ELA) and math. The Arizona College and Career Ready Standards are designed to ensure that students across grades are receiving the instruction they need to be on track for college and career by the time they graduate.[1] In spring 2015, the Arizona Department of Education (ADE) administered Arizona's Measurement of Educational Readiness to Inform Teaching (AzMERIT) to assess proficiency on the new Arizona College and Career Ready Standards for the first time. The AzMERIT measures English language arts (ELA) and math in grades 3-8 and following completion of high school coursework in ELA Grade 9, ELA Grade 10, ELA Grade 11, Algebra I, Geometry, and Algebra II.

Because directly measuring student achievement against each benchmark in the Arizona College and Career Ready Standards would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the ACCRS.[2] To ensure that each student is assessed on the intended breadth and depth of the ACCRS, test construction is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark.[3] Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand

[1] Standard 1.1 The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.

[2] Standard 4.0 Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

[3] Standard 4.1 Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical. ADE has published the AzMERIT ELA and math test blueprints that specify the distribution of items across reporting strands and depth of knowledge levels. The ELA and math blueprints are also provided as an attachment in Appendix A.

While the blueprints ensure that the full range of the intended measurement construct is represented in each test administration, tests may also inadvertently measure attributes that are not relevant to the construct of interest. For example, when a high level of English language proficiency is necessary to access content in other subject area assessments such as math or science, language proficiency may unnecessarily limit the student's ability to demonstrate achievement in those subject areas. Thus, while such tests may measure achievement of relevant subject area content standards, they may also measure construct irrelevant variation in language proficiency, limiting the generalizability of test score interpretations for some student populations.

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement.[4] Universal design removes barriers to access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

- Inclusive assessment population
- Precisely defined constructs
- Accessible, non-biased items
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

Test development specialists receive extensive training on the principles of universal design and apply these principles in the development of all test materials, including items and accompanying stimuli. In the review process, adherence to the principles of universal design is verified.

In addition, the AzMERIT test delivery system provides a range of accessibility tools and accommodations for reducing construct irrelevant barriers to accessing test content for virtually all students. [5] The range of accommodations, provided in the online testing environment, far exceed the typical accommodations made available in paper-based test administrations. Exhibits 1.2.1-1.2.5 list the accommodations and accessibility supports currently available for students taking the AzMERIT assessments online. Paper test forms are available as

---

[4] Standard 3.0 – All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.
[5] Standard 3.1 – Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
Standard 3.2 – Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.
Standard 12.3 – Those responsible for the development and use of educational assessments should design all relevant steps of the testing process to promote access to the construct for all individuals and subgroups for whom the assessment is intended.

an accommodation for students testing in online schools should the accommodations provided online not be sufficient to remove barriers to accessing test content. These include both larger print and Braille forms, which are also available, for students who need them, in schools administering AzMERIT as a paper-based assessment.. Section 6.3 describes available testing tools and accommodations for students testing online and on paper.

Test administrators are required to provide students with an appropriate testing location that is comfortable and free from distractions. Universal test administration conditions are specific testing situations and environments that may be offered to any student in order to provide a more comfortable and distraction-free testing environment. Universal test administration conditions are available for both paper-based test (PBT) and computer-based testing (CBT) modes. Universal test administration conditions include:

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel,
- Being seated in a specific location within the testing room or being seated at special furniture,
- Having the test administered by a familiar test administrator,
- Using a special pencil or pencil grip,
- Using a place holder,
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting,
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT),
- Using devices that allow the student to hear the test directions: hearing aids and amplification,
- Wearing noise buffers after the scripted directions have been read,
- Signing the scripted directions,
- Having the scripted directions repeated (at student request),
- Having questions about the scripted directions or the directions that students read on their own answered,
- Reading the test quietly to himself/herself as long as other students are not disrupted, and
- Extended time. (Testing session must be competed in the same school day it was started. No student is expected to need more than twice the estimated testing time.)

While some of the items listed as universal test administration conditions might be included in a student's individualized education plan as an accommodation, for AzMERIT testing purposes these are not considered testing accommodations and are available to any student who needs them not just to students with IEPs.

Exhibit 1.2.1 summarizes the Universal Test Tools are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

Exhibit 1.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Allows student to click anywhere on the selected response text or button for multiple choice options.
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Help	View the on-screen *Test Instructions and Help*.

Universal Test Tool	Description
Highlighter	Highlight text in a passage or item.
Line Reader	Allows student to track the line he or she is reading.
Mark (Flag) for Review	Mark an item for review so that it can be easily found later.
Notes/Comments	Allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	Allows the session to be paused at any time and restarted and taken over a one day period. For test security purposes, visibility on past items is not allowed when paused longer than 20 minutes.
Review Test	Allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

AzMERIT testing requires specific subject area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 1.2.2.

Exhibit 1.2.2 Subject Area Tools/Resources Available to All Students

Tool	Applicable Subject Area	Description of Tool
Dictionary/Thesaurus	Writing	CBT – Students have access to the dictionary/thesaurus tool. Students may opt to use a published, paper dictionary or thesaurus instead of using this tool.

Tool	Applicable Subject Area	Description of Tool
		PBT – Schools must make published, paper dictionaries and thesauruses available to students. Students with a visual impairment may use an electronic dictionary and thesaurus with other features turned-off.
Writing Guide	Writing	CBT – Students have access to the writing guide tool. PBT – The writing guide is included within the test booklet.
Scratch Paper	Writing and Math	CBT – Schools must provide scratch paper (plain, lined, or graph) to students. PBT – Schools must provide scratch paper (plain, lined, or graph) to students.
Calculator Grades 7-8 (Part 1 only): scientific calculators are acceptable EOC (entire test): graphing calculators are acceptable	Math	CBT – Students have access to the calculator tool when calculator use is permitted. Students may opt to use an acceptable handheld calculator instead of this tool when calculator use is permitted. PBT – Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.

Accommodations are provisions made in how a student accesses and demonstrates learning that do not substantially change the instructional level, the content, or the performance criteria. Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. Testing accommodations provide more equitable access during assessment but do not alter the validity of the assessment, score interpretation, reliability, or security of the assessment. For a student with disabilities, accommodations are intended to reduce or even eliminate the effects of the student's disability. For an English Language Learner or a Fluent English Proficient Year 1 or Year 2 student, accommodations are intended to allow the student the opportunity to demonstrate content knowledge even though the student may not be functioning at grade level in English.

Research indicates that more accommodations are not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. There should be a direct connection between a student’s disability, special education need, or language need and the accommodation(s) provided to the student during educational activities, including assessment. Test administrators are instructed to make accommodation decisions based on individual needs, and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation may be put in place for an AzMERIT test that is not already used regularly in the classroom.

Testing accommodations may not violate the construct of a test item. Testing accommodations may not provide verbal or other clues or suggestions that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, writing prompt, or answer option. The accommodations available to students while testing on AzMERIT are generally limited to those listed in *AzMERIT Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section. Arizona takes care to ensure allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student's individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact ADE for guidance.

Students with an injury, such as a broken hand or arm, that would make it difficult to participate in AzMERIT may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. There are no specific CBT tools to support these accommodations.

Exhibit 1.2.3 Accomodations for Students with an Injury

Accommodation	Description
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	Use of assistive technology for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.
Rest/Breaks	Student may take breaks during testing sessions to rest.

Students who are not proficient in English, as determined by the Arizona English Language Learner Assessment (AZELLA), may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. This includes English Language Learner (ELL) students and students withdrawn from English language services at parent request. Reclassified Fluent English Proficient (FEP) students are monitored for two school years. These FEP Year 1 and FEP Year 2 students also may use, as appropriate, any of the universal test administration conditions and any of the following accommodations.

The *upon student request* accommodations are required to be administered in a setting that does not disturb other students such as in a one-on-one or very small group setting.

Exhibit 1.2.4 summarizes accommodations that may be provided for ELL and FEP students.

Exhibit 1.2.4 Allowable Accommodations for ELL and FEP Students

Accommodation	Description of Use
Read Aloud Test Content	CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test. PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test upon student request. Reading aloud the content of the Reading portion of the ELA test is prohibited.
Rest/Breaks	Student may take breaks during testing sessions to rest.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	Exact oral translation, in the student's native language, of the scripted directions or the directions that students read on their own upon student request. Translations that paraphrase, simplify, or clarify directions are not permitted. Written translations are not permitted. Translation of test content is not permitted.
Translation Dictionary	Provide a word-for-word published, paper translation dictionary. Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned-off.

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 1.2.5, as designated in their IEP or 504 plan.

Exhibit 1.2.5 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	Use of assistive technology, including Braille writer, for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.
Braille Test Booklet	Provide a paper Braille test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.
Large Print Test Booklet	CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet. PBT – Provide a Large Print test booklet. A PBT Large Print test booklet requires Adult Transcription on a regular size paper test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.
Paper Test Booklet	CBT – Provide a regular size paper test booklet for a student at a school administering the CBT. If a paper test booklet is ordered as an accommodation for a student at a CBT school, the student must use the paper test booklet and may not participate in computer-based testing.

1.3 EVIDENCE BASED ON TEST CONTENT

Because the AzMERIT assessments are designed to measure student progress toward achievement of the ACCRS the validity of AzMERIT test score interpretations critically depend on the degree to which test content is aligned with expectations for student learning specified in the academic standards.[6]

Alignment of content standards is achieved through a rigorous test development process that proceeds from the content standards and refers back to those standards in a highly iterative test development process that includes the state department of education, test developers, and educator committees. Items used to develop the spring 2016 operational test forms were mainly drawn from the AIRCore pool of items developed to align with the Common Core State Standards. These items were also reviewed by Arizona content experts and educators prior to field-testing in spring 2015 and subsequent operational test administration in spring 2016. Only items that were found to align well with the ACCRS were used. To supplement the AIRCore pool of items, a few previously developed Arizona items that also aligned to the ACCRS were used. In subsequent years, test forms will be constructed using items developed directly with Arizona, meaning ADE and Arizona educator committees act as reviewers throughout the item development cycle.

In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards will be covered in each test administration.[7] Thus, the test blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the ACCRS is evaluated, alignment of test blueprints with the content standards is critical.

With the desired alignment of test blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet the blueprints. Developing test forms is difficult because test blueprints can be highly complex, specifying not only the range of items and points for each strand and standard, but also cross-cutting criteria such as distribution across item types, depth of knowledge, writing genre, and so on. And in addition to meeting complex blueprint requirements, test developers must work to meet psychometric goals so that alternate test forms measure equivalently across the range of ability.

Following a standard item review process, item reviews proceeded initially through a series of internal reviews before items were eligible for external review by Department staff and educator committees. Most of AIR's content staff members, who are responsible for conducting internal reviews, are former classroom teachers who

---

[6] Standard 12.4 – When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent.

[7] Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

hold degrees in education and/or their respective content areas. Each item passed through four internal review steps before it was eligible for external review. Those steps include

- Preliminary review, conducted by a group of AIR content area experts
- Content Review 1, performed by an AIR content specialist
- Edit, in which a copyeditor checks the item for correct grammar/usage
- Senior Content Review, by the lead content expert.

At every stage of the item review process, beginning with preliminary review, AIR's test developers analyze each item to ensure that

- The item is well-aligned with the intended content standard
- The item conforms to the item specifications for the target being assessed
- The item is based on a quality idea (i.e. it assesses something worthwhile in a reasonable way);
- The item is properly aligned to a depth of knowledge (DOK) level;
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter, and takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward
- Any accompanying graphic and stimulus materials are actually necessary to answer the question
- The item stem is clear, concise, and succinct, meaning it contains enough information to know what is being asked, is stated positively (and does not rely on negatives such as no, not, none, never, unless absolutely necessary), and it ends with a question
- For selected response items, the set of response options are succinct; parallel in structure, grammar, length, and content; sufficiently distinct from one another; all plausible, all non-keyed response options are unambiguously incorrect;
- There is no obvious or subtle cluing within the item
- The score points for constructed-response items are clearly defined
- For machine-scored constructed-response (MSCR) items, that item responses yield the intended score points based on the rubric, and
- For human scored constructed response items, the scoring rubric clearly explains what characterizes responses at each possible level of achievement.

In addition, rubric-scored items, both machine-scored and human-scored, are validated following field test administration. Machine-scored items go through a rubric validation process wherein samples of student responses are reviewed, along with resulting scores, to ensure that rubrics are enacted as intended. This process is described in Section 10.1.1. Human-scored items go through a rangefinding process prior to scoring where samples of item responses are used to create scorer training materials and ensure the scoring rubric is appropriate, as described in Section 10.1.2.

Based on their review of each item, the test developer may accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to the Department for their review. At this stage, items may be further revised based on any edits or changes requested by the Department, or rejected outright. Items passing through the Department review level then have to pass through a stakeholder review in which a committee of educators reviews each item's accuracy, alignment to the intended standard and DOK level, as well as item fairness and language sensitivity. Thus, all items considered for inclusion in the AzMERIT item pools were initially reviewed by an educator committee which checked to ensure that each item and associated stimulus materials was:

- aligned to the content standards
- appropriate for the grade level
- accurate
- presented online in a way that is clear and appropriate
- free from bias, sensitive issues, controversial language, stereotyping, and statements that reflect negatively on race, ethnicity, gender, culture, region, disability, or other social and economic conditions and characteristics.

Items were also passed through to a parent/community sensitivity review committee to ensure that test content did not violate community standards. Items successfully passing through both the educator and parent/community review process were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is therefore an important step in constructing valid and equivalent operational test forms.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning analyses).

Items flagged for review based on their statistical performance must pass a three-stage review to be included in the final item pool from which operational forms were created. In the first stage of this review, a team of psychometricians reviewed all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct and there are no other obvious problems with the items.

ADE content staff then re-evaluated flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, ADE determined that certain flagged items must be rejected, or deemed the item eligible for inclusion in operational test administrations.

1.4 EVIDENCE FOR INTERPRETATION OF PERFORMANCE STANDARDS

Alignment of test content to the Arizona College and Career Ready Standards ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the Arizona College and Career Ready Standards. However, the interpretation of the AzMERIT test scores rests fundamentally on how test scores relate to performance standards which define the extent to which students have achieved the expectations defined in the Arizona standards. AzMERIT test scores are reported with respect to four proficiency levels, demarcating the degree to which Arizona students have achieved the learning expectations defined by the Arizona College and Career Ready Standards. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade level expectations for achievement of the Arizona standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standard for the AzMERIT assessments are therefore central to the validity of test score interpretations.[8]

Following the first operational administration of the AzMERIT in spring 2015, a standard setting workshop was conducted to recommend to ADE a set of performance standards for reporting student achievement of the Arizona College and Career Ready Standards. Arizona educators, serving as standard setting panelists, followed a standardized and rigorous procedure to recommend performance level cut scores. The workshops employed the Bookmark procedure, a widely used method in which standard setting panelists used their expert knowledge of the Arizona College and Career Ready Standards and student achievement to map the performance level descriptors adopted by Arizona onto an ordered item book comprising the spring 2015 operational test form and augmented with items administered in the embedded field test slots to minimize information gaps in the operational test form.[9]

Panelists were also provided with contextual information to help inform their primarily content driven cut score recommendations. For each assessment, panelists were provided the approximate location of performance standards for other important assessment systems. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3-8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3-8 and 11 assessments in ELA and math to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

In addition, panelists were provided with feedback about the vertical articulation of their recommended performance standards so that they could view the relationship between the locations of recommended cut scores for each grade level assessment to the cut score recommendations at the other grade levels. This approach allowed panelists to view their cut score recommendations as a coherent system of performance standards, and further reinforces the interpretation of test scores as indicating not only achievement of current grade level standards, but also preparedness to benefit from instruction in the subsequent grade level.

---

[8] Standard 4.22 – Test developers should specify the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.

[9] Standard 1.18 – When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.

Following recommendation of final performance standards, the recommended cut scores were presented to the Arizona State Board of Education for review and adoption. The Board adopted the recommended performance standards in August 2015.

Based on the adopted performance standards, Exhibit 1.4.1 shows the estimated percentage of students meeting the AzMERIT proficient standard for each assessment in spring 2015. Exhibit 1.4.1 also shows the approximate percentage of Arizona students that would be expected to meet the ACT college ready standard, and the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8. Exhibit 1.4.1 also presents the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring 2014 field test administration. As Exhibit 1.4.1 indicates, the performance standards recommended AzMERIT assessments are quite consistent with relevant ACT college ready, and the NAEP and Smarter Balanced proficient, benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Exhibit 1.4.1 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

Test	Percent of Students Meeting Standard			
	AzMERIT Proficient	Arizona ACT College Ready	Arizona NAEP Proficient	Projected SBAC
		ELA		
3	41%			38%
4	38%		28%	41%
5	30%			44%
6	34%			41%
7	33%			38%
8	32%		28%	41%
9	27%			
10	30%			
11	25%	34%		41%
		Math		
3	42%			39%
4	42%		42%	38%
5	40%			33%
6	32%			33%
7	31%			33%
8	33%		32%	32%
Algebra I	32%			
Geometry	30%			
Algebra II	29%	36%		33%

Although AIR previously identified ACT college ready cut scores on the AzMERIT ELA and math scales, that study involved an indirect linkage. In that study, student performance on the grade 10 AIMS was used to predict subsequent student performance on the ACT tests, and then a linking study between the AIMS and AzMERIT allowed for the identification of the ACT cut scores on the AIMS scale to be represented onto the AzMERIT scale.

To examine directly the relationships between the AzMERIT and ACT assessments, ADE obtained the ACT test scores for Arizona students graduating high school in spring 2016. Although AzMERIT is offered as a series of end of course tests in high school, most students take the Algebra II assessment at grade 11, so the focus of this investigation will be on the grade 11 ELA and Algebra II AzMERIT assessments administered in spring 2015.

Because a selected sample of fewer than half of Arizona students takes the ACT, a two-step approach is typically adopted to impute missing data in the analysis of the relationship between the AzMERIT and ACT test scores. However, previous investigations with other state assessments, including the Arizona AIMS, demonstrated that imputing or deleting the missing records did not impact the linkage identified between state assessments and the ACT test. For this study we instead divided the complete sample of merged records into model building and cross-validation samples of equal size. The cross-validation sample allows for better estimation model fit. Because the model is built using a sample independent from that used to evaluate model fit, estimates of model fit exclude sample dependent idiosyncrasies that would be reflected as model overfit in the model development sample.

Exhibit 1.4.2 shows the location of the ACT college ready cut scores for math and reading on the AzMERIT scale. The first column shows the location as identified via indirect linkage through AIMS and that was provided as benchmark information to AzMERIT standard setting panelists. The second column shows the location of the ACT college ready cut scores as identified via direct linkage between ACT and AzMERIT described here. The third column shows the location of the AzMERIT meets performance standard on the Algebra II and Grade 11 ELA assessments. As indicated in the table, the location of the ACT college ready cut scores on the AzMERIT scale were reasonably consistent across methods, especially for ELA. Importantly, the results affirm that the location of adopted AzMERIT performance standards are consistent with the ACT college ready criteria.

Exhibit 1.4.2. Location of the ACT College Ready Cut Scores on the AzMERIT Scales.

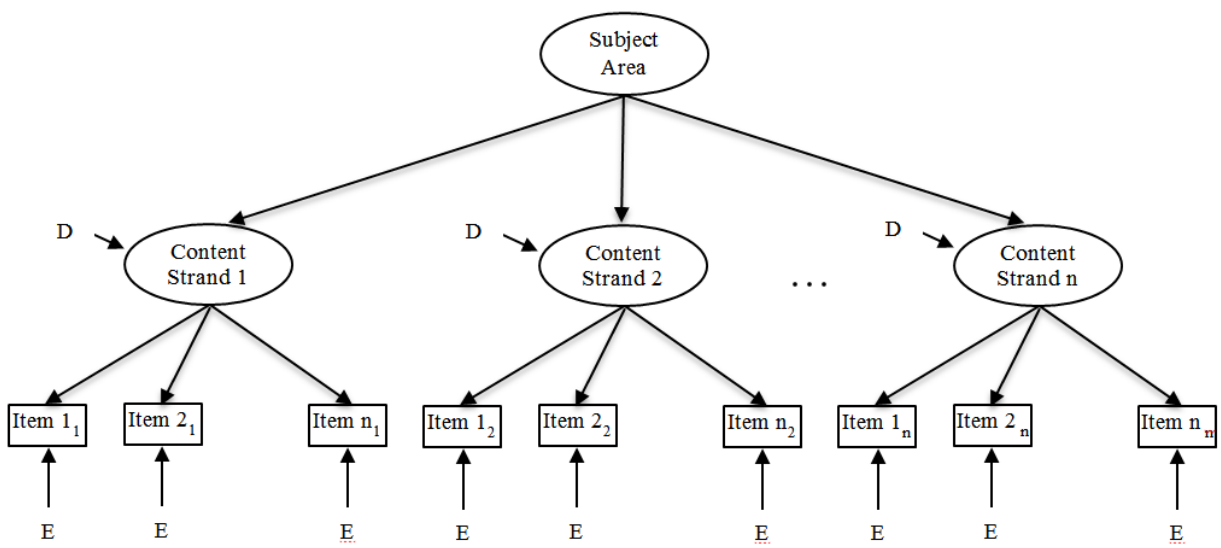
	Location of ACT College Ready Cut on AzMERIT Scale		**AzMERIT Meets Performance Standard**
	Via Indirect Linkage through AIMS	**Via Direct Linkage with AzMERIT**	
Algebra II	3704	3727	3711
Grade 11 ELA	2579	2585	2585

The equipercentile equating method was used to verify the linkage between ACT and AzMERIT test scores. The AzMERIT scale score associated with the ACT college ready cut scores in reading was 2586 on the AzMERIT ELA scale. The location of the ACT college ready cut score in math was 3727 for the AzMERIT math scale. Results from the equipercentile approach were thus consistent with the cut scores identified using regression models.

1.5 EVIDENCE BASED ON INTERNAL STRUCTURE

Arizona's AZMERIT assessment represents a structural model of student achievement in grade level and course specific content areas. Within each subject area (e.g., ELA), items are designed to measure a single content strand (e.g., Reading Information, Reading Literature, Language, Writing). Content strands within each subject area are, in turn, indicators of achievement in the subject area. The form of the second-order confirmatory factor analyses is illustrated in Exhibit 1.5.1. As the exhibit illustrates, each item is an indicator of an academic content strand. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each academic content strand serves as an indicator of achievement in a subject area. As at the item level, the content strands include an error term indicating that the content strands are not pure indicators of overall achievement in the subject area. The paths from the content strands to the items represent the first-order factor loadings, the degree to which items are correlated with the underlying academic content strand construct. Similarly, the paths from subject area achievement to the content strands represent the second-order factor loading, indicating the degree to which academic content strand constructs are correlated with the underlying construct of subject area achievement.

Exhibit 1.5.1 Second-Order Structural Model for AzMERIT Assessments



Following the first operational test administration in spring 2015, confirmatory factor analysis was used to evaluate the fit of this structural model to student response data.[10] For each of test forms administered in spring 2015, we examined the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a χ^2 statistic, with good model fit indicated by a non-significant χ^2 statistic. The χ^2 statistic is sensitive to sample size, however, so even well-fitting models will demonstrate highly significant χ^2 statistics given a very large number of students. Therefore, fit indices, such as the Comparative Fit Index (CFI; Bentler, 1990), the Tucker-Lewis Index (Tucker & Lewis, 1973), and the Root Mean Square of Approximation (RMSEA) were also used to evaluate model fit.

The AzMERIT assessments also claim to measure subject area achievement using test items that probe student knowledge and skills across multiple depth of knowledge levels. As with the content standards, the classification of items by depth of knowledge also represents a structural model that can be evaluated using confirmatory factor analysis.[11] In this case, each item is an indicator of a depth of knowledge level first-order factor, and each depth of

[10] Standard 1.13 – If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.

[11] Standard 1.12 – If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.

knowledge level is in turn an indicator of subject area achievement. Thus, confirmatory factor analysis was used to evaluate the fit of this depth of knowledge structural model to student response data from the spring 2015 AzMERIT test administrations.

Exhibit 1.5.2 Guidelines for Evaluating Goodness of Fit

Goodness-of-Fit Index	Indication of Good Fit
CFI	≥ .95
TLI	≥ .95
RMSEA	≤ .05

In addition to testing the fit of the hypothesized AzMERIT second-order confirmatory factor analysis model, we examined the degree to which the second-order model improved fit over the more general one-factor model of academic achievement in each subject area. Because the second-order model was nested within the one-factor, general achievement model, a simple likelihood ratio test was used to determine whether the added information provided by the structure of the ACCRS frameworks improved model fit over a general achievement model. Results indicating improved model fit for the second-order factor model provide support for the interpretation of content standard performance above that provided by the overall subject area score.[12]

1.5.1 ELA CONTENT MODEL

We began by evaluating the fit of the first-order, general achievement model in which all items are indicators of a common subject area factor. This model importantly evaluates the assumption of unidimensionality of the subject area assessments, and provides a baseline for evaluating the improvement of fit for the more differentiated second-order model. The goodness-of-fit statistics for the first-order, general achievement models in ELA are shown in Exhibit 1.5.1.1. All of the statistics indicate the general achievement factor model fit the data well. This pattern was true across all grades. The CFI and TLI values were all greater than 0.9 and generally equal to or greater than 0.95, and the RMSEA values were all below .05, indicating good fit for the base model.

Exhibit 1.5.1.1 Goodness-of-Fit for the AzMERIT ELA First-Order Model

First-Order Models			
Grade	CFI	TLI	RMSEA
3	0.934	0.931	0.047
4	0.949	0.946	0.033
5	0.966	0.964	0.039
6	0.955	0.953	0.043
7	0.974	0.972	0.037
8	0.964	0.963	0.048
9	0.924	0.921	0.039
10	0.948	0.945	0.042
11	0.928	0.925	0.034

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.5.1.2. All of the statistics indicate the second-order models posited by the AzMERIT assessments fit the data well. This pattern was true across all grades. As with the general factor model, the CFI and TLI values for the second-order

[12] Standard 1.14 – When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

models were all equal to or greater than .95, with RMSEA values well below the .05 threshold used to indicate good fit.

Exhibit 1.5.1.2 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Second-Order Models			
Grade	CFI	TLI	RMSEA
3	0.958	0.956	0.038
4	0.970	0.969	0.025
5	0.980	0.979	0.030
6	0.973	0.972	0.033
7	0.983	0.982	0.029
8	0.980	0.979	0.036
9	0.962	0.960	0.028
10	0.972	0.970	0.031
11	0.949	0.947	0.029

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 1.5.1.3. We note that model fit for first-order model of general achievement are also very high and provide evidence for the unidimensionality of the subject area assessments. The purpose of these analyses is to determine whether the posited second-order reporting model adds information beyond that provided by the first-order model. The chi-square difference test shows that across grade levels, the strand-based second-order model showed significantly better fit than the general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 1.5.1.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	13560.7	3	$p < .001$
4	8460.9	3	$p < .001$
5	10944.7	3	$p < .001$
6	12019.8	3	$p < .001$
7	8848.6	3	$p < .001$
8	15590.1	3	$p < .001$
9	8896.6	3	$p < .001$
10	9084.7	3	$p < .001$
11	4412.8	3	$p < .001$

1.5.2 ELA DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the hypothesized AzMERIT second-order models in ELA are shown in Exhibit 1.5.2.1. Across all grades, results indicate the second-order models posited by the AzMERIT assessments fit the data well. The CFI and TLI values were all .97 to .99, RMSEA values are all .03 or lower.

Exhibit 1.5.2.1 Goodness-of-Fit for the AzMERIT ELA Second-Order Model

Second-Order Models			
Grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.99	0.99	0.02
6	0.98	0.98	0.03

7	0.99	0.99	0.02
8	0.99	0.99	0.02
9	0.98	0.98	0.02
10	0.98	0.97	0.02
11	0.98	0.98	0.02

The results of the comparison between the hypothesized AzMERIT model and the more general achievement model are presented in Exhibit 1.5.2.2. The chi-square difference test shows that across grade levels, the DOK-based second-order model showed significantly better fit than the general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels.

Exhibit 1.5.2.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	21402.6	4	$p < .001$
4	12053.6	4	$p < .001$
5	17102.9	4	$p < .001$
6	18192.1	4	$p < .001$
7	16351.4	4	$p < .001$
8	25454.7	4	$p < .001$
9	14989.3	4	$p < .001$
10	14920.9	4	$p < .001$
11	8075.1	4	$p < .001$

1.5.3 MATH CONTENT MODEL

As with ELA, structural analyses of the math assessments began with an evaluation of fit for the first-order, general achievement model in which all items are indicators of a common math subject area factor. This model provides for an evaluation of the unidimensionality assumption of the subject area assessments, and provides a baseline for evaluating the improvement of fit for the more differentiated second-order model. The goodness-of-fit statistics for the general achievement models in math are shown in Exhibit 1.5.3.1. All of the statistics indicate the general achievement factor model fit the data well. This pattern was true across all grades. The CFI and TLI values were all equal to or greater than .95, and the RMSEA values are all below .05, indicating good fit for the base model.

Exhibit 1.5.3.1 Goodness-of-Fit for the AzMERIT Math First-Order Model

First-Order Models			
grade	CFI	TLI	RMSEA
3	0.975	0.973	0.027
4	0.976	0.975	0.024
5	0.976	0.975	0.026
6	0.975	0.973	0.023
7	0.982	0.981	0.021
8	0.969	0.967	0.026
Algebra I	0.976	0.975	0.023
Algebra II	0.973	0.971	0.021
Geometry	0.986	0.985	0.018

The goodness-of-fit statistics for the strand-based second-order models are shown in Exhibit 1.5.3.2. The models show very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their.05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 1.5.3.2 Goodness-of-Fit for the AzMERIT Math Second-Order Model

Second-Order Models			
grade	CFI	TLI	RMSEA
3	0.979	0.978	0.024
4	0.978	0.977	0.024
5	0.978	0.977	0.025
6	0.976	0.975	0.023
7	0.983	0.982	0.021
8	0.970	0.969	0.026
Algebra I	0.978	0.977	0.022
Algebra II	0.974	0.972	0.020

Geometry	0.987	0.986	0.017

The results of the comparison between the second-order, strand-based model and the first-order, general achievement model are presented in Exhibit 1.5.3.3. Again, model fit for the general achievement first-order model is very high, providing evidence for the unidimensionality of the subject area assessments. The purpose of these analyses is to determine whether knowledge of the depth of knowledge level of items provides information beyond that provided by the more general model. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 1.5.3.3 Difference in Fit Between Content Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	3225.0	3	$p < .001$
4	1326.3	3	$p < .001$
5	1427.0	3	$p < .001$
6	1036.2	4	$p < .001$
7	559.8	4	$p < .001$
8	1039.3	4	$p < .001$
Algebra I	750.9	3	$p < .001$
Algebra II	246.5	3	$p < .001$
Geometry	269.7	4	$p < .001$

1.5.4 MATH DEPTH OF KNOWLEDGE

The goodness-of-fit statistics for the DOK-based second-order models are shown in Exhibit 1.5.4.1. The models demonstrate very good fit, with all CFI and TLI fit indices above .97, and with RMSEA estimates are well below their.05 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 1.5.4.1 Goodness-of-Fit for the AzMERIT Math Second-Order Model

Second-Order Models			
grade	CFI	TLI	RMSEA
3	0.98	0.98	0.03
4	0.98	0.98	0.02
5	0.98	0.98	0.03
6	0.98	0.97	0.02
7	0.98	0.98	0.02
8	0.97	0.97	0.03
Algebra I	0.98	0.98	0.02
Algebra II	0.99	0.99	0.02
Geometry	0.97	0.97	0.02

The results of the comparison between the second-order, DOK-based model and the first-order, general achievement model are presented in Exhibit 1.5.4.2. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels.

Exhibit 1.5.4.2 Difference in Fit Between DOK Derived Second-Order and General Achievement First-Order Model

grade	χ^2	*df*	*p* value
3	331.4	3	$p < .001$
4	309.5	3	$p < .001$
5	14.9	3	$p < .001$
6	14.5	3	$p < .001$
7	236.6	3	$p < .001$
8	79.2	3	$p < .001$
Algebra I	20.1	3	$p < .001$
Algebra II	26.4	3	$p < .001$
Geometry	20.9	3	$p < .001$

1.6 EVIDENCE FOR RELATIONSHIPS WITH CONCEPTUALLY RELATED CONSTRUCTS

Validity evidence based on relations to other variables can address a variety of questions. At its core, this type of validity addresses the relationship between test scores and variables of interest that are derived outside the testing system. One type of validity evidence based on relations to other variables is evidence for convergent and discriminant validity. Evidence for convergent validity is based on the degree to which test scores correlate with other measures of the same attribute—scores from two tests measuring the same attribute should be correlated. Conversely, evidence for discriminant validity is obtained when test scores are not correlated with measures of construct irrelevant attributes.[13]

Observed correlations between alternate indicators of student achievement of course objectives, such as locally administered assessments of student achievement and AzMERIT, should be limited only by the unreliability of the measures. When both assessments measure student achievement in common subject areas, as with for example, locally administered and statewide assessments of math achievement, we expect test scores between the common subject area assessments to be substantially correlated. In addition, we expect that the magnitude of observed correlations between test scores in different subject areas will be lower than correlations between test scores in a common subject area. Because the content domains assessed in ELA and math tests are quite different, AzMERIT ELA test scores should correlate less well with locally administered assessments of math than ELA. It is important to note, however, that test scores across subject areas and test systems are nevertheless expected to be highly correlated. This is because even though subject area test scores measure different academic content domains, student achievement across subject areas is influenced by factors both internal (e.g., general intelligence) and external (e.g., socioeconomic status) to the student that contribute to student achievement across all academic subject areas so that student test scores across subject areas tend to be highly intercorrelated. So while we certainly do expect correlations between test scores across subject areas to be lower than correlations between test scores within a subject area, we nevertheless expect test scores across subject areas to be quite high.

Exhibit 1.6.1 shows the correlations between student test scores on the spring 2015 statewide AzMERIT assessment with corresponding test scores on a district-wide administration of the Northwest Evaluation Association (NWEA)

[13] Standard 1.16 – When validity evidence includes empirical analyses of responses to test items together with data on other variables, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent.

assessment. Sample sizes range from more than 1,400 students taking the grade 3 assessments, to nearly 1,100 students taking the middle school assessments, so the observed correlations are expected to be stable. Convergent correlations are quite high, ranging from 0.82 to 0.84 between AzMERIT ELA (assessing reading, writing, and listening) and NWEA reading. Correlations between AzMERIT and NWEA math scores are even higher, ranging from 0.85 to 0.89.

Exhibit 1.6.1 Correlations between AzMERIT and Locally Administered NWEA Test Scores

	AzMERIT ELA/NWEA Reading		AzMERIT Math/NWEA Math	
Grade	**Sample Size**	**Correlation**	**Sample Size**	**Correlation**
3	1426	0.82	1429	0.86
4	1214	0.84	1214	0.88
5	1303	0.84	1303	0.88
6	1119	0.82	1115	0.85
7	1081	0.82	1082	0.89
8	1090	0.82	1091	0.89

Exhibit 1.6.2 shows the discriminant correlations between AzMERIT and the locally administered NWEA assessment. As expected, correlations across subject area assessments remain quite high, indicating considerable consistency in student achievement across subject area assessments. Nevertheless, correlations across subject area assessments are systematically lower than within subject correlations, indicating that the subject area assessments are measuring domain specific knowledge and skills in addition to common factors underlying student achievement.

Exhibit 1.6.2 Correlations between AzMERIT and Locally Administered NWEA Test Scores

	AzMERIT ELA/NWEA Math		AzMERIT Math/NWEA Reading	
Grade	**Sample Size**	**Correlation**	**Sample Size**	**Correlation**
3	1426	0.72	1428	0.70
4	1211	0.76	1217	0.72
5	1303	0.75	1303	0.72
6	1117	0.73	1117	0.71
7	1081	0.77	1080	0.74
8	1088	0.75	1093	0.71

Convergent correlations between AzMERIT and locally administered assessments were also reported by Estrada and colleagues (Estrada, Burnham, Feld, Bergan, and Bergan, 2015). These researchers reported the mean correlations between a variety of local assessments and AzMERIT test scores for ELA and math assessments in grades 3-8. Mean correlations between AzMERIT and various local assessments of ELA ranged from .77 to .79 across the grade levels investigated. Mean correlations between AzMERIT and local assessments of mathematics ranged from .71 to .75 across grade levels 3 through 8. These results likewise show good convergence between AzMERIT and other locally administered assessments purporting to measure the same constructs.

1.7 MEASUREMENT INVARIANCE ACROSS SUBGROUPS

Measurement invariance occurs when the likelihood of correct responding conforms to the measurement model and is independent of group membership and the parameters of a measurement model are statistically equivalent across groups.[14] The parameters of interest in measurement invariance testing are the factor loadings and intercepts/thresholds. Invariance in residual variances or scale factors can also be tested, but there is consensus that it is not necessary to demonstrate invariance across groups on these parameters. In general, measurement invariance testing can be conducted using a series of multiple-group confirmatory factor analysis (CFA) models, which impose identical parameters across groups. The measurement model parameters, including factor patterns (configural invariance), factor loadings (metric or weak invariance), latent intercepts/thresholds (scalar or strong invariance), and unique or residual factor variances (strict invariance), are tested across groups in that sequential order. When factor loadings and intercepts/thresholds are invariant across groups, scores on latent variables can be validly compared across the groups.

Appendix B shows the results of measurement invariance testing by subgroups for ELA and Math. Items comprising the spring 2016 operational test administration were used to investigate measurement invariance across subgroups. The full set of tables associated with these analyses is provided for each of the grade and subject area assessments. The series "a" tables (e.g., 1a, 2a, etc.) present the global model fit indices for the measurement invariance tests for each assessment. Following the sequence of tests of measurement invariance (Millsap & Cham, 2012), we tested configural, metric, and scalar invariance models using $\chi 2$ difference test (at $\alpha \leq 0.05$) and the examination of significant differences of the Root Mean Square of Approximation (RMSEA, change in RMSEA ≤ 0.015; Chen, 2007) between the two nested invariance models. Measurement invariance was investigated across the following subgroups: gender (Model A), ethnicity including African American vs. White (Model B-1), Hispanics vs. White (Model B-2), Asian vs. White (Model B-3), American Indian vs. White (Model B-4), and Multi-Ethnics vs. White (Model B-5), special education program status (SPED; Model C), economic disadvantage status (Low Income; Model D), limited English proficiency status (LEP; Model E), and accommodated test forms (Accommodation, Model F). Invariance tests of subgroups were investigated separately for each grade and subject area test. Since in each ELA assessment, students were randomly assigned to one of six writing prompts for administration, the missing responses on the writing items resulted in unsuccessful model convergence. Thus, to achieve model convergence, we included the students who took a common writing prompt between online and paper in each ELA assessment.

The null hypothesis of the $\chi 2$ difference test is that the more restricted invariance model (e.g., metric) fits the data equally as well as the less restricted invariance model (e.g., configural). Given that the sensitivity of the $\chi 2$ difference tests to sample size, we additionally examined significant differences on this test with an examination of the RMSEA. A small change in the RMSEA between the more restricted and less restricted invariance models supports retention of the more restricted invariance model (Chen, 2007).

The series "b" tables (e.g., 1b, 2b, etc.) show the model fit indices of scalar invariance models assuming same factor pattern + identical factor loadings + identical latent intercept/threshold across subgroups. Global model fit indices included the Comparative Fit Index (CFI; Bentler, 1990) and Root Mean Square of Approximation (RMSEA). CFI values ≥ 0.90 and RMSEA values ≤ 0.08 were used to evaluate acceptable model fit. The model fit indices of the scalar invariance models for all tests suggested acceptable fit to the data. For ELA, CFI ranged from 0.870 to 0.990 and RMSEA ranged from 0.012 to 0.044. For Math, CFI values ranged from 0.905 to 0.990 and RMSEA ranged from 0.010 to 0.058.

Although the $\chi 2$ difference test should ideally be nonsignificant, all $\chi 2$ difference tests were significant at α =.05 due to large sample sizes except Model B-4, where the $\chi 2$ difference tests for most grades was nonsignificant or marginally significant at α =.05. In spite of significant $\chi 2$ difference tests for most models, we found that changes of the RMSEA between the two nested invariance models were very small (ranging from 0.000 to 0.002 for both ELA and MATH). Based on the similar magnitudes of the RMSEA (i.e., no material change across all tested models;

[14] Standard 3.15 – Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

Cheung & Rensvold, 2002) and the acceptable fit indices of the scalar invariance model to the data, ELA and MATH test scores have the same measurement structure across gender, ethnicity (African American vs. White, Hispanics vs. White, Asian vs. White, American Indian vs. White, and Multi-Ethnics vs. White), special education program status, economic disadvantage status, limited English proficiency status, and accommodation test forms.

1.8 DIFFERENTIAL MODE EFFECTS ACROSS SUBGROUPS

To explore the possibility that mode of test administration may exert differential effects across subgroups, we began by identifying matched samples of students participating online and on paper. For students administered assessments on paper, observed test scores were regressed on prior achievement and demographic variables to obtain regression weights. The resulting prediction equation was then applied to all students to yield predicted paper test scores. The predicted paper scores were used to identify matched samples of online and paper test takers.

To identify possible differential effects of mode across subgroups, we used the observed test score as dependent variable and then covaried the predicted test score to isolate the effects of mode. We then entered the dummy-coded demographic variables including gender, English language learner status, special education status (SPED), free reduced lunch status (FRL), migrant status, and six ethnicity subgroups as predictors. Significant interactions between mode of test administration and the demographic subgroup comparisons indicate differential mode effects between the specified demographic subgroups.

Given the very large sample sizes, many effects achieve conventional levels of statistical significance, but are nevertheless quite small. Thus, Exhibit 1.8.1 shows the parameter estimates for the differential mode effects by subgroup interaction only for effects where $p < .0001$.

Results indicated that mode effects were more pronounced for special education students relative to general education population. Especially for the high school EOC tests, AzMERIT tests were more difficult for special education students when administered on paper than online.

Mode effects were more pronounced for low income students with respect to the math assessments. Math tests were generally more difficult for low income students when administered online than on paper.

Mode effects were also more pronounced for LEP students than for the general education population in math but not ELA. However, the direction of this effect was not consistent across grades. Online math tests were more difficult than paper for LEP students in the lower grades, while paper math tests were more difficult than online tests for LEP students in the higher grades.

Exhibit 1.8.1 Parameter Estimates for Differential Mode Effects by Subgroups Interactions

Test	Gender	White	Black	Asian	Hawaii/Pacific	Hispanic/Latino	American Indian	Special Education	Limited English Proficiency	Free/Reduced Lunch	Migrant
ELA											
G3E	0.49									0.27	
G4E											

G5E											
G6E								-0.61			
G7E								0.5			
G8E					1.66	-0.34					
G9E	0.45							-0.74			
G10E								-1.23		-0.41	
G11E	-0.33					0.36		-0.58			
MATH											
G3M								0.57			
G4M									0.52	-	-4.46
G5M							-0.89			0.34	
G6M		1.15	0.96				0.69		0.6	-0.31	
G7M	-0.26									0.25	-2.87
G8M		0.89					0.86		-0.58		
Algebra I						0.73		-0.8	-0.95	0.5	
Geometry						-0.44		-1.32		1.11	
Algebra II							-1.07	-0.75		0.63	

1.9 EVIDENCE FOR STUDENT GROWTH – OVERALL AND BY SUBGROUPS

The AzMERIT assessments report student test scores on a vertical scale, allowing families and teachers to make inferences about student growth across school years. The validity of test score interpretations about student growth over time depend strongly on the vertical linking design used to develop the vertical scale. But even when test score interpretations are appropriate to the scaling design, it is important to examine whether student gains may be interpreted consistently across subgroups or whether differential gain rates across subgroups limit the inferences that can be made about test score gains over time.[15] To address this issue, we examined rates of student growth across student gender, race/ethnicity, students with disabilities (SPED), English language learners (LEP), and low income status (Low Income).

Exhibit 1.9.1 shows the mean test scores on the spring 2015 and spring 2016 administrations of AzMERIT for students participating in both test administrations, as well as the correlation between test scores across the two assessment occasions. Correlations between test scores are quite high and indicate substantial consistency in rank ordering of student achievement between the two test administrations. The correlation between student achievement in grade 8 math and Algebra I is attenuated somewhat, and further that the distribution of student ability is somewhat less variable for this cohort, especially with respect to the spring 2016 Algebra I performance. We note that in spring 2015, grade 8 students enrolled in Algebra I were required to participate in both

[15] Standard 3.15 – Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.
Standard 3.17 – When aggregate scores are publicly reported for relevant subgroups— for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults— test users are responsible for providing evidence of comparability and for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups.

assessments, but in spring 2016, those high achieving students would likely have participated in the Geometry assessment and would not have been included in these analyses. The resulting restriction of range could be responsible for the attenuated correlation.

The exhibit also shows that rate of achievement gain is somewhat higher for math than ELA, and that while gain rates decelerate across the school years, the rate of gains diminishes more rapidly for ELA than math over time. For math, large gains, typically ¾ standard deviation or more (e.g., average gain of 33 points in grade 3 math is 80% of the 40 point standard deviation of student test scores), are observed through the middle school grades, dropping to about 1/3 standard deviation between administrations of the high school end-of-course assessments. For ELA, while elementary school gains are strong, by middle school, annual gains are between 1/3 to 1/2 standard deviation, and by high school drop to about ¼ standard deviation, with no growth observed between grade 10 and 11.

Exhibit 1.9.1 Test Score Stability and Performance Gains Overall

Assessment Subject_2015 _2016	N	Spring 2015		Spring 2016		Change from 2015 to 2016		Percent Scoring Lower		Correlation
		Mean	Std Dev	Mean	Std Dev	Mean	IRT based Standard Error	Expected	Observed	
ELA										
ELA_G3E_G4E	80245	2501	29.99	2518	33.59	17	14.43	0.23	0.18	0.82
ELA_G4E_G5E	79662	2514	28.61	2537	34.08	23	14.32	0.16	0.11	0.82
ELA_G5E_G6E	78965	2529	27.9	2541	33.59	13	14.06	0.29	0.24	0.83
ELA_G6E_G7E	78273	2541	29.81	2553	30.88	11	13.90	0.30	0.25	0.84
ELA_G7E_G8E	76782	2548	28.85	2556	32.21	8	13.74	0.35	0.32	0.84
ELA_G8E_G9E	69223	2561	28.67	2566	30.81	6	13.72	0.40	0.36	0.82
ELA_G9E_G10E	61972	2562	26.31	2567	28.68	5	13.54	0.41	0.38	0.80
ELA_G10E_G11E	53924	2571	26.48	2569	31.19	-2	13.31	0.53	0.54	0.80
MATH										
Math_G3M_G4M	80875	3522	39.27	3553	40.21	31	16.38	0.14	0.10	0.81
Math_G4M_G5M	80277	3553	37.49	3589	41.29	36	16.11	0.11	0.07	0.81
Math_G5M_G6M	79107	3585	38	3616	42.37	31	16.05	0.14	0.09	0.83
Math_G6M_G7M	76152	3614	34.1	3633	35.48	20	15.22	0.21	0.16	0.83
Math_G7M_G8M	64887	3625	33.3	3652	35.99	27	15.48	0.14	0.09	0.82
Math_G8M_Algebra I	50295	3651	29.23	3665	28.79	13	14.98	0.30	0.25	0.74
Algebra I_Geometry	54203	3674	33.74	3685	35.71	11	15.72	0.33	0.29	0.80
Geomery_Algebra II	43737	3687	33.22	3697	33.17	11	16.37	0.34	0.31	0.77

To evaluate differential growth across demographic subgroups, a series of regression analyses were conducted to predict 2016 test scores from 2015 test scores, controlling for demographic subgroup membership. To compare ethnic subgroup performance, we created six dummy variables contrasting white students with each of other ethnic groups (e.g., white/Hispanic, white/African American). Gender was coded 1 for female. SPED, LEP, and Low

Income students were coded as 1 to contrast with students who were not identified with those needs who were coded as 0.

Appendix C shows the regression model parameter estimates for the ELA and math assessments. The 2015 test scores were centered on the reference group mean so that the intercept values at the top of the table represent the mean performance of white males on the 2015 assessment, with group parameters reflecting differences from the reference group on the spring 2016 assessment. Results indicate that females generally performed better than males for both ELA and math across grades. With respect to ethnicity, Asian students generally performed better than white students in both ELA and math. For all other ethnic group comparisons, the focal groups generally performed less well than whites. Special education students, limited English proficient students, and low income students all performed less well than the general education population in both ELA and math.

The slope represents the association between 2015 and 2016 test scores, controlling for demographic subgroups. The overall positive slope parameter indicates reference group gains in test scores between 2015 and 2016. The group specific slope parameters indicate differential gains between contrasted groups.

Although many individual effects attained conventional levels of statistical significance due to large sample sizes, we focus here only on highly significant effects that are associated with more practically significant effect sizes and that may point to trends across grade level and/or subject area assessments.

While females tended to score higher across assessments, differential gain rates by gender were small and inconsistent.

With respect to ethnicity, differential gain rates were small and inconsistent in the elementary and middle school grade assessments. Asian students did, however, show higher gain rates than whites only in the high school years. And African American and Hispanic students showed lower gain rates than whites in the high school math assessments.

Special education students generally showed lower rates of gain than general education students, although pattern was reversed for elementary school math, with special education students showing greater rates of gain from grade 3 to 4, and from grade 4 to 5.

Limited English proficient students showed lower rates of gain in both ELA and math, but this effect seems to moderate in the high school grades where differential gain rates were much less pronounced.

Differential gain rates for low income students were observed for ELA, but were inconsistent across grades. Low income students generally showed lower gain rates in math, but this effect was more apparent at the middle and high school grades.

1.10 DAY, WEEK, AND TIME OF DAY EFFECTS ON PERFORMANCE

Administration of Arizona’s new AzMERIT online tests is untimed and schools may flexibly schedule students to take the tests in computer labs throughout the testing window. Thus, students taking the same grade level or end of course (EOC) test are not required to test on the same day. This is a marked departure from Arizona’s administration of AIMS in which students statewide were administered tests on the same date and nearly the same time, ensuring that any effects of test administration time or day were held constant.

Because the days and times on which tests can be administered is variable, the possibility arises that performance factors associated with time of day or day of week may influence student test scores.

A series of regression models were developed to predict student performance using the day of the week and time of the day variables, as well as the duration of the test administration from test start to test end. The dependent variable for these analyses was the spring 2016 AzMERIT scale score. To control for student achievement, we first covaried previous achievement using spring 2015 AzMERIT test scores. Because of the need to covary previous achievement, the analyses were limited to students participating in the grade 4 to 8 and high school EOC assessments in mathematics and ELA tests, and for whom 2015 test scores were available. The day of the week was coded as 1 to 5 (1 for Monday, 2 for Tuesday, and so on). For the regression analyses, the time of day and the duration were continuous variables using the actual time. Time of day effects were further evaluated using paired comparisons between early morning, late morning, early afternoon, and late afternoon.

Exhibit 1.10.1 shows the standardized regression coefficient estimates of the time effect on student's performance only for effects where $p < .05$. Results indicate generally that starting tests earlier in the week resulted in higher test performance. Tests started on Friday were consistently associated with impaired performance. There were some exceptions to this. For example, students beginning the grade 7 ELA tests on Monday scored lower than students beginning on any other day than Friday. But generally the pattern was pronounced.

Conversely, assessments which were completed earlier in the week were associated with lower test scores. Tests ending on any other day than Monday were associated with higher test scores. And this effect was generally true for tests ending on Tuesday. That said, students appeared to perform better on tests ending Wednesday or Thursday than on Friday, although there were exceptions to this as well (e.g., grade 9 and grade 10 ELA where Friday end dates were associated with greater performance).

Time of day effects were less consistent. For ELA, morning start times were associated with greater performance than afternoon start times for high school students. For middle school students later morning start times were associated with poorer performance than early morning or late afternoon. And at grade 5, ELA tests with morning start times were associated with lower performance than tests with afternoon start times.

For math tests, later start times were generally associated with better performance. An exception to this pattern was observed for Algebra I, where students beginning testing late morning performed better than students starting at any other time.

Tests ending early in the afternoon were generally associated with higher performance than tests ending earlier in the day, although grade 6 ELA proved an exception with tests ending early morning associated with the highest scores.

In addition, longer test administrations were associated with higher performance.

Exhibit 1.10.1. Standardized Regression Coefficients of Time Effect on Student's Performance

Test	Start Day	End Day	Start Time	End Time	Duration
ELA					
Grade 4 ELA		0.02	–0.01	0.03	–0.01
Grade 5 ELA	–0.01	0.01	-0.01	0.02	

Grade 6 ELA	0.02		0.01		
Grade 7 ELA	0.01	0.03	–0.01	–0.01	0.01
Grade 8 ELA		0.02	–0.01		0.02
Grade 9 ELA		0.01	–0.06	0.02	0.01
Grade 10 ELA	–0.02		–0.08	0.03	0.01
Grade 11 ELA	–0.03		–0.08	0.05	0.01
MATH					
Test	**Start Day**	**End Day**	**Start Time**	**End Time**	**Duration**
Grade 4 MATH	–0.01	0.02	–0.02		
Grade 5 MATH	–0.02	0.01	–0.03	0.04	0.01
Grade 6 MATH	–0.03	0.01		0.03	0.01
Grade 7 MATH	–0.01	0.01	–0.04	0.06	
Grade 8 MATH		0.01	–0.01	0.04	
Algebra I	–0.05	0.01	–0.12	0.08	0.04
Geometry		0.03	–0.11	0.10	0.03
Algebra II	–0.04	0.04	–0.13	0.12	0.05

1.11 SUMMARY OF VALIDITY OF TEST SCORE INTERPRETATIONS

Evidence for the validity of test score interpretations is strengthened as evidence supporting test score interpretations accrues. In this sense, the process of seeking and evaluating evidence for the validity of test score interpretation is ongoing. Nevertheless, there currently exists sufficient evidence to support the principle claims for the test scores, including that AzMERIT test scores indicate the degree to which students have achieved the Arizona College and Career Ready Standards at each grade level, and that students scoring at the proficient level or higher demonstrate levels of achievement consistent with national benchmarks indicating that they are on track to college readiness. These claims are supported by evidence of a test development process that ensures alignment of test content to the ACCRS, a standard setting process that yielded performance standards consistent with those of rigorous, national benchmarks. Confirmatory factor analyses indicate that the subject area assessments are unidimensional and therefore consistent with the measurement model, but also that the hypothesized reporting strand structure of the AzMERIT provides significant additional information about student achievement. In addition, test scores on the AzMERIT correlate strongly with other measures of subject area achievement, and demonstrate differential relationships across subject area assessments.

2. BACKGROUND OF ARIZONA STATEWIDE ASSESSMENTS

In November 2014, the Arizona State Board of Education adopted Arizona's Measurement of Educational Readiness to Inform Teaching, or AzMERIT, to measure student mastery of the Arizona academic standards and progress toward college and career readiness. The AzMERIT measures English language arts in grades 3-11, and math in grades 3-8 and following completion of high school coursework in Algebra I, Geometry, and Algebra II. The

Arizona Department of Education worked with the American Institutes for Research to develop and administer the AzMERIT beginning in the spring of 2015. In accordance with state requirements, the AzMERIT was designed to[16]:

- Align to the academic standards adopted by the Arizona State Board of Education in 2010 (Arizona College and Career Ready Standards, or ACCRS)
- Supply criterion referenced summative assessments for grades 3 through 8, and criterion referenced end of course assessments in identified high school math and English language arts courses for implementation beginning in the 2014-15 school year
- Assess, without bias, a range of basic knowledge and lower level cognitive skills and higher order, analytical thinking skills in writing, analysis, and problem-solving across subjects, using multiple assessment methods
- Provide valid, reliable and timely data to educators and policy makers to advance the academic success of Arizona students and inform the State's accountability measures
- Communicate results to students, parents and educators, in a clear and timely manner to guide instruction
- Provide an accurate perspective of the quality of learning occurring within classrooms and schools
- Offer educators, students, and families critical tools to improve student achievement, including, but not limited to, formative and interim assessments, sample items and practice tests
- Allow meaningful national or multistate comparisons of school and student achievement
- Use 21st Century technology to deliver the assessment, as available infrastructure allows
- Ensure clarity, transparency, accuracy and security in all aspects of assessment development, deployment, scoring and reporting
- Provide for content and psychometric evaluation and validation
- Establish the involvement of Arizona stakeholders – educators, students, parents, institutions of higher education, and business – in the development of the test, test related materials, and achievement levels indicative of college and career readiness
- Demonstrate accessibility for all students, with optimal access for English language learners and students with special needs
- Respect Arizona's local control of the selection of classroom instructional materials
- Satisfy assessment goals in a cost-efficient manner

The AzMERIT was first administered in spring 2015, assessing proficiency in ELA in grades 3 through 11, math in grades 3 through 8, and following completion of Algebra I, Geometry, and Algebra II (or similar) coursework. Following the initial administration, the AzMERIT in grades 3 through 8 will be administered in the spring of each academic year; tests assessing high school end-of-course tests will be administered in the fall, spring, and summer of each academic year.

The Rasch model, and Masters' (1982) Partial Credit Model, an extension of the one parameter Rasch model that allows for graded responses, was used to estimate item parameters for the AzMERIT. Item pools for grade level summative and End of Course assessments were calibrated following the first operational administration in spring

[16] Standard 7.1 – The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. When particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.
Standard 7.2 – The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

2015. A vertical linking design was also implemented to produce a common vertical scale across grade levels to monitor student growth across grades 3 through 8, as well as the high school EOC assessments. In subsequent years, pre-equated bank item parameter estimates will be applied directly for final scoring and reporting, a strategy that allows for more rapid reporting of tests administered online.

2.1 DEVELOPMENT OF ARIZONA COLLEGE AND CAREER READY STANDARDS (ACCRS)

In 2010, the Arizona State Board of Education adopted new academic content standards in ELA and mathematics that reflect high expectations all Arizona students and strive to ensure that high school graduates are college and career ready. The Arizona College and Career Ready Standards (ACCRS) in mathematics describe expectations for learning in grades K-8 and the first three high school courses (Algebra I, Geometry, Algebra II, or Mathematics 1,2,3) plus specific standards that could be included in a fourth high school credit mathematics course. The ACCRS in ELA describe the reading, writing, language, and speaking and listening skills that students should acquire from grade K-12. The standards can be found at http://www.azed.gov/standards-practices/.

2.2 AZMERIT TEST DESIGN

The AzMERIT is a series of fixed form assessments that are intended to be administered online, although the assessment is offered as a dual mode, online and paper, assessment to accommodate schools that are not yet ready to transition to the online testing environment. A common operational base form is administered to all students within a given test grade and subject. Each assessment is comprised of two to three discrete test sessions. The AzMERIT operational item pools include a variety of selected response, machine-scored constructed responses (MSCR), and some hand-scored constructed-response items in the paper math forms where MSCR items could not readily be rendered for paper test administration. AzMERIT also includes essay responses. In spring 2016, a sample of online responses was hand-scored (100% double scoring with resolution of all discrepancies) for purposes of developing statistical models for machine scoring the remaining online responses. Essays captured on paper were hand-scored.

Six types of MSCR items were included in the AzMERIT forms: graphic response, natural language, equation response, hot text, and table input items. The graphic response item types require students to place objects or move objects around in the answer space. A student can also plot points, draw lines, and draw shapes. The natural language item types require students to type an English language answer. The equation response items require students to enter a value or equation. Hot text items ask students to select or rearrange sentences or phrases in a passage. The table input item types require students to input numerical values into a table. The validity of computer-assigned scores for constructed-response items was evaluated following the spring 2015 online administration of the embedded field test items. Rubric validation for all operational test items was completed prior to test construction and was based on the previous field test administration of those items.

Each ELA assessment included one writing essay prompt that required an extended essay response. For the online test administrations, students were randomly administered one of six writing tasks. A random sample of student responses to each writing task was selected for human scoring. These responses were scored by two human raters on three distinct scoring dimensions or rubrics: Statement of Purpose/Focus and Organization, Evidence/Elaboration, and Conventions/Editing, with any discrepancy adjudicated in a resolution score. This sample of essay responses and writing scores were used to develop the statistical models used for machine scoring the remaining online essay responses. All essay responses captured from paper tests were hand scored. In addition, hand-scoring was required for a subset of math items administered on paper, generally equation items, where it was not possible to represent the item on paper in a way that allowed machine-scoring.

3. SUMMARY OF FALL 2015 OPERATIONAL TEST ADMINSTRATION

The following tests were administered in fall 2015:

- ELA (reading and writing) in grades 9 through 11
- Math in grades 9 through 11, following completion of Algebra I, Geometry, and Algebra II, or similar, coursework

Online administration of the AzMERIT occurred from October 26 through December 4, 2015. The paper version of the AzMERIT was administered between October 26 and November 6, 2015.

The scoring and reporting of the fall 2015 assessments used the items parameters calibrated following the spring 2015 administration and vertical scale and performance standards established in summer 2015. This section summarizes the operational test results for the fall 2015 administration of the AzMERIT.

3.1 STUDENT POPULATION AND PARTICIPATION

Assessment data for operational analyses included Arizona students who meet minimum attemptedness requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and math is presented in Exhibits 3.1.1 and 3.1.2 by assessment and subgroup.[17] Tables in Appendix D show the demographic composition of test takers by mode of test administration.

Exhibit 3.1.1 Number of Students Participating in ELA Assessments, by Test

Group	ELA 9	ELA 10	ELA 11
All Students	3798	4043	6093
Female	1765	1853	2965
Male	2033	2190	3128
Unknown	N/A	N/A	N/A
African American	237	235	404
Asian	68	85	126
Native Hawaiian/Pacific Islander	14	17	21
Hispanic/Latino	1711	1713	2886
American Indian or Alaskan	306	255	330
White	1376	1598	2198
Multiple Ethnicities	83	133	125
Limited English Proficiency	79	79	79
Special Education	523	523	523

Exhibit 3.1.2 Number of Students Participating in Math Assessments, by Test

Group	Algebra I	Geometry	Algebra II
All Students	7359	5887	5412
Female	3514	2730	2738

[17] Standard 1.8 – The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio demographic and developmental characteristics.

Group	Algebra I	Geometry	Algebra II
Male	3845	3157	2674
Unknown	N/A	N/A	N/A
African American	532	450	338
Asian	212	110	175
Native Hawaiian/Pacific Islander	35	19	26
Hispanic/Latino	3066	2811	2160
American Indian or Alaskan	285	222	245
White	3069	2098	2315
Multiple Ethnicities	155	176	153
Limited English Proficiency	118	91	75
Special Education	449	517	259

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

The state summary results for the average scale scores, standard deviation, and minimum and maximum observed scale scores are presented in Exhibit 3.2.1.

Exhibit 3.2.1 Test Score Summary Statistics

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Min.	Max
ELA					
9	3798	2551.01	26.60	2454.00	2638.34
10	4043	2558.93	24.41	2458.47	2648.91
11	6093	2562.52	24.77	2474.47	2652.49
Math					
Algebra I	7359	3674.21	35.40	3577.00	3786.95
Geometry	5887	3671.83	28.24	3609.46	3802.43
Algebra II	5412	3692.78	33.45	3629.35	3828.53

The percentage of students in each performance level by grade and content area, as well as the percent of students at or above Proficient are presented in Exhibit 3.2.2.

Exhibit 3.2.2 Percentage of Students in Performance Levels

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
ELA						
9	3798	56	25	17	2	19
10	4043	60	21	15	3	19
11	6093	60	19	16	5	20
Math						
Algebra I	7359	40	21	27	12	39
Geometry	5887	54	29	15	2	17

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
Algebra II	5412	46	24	24	6	30

3.3 STUDENT PERFORMANCE BY SUBGROUP

Exhibit 3.3.1 and 3.3.2 present the number and percentage of students in each grade and subject at each performance level, by gender and ethnicity, including female, male, African American, Asian, Alaskan/Hawaiian Native Hispanic/Latino, American Indian, White, Multiple Ethnicities, limited English proficiency and special education.

Exhibit 3.3.1 Number of Students At Each Performance Level by Subgroups-Fall 2015

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency
							ELA						
9	Min Proficient	2127	883	1240	147	22	4	1163	214	537	28	301	81
	Part. Proficient	950	494	447	57	23	8	342	64	427	20	32	4
	Proficient	646	335	305	31	14	2	171	28	358	24	14	0
	Highly Proficient	76	53	41	0	9	0	17	0	55	11	0	0
10	Min Proficient	2426	1038	1402	155	29	6	1233	186	767	59	315	46
	Part. Proficient	849	408	438	42	23	4	308	43	384	40	33	2
	Proficient	606	315	307	31	28	4	154	20	352	29	11	1
	Highly Proficient	121	74	66	7	5	3	34	5	96	5	7	0
11	Min Proficient	3656	1720	1971	283	44	16	2049	241	967	66	455	76
	Part. Proficient	1158	623	563	65	38	1	462	50	528	30	42	2
	Proficient	975	474	469	48	33	4	289	30	528	24	26	1
	Highly Proficient	305	148	125	8	11	0	58	10	176	5	5	0
							Mathematics						
Algebra I	Min Proficient	2944	1300	1653	261	32	9	1594	154	829	64	301	89
	Part. Proficient	1545	773	769	122	25	8	644	60	614	28	49	17
	Proficient	1987	1019	961	112	78	11	613	54	1074	40	67	13
	Highly Proficient	883	422	500	32`	76	7	184	17	552	23	27	1
Geometry	Min Proficient	3179	1474	1673	294	36	7	1715	124	902	88	367	71
	Part. Proficient	1707	819	884	117	32	5	759	56	692	58	83	15
	Proficient	883	382	505	41	34	3	309	38	441	26	52	5
	Highly Proficient	118	55	95	0	8	4	28	4	84	4	16	0
Algebra II	Min Proficient	2490	1232	1257	166	39	10	1274	172	741	69	194	65
	Part. Proficient	1299	712	615	85	32	4	475	54	625	21	36	6
	Proficient	1299	657	642	81	74	6	346	17	741	47	26	4

	Highly Proficient	325	137	187	7	32	6	43	0	208	14	3	0

Exhibit 3.3.2 Percentage of Students At Each Performance Level by Subgroups—Fall 2015

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency
ELA													
9	Min Proficient	56	50	61	62	32	29	68	70	39	34	86	95
	Part. Proficient	25	28	22	24	34	57	20	21	31	24	9	5
	Proficient	17	19	15	13	21	14	10	9	26	29	4	0
	Highly Proficient	2	3	2	0	13	0	1	0	4	13	0	0
10	Min Proficient	60	56	64	66	34	35	72	73	48	44	86	94
	Part. Proficient	21	22	20	18	27	24	18	17	24	30	9	4
	Proficient	15	17	14	13	33	24	9	8	22	22	3	2
	Highly Proficient	3	4	3	3	6	18	2	2	6	4	2	0
11	Min Proficient	60	58	63	70	35	76	71	73	44	53	87	96
	Part. Proficient	19	21	18	16	30	5	16	15	24	24	8	3
	Proficient	16	16	15	12	26	19	10	9	24	19	5	1
	Highly Proficient	5	5	4	2	9	0	2	3	8	4	1	0
Mathematics													
Algebra I	Min Proficient	40	37	43	49	15	26	52	54	27	41	67	75
	Part. Proficient	21	22	20	23	12	23	21	21	20	18	11	14
	Proficient	27	29	25	21	37	31	20	19	35	26	15	11
	Highly Proficient	12	12	13	6	36	20	6	6	18	15	6	1
Geometry	Min Proficient	54	54	53	66	33	37	61	56	43	50	71	78
	Part. Proficient	29	30	28	26	29	26	27	25	33	33	16	16
	Proficient	15	14	16	9	31	16	11	17	21	15	10	5
	Highly Proficient	2	2	3	0	7	21	1	2	4	2	3	0
Algebra II	Min Proficient	46	45	47	49	22	38	59	70	32	45	75	87
	Part. Proficient	24	26	23	25	18	15	22	22	27	14	14	8
	Proficient	24	24	24	24	42	23	16	7	32	31	10	5
	Highly Proficient	6	5	7	2	18	23	2	0	9	9	1	0

Note: Part. = Partially; Min. = Minimally; Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander.

3.4 RELIABILITY

Reliability refers to the consistency or precision of test scores and performance level classifications, and essentially addresses the question of how likely would a student be to achieve the same score, or be classified in the same performance level, across multiple administrations of equivalently constructed and administered test forms. As part of each test administration, the reliability of test scores and performance classifications is evaluated from a variety of perspectives. The reliability evidence of the AzMERIT ELA and math are provided with respect to both classical and IRT indices of internal consistency of test scores, and decision accuracy and consistency of performance level classifications.[18]

Test score reliability is traditionally estimated using both classical and IRT approaches. Classical estimates of test reliability such as coefficient alpha, provide a general index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. The equations and formula for estimating reliability are presented in Appendix E.[19]

3.4.1 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Classical estimates of test reliability such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. Exhibit 3.4.1.1 shows the Cronbach's alpha internal consistency estimates for each of the AzMERIT ELA and math assessments. Internal consistency estimates are uniformly in the 0.9 range, consistent with most similar length achievement tests.

Exhibit 3.4.1.1 Internal Consistency Reliabilities for AzMERIT Scores

Grade/Course	ELA		Math	
	Reliability	Variance	Reliability	Variance
9/Algebra I	0.87	707	0.91	1264
10/Geometry	0.86	589	0.81	735
11/Algebra II	0.86	599	0.87	1111

3.4.2 STANDARD ERROR OF MEASUREMENT

Because measurement error is conditional on test information, the precision of test scores varies with respect to the information value of the test at each location along the ability distribution. Precision of individual test scores is critically important to valid test score interpretation. Test scores are most precise in locations where test information is greatest. Because relatively little test information is targeted to measurement of very low and high performing students, the precision of test scores decreases near the tails of the ability distribution.

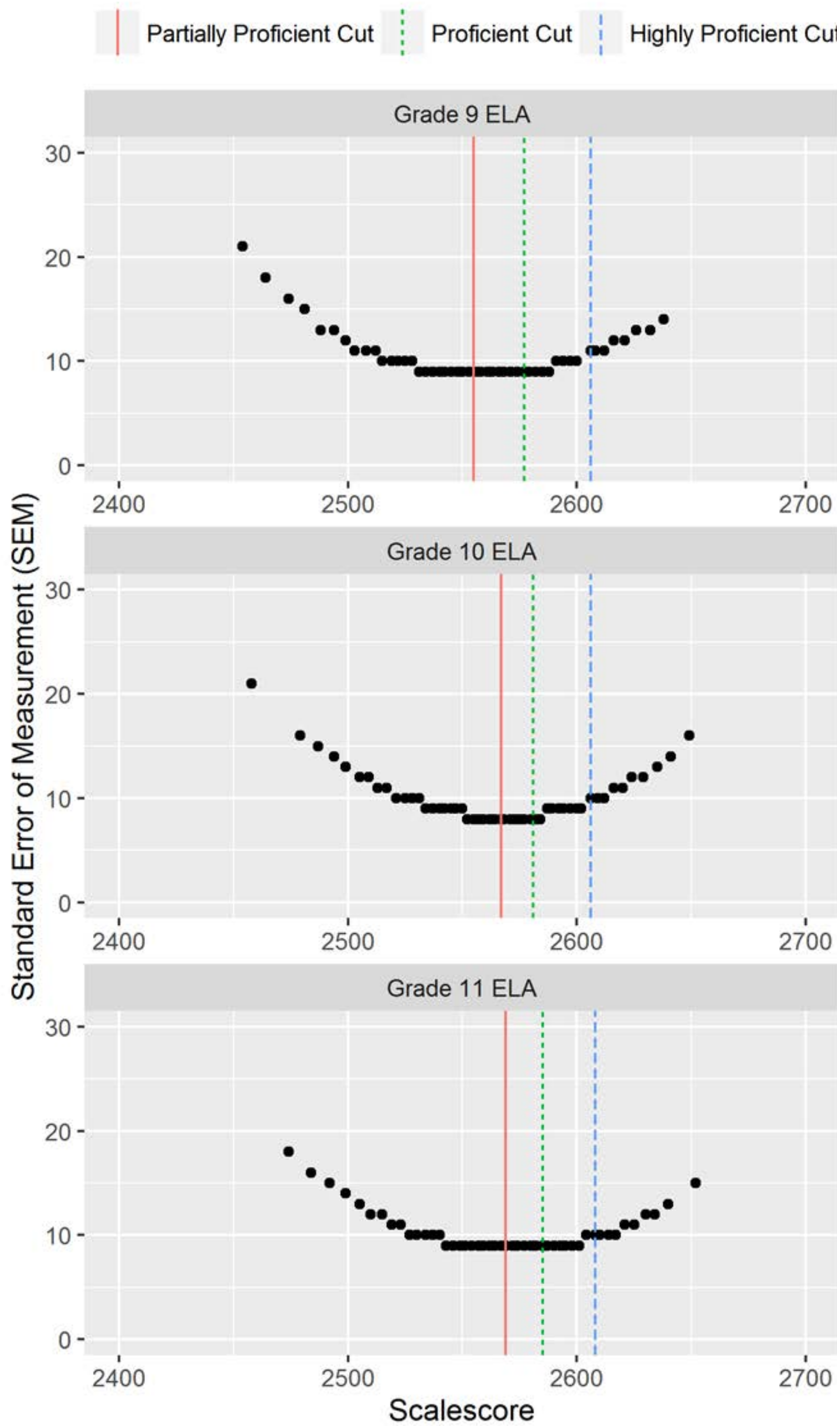
The figure in Exhibit 3.4.2 present graphically the standard errors of measurement for the AzMERIT ELA and math assessments. Each figure also includes the location of the three AzMERIT performance standards. As the figures indicate, the AzMERIT test scores are most precise near the middle of the ability distribution, and especially near

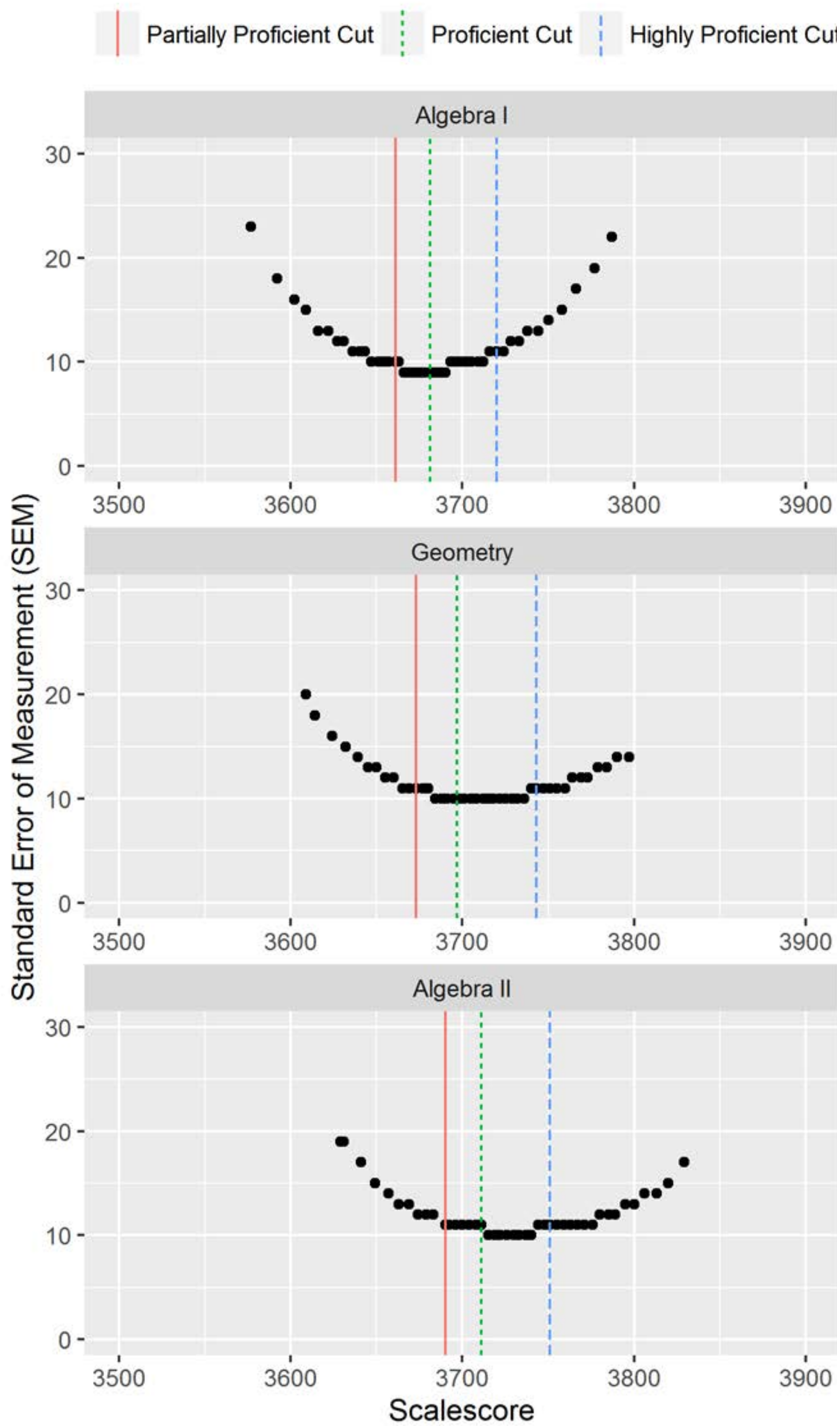
[18] Standard 2.2 – The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.
Standard 2.3 – For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.
[19] Standard 2.19 – Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.

the Partially Proficient and Proficient performance standards.[20] Test scores near the tails of the ability distribution are somewhat less precise, as expected. An SEM of .3 on the theta metric is consistent with an internal consistency of 0.9. The tables in Appendix F show the mean SEMs for students scoring in each of the performance levels on the AzMERIT reporting scale. While these tables also indicate that the AzMERIT test scores are somewhat more precise for test scores near the middle of the scale, they also show that test scores remain precise even for students in the lowest and highest performance level classifications.

[20] Standard 2.14 – When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Exhibit 3.4.2.1 Overall Standard Error of Measurement for ELA and Math





Partially Proficient Cut
Proficient Cut
Highly Proficient Cut
Algebra I
Geometry
Algebra II
Standard Error of Measurement (SEM)
30
20
10
0
3500
3600
3700
3800
3900
Scalescore

3.4.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index is computed in terms of the probabilities of consistent classification of students as specified in standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).[21] This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications are typically estimated on a single-form test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995).

The classification index can be examined for decision accuracy and decision consistency. Decision accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Decision consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate, equivalently constructed test form—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test administrations.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with measurement error.

3.4.4 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution, we can estimate the above probabilities directly using the likelihood function. The likelihood function of the achievement attribute, designated θ, given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated theta is below the cut score, the probability of *at or above* the cut score is an estimate of the chance that this student is misclassified as below the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

In Exhibit 3.4.4.1, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassifications, false positives and false negatives, occur when students' true score classifications differ from their observed score classifications (e.g., a student whose true score results in a Proficient level classification, but is classified incorrectly as Partially Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of

[21] Standard 2.16 – When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

students falsely above the cut score; N_{00} represents the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

Exhibit 3.4.4.1 Classification Accuracy

		Classification on a Form Actually Taken	
		At or Above the Cut Score	**Below the Cut Score**
Classification on True Score	**At or Above the Cut Score**	N_{11} (Truly above the cut)	N_{10} (False negative)
	Below the Cut Score	N_{01} (False positive)	N_{00} (Truly below the cut)

3.4.5 CLASSIFICATION CONSISTENCY

As shown in Exhibit 3.4.5.1, consistent classification occurs when two forms agree on the classification of a student as either *at and above* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit 3.4.5.1 Classification Consistency

		Classification on the 2nd Form Taken	
		Above the Cut Score	**Below the Cut Score**
Classification on the 1st Form Taken	**At or Above the Cut Score**	N_{11} (Consistently above the cut)	N_{10} (Inconsistent)
	Below the Cut Score	N_{01} (Inconsistent)	N_{00} (Consistently below the cut)

3.4.6 CLASSIFICATION RELIABILITY ESTIMATES

Exhibit 3.4.6.1 presents the classification accuracy and consistency indexes for the fall 2015 administration of AzMERIT. Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error, while the accuracy index assumes only a single test score and a true score, which does not include measurement error.

Exhibit 3.4.6.1 Classification Accuracy and Consistency Indexes for Performance Standards

	Accuracy			Consistency		
Grade	**Partially Proficient**	**Proficient**	**Highly Proficient**	**Partially Proficient**	**Proficient**	**Highly Proficient**
			ELA			
9	0.91	0.94	0.98	0.87	0.91	0.98
10	0.90	0.93	0.98	0.86	0.90	0.97
11	0.90	0.92	0.97	0.86	0.90	0.96

	Accuracy			Consistency		
Grade	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
			MATH			
Algebra I	0.91	0.93	0.96	0.88	0.90	0.94
Geometry	0.86	0.94	0.99	0.81	0.91	0.99
Algebra II	0.89	0.92	0.97	0.86	0.88	0.96

3.4.7 RELIABILITY FOR SUB-GROUPS IN THE POPULATION

Exhibit 3.4.7.1 and 3.4.7.2 shows the mean reliability for each of the subgroups: African Americans, Asian, Native Hawaiians/Pacific Islanders, Hispanic/Latinos, American Indian or Alaskans, Whites, Multiple Ethnicities and females and males (regardless of racial or ethnic group).[22] Each racial and/or ethnic group was composed of approximately equal numbers of males and females. As the Exhibit indicates, internal consistency reliabilities are consistent across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups. Where reliability estimates are attenuated, there is an associated decrease in variance within the subgroup population, indicating that the decrease in reliability is likely due to a restriction in range.

Exhibit 3.4.7.1 Internal Consistency Reliability by Subgroup– ELA

	Grade 9 ELA		Grade 10 ELA		Grade 11 ELA	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.87	707	0.86	589	0.86	599
Female	0.87	674	0.86	544	0.85	571
Male	0.87	708	0.87	614	0.86	623
African American	0.85	614	0.86	573	0.85	601
Asian	0.89	884	0.84	480	0.85	554
Native Hawaiian/Pacific Islander	0.79	395	0.87	573	0.74	306
Hispanic/Latino	0.85	590	0.84	494	0.84	530
American Indian or Alaskan	0.83	524	0.85	536	0.82	489
White	0.87	651	0.87	593	0.85	576
Multiple Ethnicities	0.90	984	0.85	514	0.86	612
Limited English Proficiency	0.68	321	0.72	334	0.63	267
Special Education	0.78	457	0.77	383	0.80	479

[22] Standard 2.11 – Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

Exhibit 3.4.7.2 Internal Consistency Reliability by Subgroup – Math

	Algebra I		Geometry		Algebra II	
	Reliability	Variance	Reliability	Variance	Reliability	Variance
All Students	0.91	1264	0.81	735	0.87	1111
Female	0.91	1204	0.79	643	0.86	997
Male	0.91	1318	0.83	811	0.88	1227
African American	0.89	997	0.75	583	0.84	910
Asian	0.91	1557	0.87	1021	0.89	1206
Native Hawaiian/Pacific Islander	0.90	1151	0.94	2003	0.91	1549
Hispanic/Latino	0.89	999	0.78	636	0.82	898
American Indian or Alaskan	0.88	948	0.83	814	0.71	605
White	0.91	1267	0.84	807	0.88	1090
Multiple Ethnicities	0.91	1180	0.78	628	0.89	1273
Limited English Proficiency	0.83	656	0.67	461	0.66	553
Special Education	0.86	861	0.65	482	0.75	764

3.4.8 SUBSCALE RELIABILITY

Coefficient alpha internal consistency reliability estimates associated with the subscales for the Fall 2015 operational forms are presented in Exhibit 3.4.8.1-3.4.8.3. As indicated in the Exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzMERIT. The only exception is the Modeling with Geometry strand in Geometry test. The reason is that nearly 40 percent of students received raw score 0 in this strand. Using marginal reliability is not a good choice because the standard deviation of the observed score is too arbitrary given the rule of estimating the extreme scores.

Exhibit 3.4.8.1 Subscale Reliabilities – ELA Grades 9-11

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 9	0.76	0.66	0.73
Grade 10	0.69	0.70	0.71
Grade 11	0.70	0.59	0.74

Exhibit 3.4.8.2 Subscale Reliabilities – Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.83	0.78	0.48
Algebra II	0.73	0.60	0.66

Exhibit 3.4.8.3 Subscale Reliabilities – Geometry

	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.43	0.54	-0.20	0.60

3.5 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibits 3.5.1-3.5.3. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.[23] The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

Where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y , r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y.

When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underling construct.

Exhibit 3.5.1 Subscale Intercorrelations and Reliability Estimates – ELA Grades 9 to 11

		Observed Correlation		Disattenuated Correlation	
Grade	**Subscale**	**Informational Text**	**Literature**	**Informational Text**	**Literature**
9	Literature	0.63		0.89	
	Writing & Language	0.63	0.55	0.84	0.80
10	Literature	0.65		0.90	
	Writing & Language	0.56	0.57	0.77	0.81
11	Literature	0.61		0.91	
	Writing & Language	0.63	0.58	0.83	0.87

Exhibit 3.5.2 Subscale Intercorrelations and Reliability Estimates – Algebra I & Algebra II

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	**Functions**	**Algebra**	**Functions**
Algebra I	**Functions**	0.80		0.99	
	Statistics	0.61	0.61	0.97	0.99
Algebra II	**Functions**	0.66		1.00	
	Statistics	0.71	0.65	1.00	1.00

Exhibit 3.5.3 Subscale Intercorrelations and Reliability Estimates – Geometry

Grade	Subscale	Observed Correlations	Disattenuated Correlations

[23] Standard 1.21 – When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.

		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	Congruence(C)	0.49			1.00		
	Modeling with Geometry (MG)	0.47	0.49		1.00	1.00	
	Similarity, Right Triangles and Trigonometry (SRTT)	0.49	0.58	0.54	0.97	1.00	1.00

4. SUMMARY OF SPRING 2016 OPERATIONAL TEST ADMINSTRATION

The following AzMERIT assessments were administered in spring 2016:

- ELA (reading and writing) in grades 3 through 11
- Math in grades 3 through 8, Algebra I, Geometry, and Algebra II.

Online administration of the AzMERIT occurred from March 28 through May 6, 2016. The paper version of the AzMERIT was administered between March 28 and April 8, 2016.

The Math tests were scored using pre-equated item parameter estimates derived from the initial online administration of the AzMERIT tests in spring 2015. Thus, no post-equating activities were conducted prior to the scoring and reporting of the Math tests. In spring 2016, in each ELA assessment, students were randomly assigned one of six writing prompts for administration. Following the test administration, all operational items including reading and writing items were concurrently calibrated, and then linked back to the AzMERIT bank scale using the mean-mean equating method. The post-equated item parameters were used for the final scoring and reporting. This section summarizes the operational test results for the spring 2016 administration of the AzMERIT. Detailed descriptions of procedures for item and test development, test administration, scaling, equating, and scoring are presented in subsequent sections.

4.1 STUDENT POPULATION AND PARTICIPATION

Assessment data for operational analyses included Arizona students who meet minimum attemptedness requirements for scoring and reporting. The demographic composition of students taking the AzMERIT in ELA and math is presented in Exhibits 4.1.1 and 4.1.2 by assessment and subgroup.[24] We note that some students participated in an end-of-course assessment rather than a grade level assessment, especially in grade 8 where more advanced students are enrolled in Algebra I courses. The tables in Appendix G show the demographic composition of test takers by mode of test administration.

Exhibit 4.1.1 Number of Students Participating in ELA Assessments, by Test Name

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
All Students	87793	86325	85425	84651	84138	82779	80130	73403	64834
Female	43134	42452	42102	41190	41418	40757	39488	36507	32208
Male	44645	43861	43315	43456	42704	42011	40594	36851	32596

[24] Standard 1.8 – The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio demographic and developmental characteristics.

Group	ELA 3	ELA 4	ELA 5	ELA 6	ELA 7	ELA 8	ELA 9	ELA 10	ELA 11
Unknown	14	12	8	5	16	11	48	45	30
African American	4437	4432	4430	4518	4487	4555	4305	3995	3331
Asian	2422	2322	2466	2542	2419	2409	2423	2302	2120
Native Hawaiian/Pacific Islander	313	349	277	272	245	228	288	207	217
Hispanic/Latino	40327	39205	38172	37569	37123	36622	35381	31489	27202
American Indian or Alaskan	4354	4333	4426	4293	4001	4063	3900	3574	3153
White	33223	33150	33418	33434	33946	33145	32485	30583	27744
Multiple Ethnicities	2703	2522	2228	2018	1901	1745	1297	1198	1031
Limited English Proficiency	8491	8311	6470	4916	4105	3347	3487	2616	1755
Special Education	9581	10183	10204	9549	8903	8553	6946	6024	4996
Free/Reduced Lunch	37747	36855	36060	35466	34799	33432	30003	26487	22191
Accommodation	13439	13017	10931	9227	7612	6470	3111	2191	1649

Exhibit 4.1.2 Number of Students Participating in Math Assessments, by Test Name

Group	Math 3	Math 4	Math 5	Math 6	Math 7	Math 8	Algebra I	Geometry	Algebra II
All Students	88303	86711	85719	84675	81829	69858	82623	71654	60900
Female	43349	42607	42221	41173	40307	34103	40425	35652	30768
Male	44941	44092	43490	43496	41506	35736	42130	35949	30090
Unknown Gender	13	12	8	6	16	19	68	53	42
African American	4488	4491	4454	4543	4437	4204	4506	3662	3196
Asian	2435	2341	2460	2353	1884	1415	2402	2356	1951
Native Hawaiian/Pacific Islander	316	348	279	271	243	200	261	222	192
Hispanic/Latino	40550	39355	38298	37738	36940	33590	37028	30642	25612
American Indian or Alaskan	4404	4353	4451	4333	4074	3803	4453	3595	2920
White	33371	33268	33528	33398	32395	25159	32569	29962	26009
Multiple Ethnicities	2726	2543	2241	2033	1840	1467	1333	1159	977
Limited English Proficiency	8622	8402	6532	4967	4143	3153	3747	2857	1662
Special Education	9731	10265	10299	9628	8993	8456	7416	5319	3576
Free/Reduced Lunch	37157	36790	35932	35315	34687	30530	31860	26108	21308
Accommodation	13518	13134	11006	9309	7611	6402	3068	1886	1247

4.2 CLASSICAL ITEM ANALYSIS

Because AzMERIT is an online assessment system, classical item analysis statistics for multiple-choice (MC) and constructed-response (CR) items reported here are calculated based on all online student responses. Classical item analysis statistics are used to monitor item behavior and investigate irregularities in item scoring throughout the test window for online assessments, and following processing of answer documents for paper test administrations. Classical item analyses ensure that the items function as intended with respect to the underlying scales. For online and paper test administrations, quality assurance reports provide the required item and test statistics for each multiple-choice and constructed-response (CR) item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item during test administration. Key statistics computed and examined include biserial/polyserial correlations for item discrimination, biserial correlations for distractors for selected response items, and proportion correct for item difficulty.

The biserial/polyserial correlations indicate the extent to which each item differentiated between those examinees who possessed the skills being measured and those who do not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The biserial correlations for multiple-choice items is calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, the mean total number correct for student scoring within each of the possible score categories is used. Items are flagged for review by test development experts if the biserial correlation for the keyed (correct) response is less than .25, or changed from previous administration. For multiple-choice items, we also compute the biserial correlation for each of the distractor response options.

The proportion correct score is the average number of available points achieved by students on the item. For dichotomous items, this is simply the proportion of students responding correctly. For polytomous items, the average score on the item is divided by the points available to produce a comparable index. The proportion correct score is commonly referred to as the *p*-value.

Exhibit 4.2.1 presents the average proportion of students responding correctly and average point biserial/polyserial correlations from the spring 2016 online administration of AzMERIT. As indicated in Exhibit 4.2.1, at the elementary school level, especially at grade 3 and grade 4, are somewhat easier than the math items. By the middle school grades, however, the pattern is reversed, with items on the ELA assessments, on average, easier than items on the math assessments. While mean difficulty of ELA items is relatively consistent across grade level assessments, average difficulty of math items increases across grade level and course assessments. The proportion of students responding correctly to test items in the end of course assessments in math was relatively quite low. Mean biserial correlations for the grade level and end of course assessments are reasonably high and consistent across assessments. Exhibit 4.2.2 shows the number of items flagged for proportion correct value, biserial/polyserial corrrelation, distractor biserial/polyserial, and DIF categories for the operational items in the spring 2016 online forms. The flagging criteria are presented in section 5.4.1 and 5.4.3.

Exhibit 4.2.1 Average Proportion Correct and Point Biserial Correlations for Operational Test Items Administered Online

Grade	Average *p*-Value	*p*-Value SD	Average Point-Biserial	Point-Biserial SD
		ELA		
3	0.48	0.18	0.48	0.13
4	0.44	0.23	0.47	0.11
5	0.48	0.21	0.49	0.13
6	0.46	0.21	0.50	0.13
7	0.47	0.18	0.48	0.12
8	0.50	0.18	0.49	0.13
9	0.49	0.18	0.46	0.13
10	0.48	0.18	0.44	0.15
11	0.48	0.15	0.48	0.14
		Math		
3	0.62	0.20	0.50	0.10
4	0.56	0.19	0.48	0.10
5	0.51	0.20	0.48	0.11
6	0.49	0.21	0.49	0.10
7	0.46	0.19	0.45	0.12
8	0.40	0.21	0.43	0.12
Algebra I	0.41	0.16	0.45	0.12
Geometry	0.33	0.17	0.44	0.09
Algebra II	0.28	0.18	0.41	0.09

Exhibit 4.2.2 Number of Items Flagged For P-value, Biserial/Polyserial or DIF for Operational Test Items Administered Online

Grade	Number of Flagged Operational Items			
	Proportion Correct	Biserial/Polyserial Correlation	Biserial Correlation for Distractor	Differential Item Functioning
ELA				
3	0	0	0	7
4	0	0	0	9
5	0	0	0	1
6	0	0	1	7
7	1	0	1	9
8	0	0	0	2
9	0	0	0	3
10	0	1	2	4
11	0	0	0	4
MATH				
3	1	0	0	1
4	0	0	0	1
5	0	0	0	0
6	0	0	1	0
7	2	1	0	0
8	0	1	0	0
AlgI	2	1	2	0
Geo	3	0	2	0
AlgII	1	1	1	2

4.3 ITEM RESPONSE THEORY ANALYSIS

Calibration is the process by which the statistical relationship between item responses and the underlying measurement construct is estimated. Traditional item response models assume a single underlying trait and assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^{n} P(z|\theta),$$

where *Z* represents the vector of item responses, and θ represents a student's true proficiency.

Traditional item response models differ only in the form of the function *P*(*Z*). The one-parameter model (also known as the Rasch model) is used to calibrate dichotomously scored AzMERIT items and takes the form

$$P\left(x_j = 1|\theta_k, b_j\right) = \frac{1}{1+e^{\left(\theta_k - b_j\right)}} = P_{j1}(\theta_k).$$

The b parameter is often called the *location* or *difficulty* parameter—the greater the value of b, the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered response categories (i.e., partial credit items), AzMERIT items are calibrated using the Rasch family Masters' (1982) partial credit model. Under Masters' model, the probability of a response in category i for an item with m_j categories can be written as

$$P\left(x_j = i|\theta_k, b_{j0} \ldots b_{jm_j-1}\right) = \frac{e^{\sum_{v=0}^{i}(\theta_k - b_{jv})}}{\sum_{g=0}^{m_j-1} e^{\sum_{v=0}^{g}(\theta_k - b_{jv})}}.$$

The tables in Appendix H provide Rasch and Masters' partial credit model item parameter estimates for the spring 2016 operational test items. Since AzMERIT is an online assessment system, bank item parameters were estimated based only on online responses to test items. Exhibit 4.3.1 presents the mean and standard deviation of the Rasch item parameters by item type for each test for items administered online. Item types include traditional four-option multiple choice (MC) items, technology-enhanced (TE) selected response items which may require students to select one or more options, and machine-scored constructed response (MSCR) items for which students' constructed responses are scored electronically using explicit rubrics. In addition, the average Rasch difficulty is presented for each scoring dimension of the writing prompt administered at each grade. As illustrated in Exhibit 4.3.1, selected-response items are, on average, less difficult than the constructed-response item types.

Exhibit 4.3.1 Rasch Summary Statistics by Item Type for Items Administered Online

Grade/Course	MC			TE Selected Response			MSCR			Writing Prompt Average Rasch		
	N	Avg Rasch	SD	N	Avg Rasch	SD	N	Avg Rasch	SD	Org	Ev/Elab	Conv
ELA												
3	28	-0.01	0.70	10	-0.12	1.26	3	0.81	0.93	1.61	1.67	-1.18
4	23	-0.14	0.75	13	0.60	1.26	5	1.01	1.12	3.70	4.09	0.00
5	19	-0.06	0.67	13	0.19	1.24	9	1.40	1.36	2.38	2.90	-0.94
6	26	-0.18	0.80	9	0.45	0.78	6	0.96	1.12	2.29	2.89	-1.27
7	22	0.12	0.69	16	0.48	0.68	3	0.80	0.72	2.36	2.66	-1.42
8	25	-0.18	0.75	10	0.21	0.97	6	1.09	1.03	1.03	1.21	-1.62
9	24	-0.13	0.68	12	0.21	0.78	7	1.01	0.36	1.35	1.63	-1.67
10	30	-0.06	0.58	10	0.57	1.27	3	1.44	0.43	0.81	1.18	-2.09
11	23	-0.09	0.38	15	0.13	0.94	5	0.17	0.55	0.25	0.73	-1.90
Mathematics												
3	16	-0.52	1.39	1	2.28	NA	28	0.23	1.33	--	--	--
4	13	-0.58	1.10	7	0.29	1.22	25	0.29	1.24	--	--	--
5	13	-0.47	0.79	4	-0.50	1.22	28	0.44	1.35	--	--	--
6	12	-0.51	1.09	5	1.16	1.48	30	0.00	1.41	--	--	--

7	24	-0.35	1.03	3	0.50	0.90	20	0.65	0.92	--	--	--
8	18	-0.91	0.93	3	0.25	0.74	26	0.51	1.35	--	--	--
Algebra I	28	-0.14	0.78	2	0.96	1.22	17	0.26	1.00	--	--	--
Geometry	19	-0.70	0.89	5	0.69	0.91	23	0.43	1.27	--	--	--
Algebra II	20	-1.04	0.77	4	0.77	0.70	23	0.78	0.92	--	--	--

Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are flagged if Infit or Outfit values are less than 0.7 or greater than 1.3. Exhibit 4.3.2 summarizes the number of online administered operational test items with Infit and Outfit statistics within the range of .7 to 1.3.

Exhibit 4.3.2 Summary of Item Fit Statistics for Items Administered Online

Grade/ Course	Infit			Outfit		
	Below 0.7	**Between .7-1.3**	**Above 1.3**	**Below 0.7**	**Between .7-1.3**	**Above 1.3**
ELA						
3	0	59	0	0	57	2
4	0	59	0	0	55	4
5	0	58	1	1	56	2
6	0	59	0	1	56	2
7	0	59	0	1	58	0
8	0	59	0	6	51	2
9	0	61	0	5	55	1
10	0	61	0	5	56	0
11	0	61	0	4	57	0
Mathematics						
3	0	44	1	0	36	9
4	0	42	3	3	35	7
5	0	42	3	1	34	10
6	0	46	1	3	35	9
7	0	45	2	4	36	7
8	0	46	1	3	36	8
Algebra I	0	45	2	3	37	7
Geometry	0	47	0	3	39	5
Algebra II	0	47	0	5	40	2

4.4 SUMMARY OF OVERALL STUDENT PERFORMANCE

The state summary results for the average scale scores, standard deviation, and minimum and maximum observed scale scores are presented in Exhibit 4.4.1. The AzMERIT bank scale was established based on the spring 2015 assessments in which the item calibrations were centered on items rather than persons, resulting in operational test forms with mean difficulty of zero and standard deviation of one. Because calibrations were not centered on persons, the standard deviation of ability estimates is not expected to be 30, as might be implied by the scaling transformation.

Exhibit 4.4.1 Test Score Summary Statistics

Test	Number Tested	Scale Score			
		Mean	Std. Dev.	Min.	Max
ELA					
3	87793	2501.38	31.81	2605	2395
4	86325	2517.62	33.89	2610	2400
5	85425	2536.70	34.41	2629	2419
6	84651	2540.77	33.74	2641	2431
7	84138	2551.77	31.12	2648	2438
8	82779	2555.26	32.38	2658	2448
9	80130	2564.11	31.43	2664	2454
10	73403	2564.75	29.05	2668	2458
11	64834	2567.09	31.49	2675	2465
Math					
3	88303	3524.19	44.86	3605	3395
4	86711	3552.27	40.62	3645	3435
5	85719	3587.75	41.57	3688	3478
6	84675	3615.26	42.60	3722	3512
7	81829	3632.28	35.61	3739	3529
8	69858	3651.11	36.11	3776	3566
Algebra I	82623	3669.80	34.18	3787	3577
Geometry	71654	3684.30	36.07	3819	3609
Algebra II	60900	3695.35	33.00	3839	3629

The percentage of students in each performance level by grade and content area, as well as the percent of students at or above Proficient are presented in Exhibit 4.4.2.

Exhibit 4.4.2 Percentage of Students in Performance Levels

Grade	Number Tested	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	% At or Above Proficient
			ELA			
3	87793	45	14	29	12	41
4	86325	40	14	34	12	46
5	85425	32	23	32	13	45
6	84651	41	21	31	6	38
7	84138	38	21	35	6	41
8	82779	44	23	25	7	33
9	80130	37	28	25	9	35
10	73403	54	17	20	10	30
11	64834	53	17	19	11	30
			Math			
3	88303	25	30	28	17	45
4	86711	29	26	34	10	44
5	85719	27	27	30	15	46
6	84675	37	24	24	14	39
7	81829	46	23	22	9	31
8	69858	51	23	17	9	26
Algebra I	82623	45	19	27	9	36
Geometry	71654	41	24	28	8	35
Algebra II	60900	46	25	22	7	29

4.5 STUDENT PERFORMANCE BY SUBGROUP

Exhibit 4.5.1 and 4.5.2 present the number and percentage, respectively, of students in each grade and subject at each performance level, by gender and ethnicity, including female, male, African American, Asian, Alaskan/Hawaiian Native, Hispanic/Latino, American Indian, White, and Multiple Ethnicities, other demographic information such as special education (SPED) and limited English proficiency (LEP).

Exhibit 4.5.1 Number of Students At Each Performance Level by Gender, Ethnicity, and Other Demographic Information

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodation
							ELA								
3	Minimally Proficient	39507	17685	21430	2485	509	141	22583	2874	9635	973	7377	7217	21138	11128
	Partially Proficient	12291	6039	6250	666	315	44	5646	610	4651	378	766	594	5662	1096
	Proficient	25460	13372	12054	1065	896	100	9275	784	12293	892	1054	509	8682	1098
	Highly Proficient	10535	6039	4911	266	678	31	2420	131	6645	460	383	85	2265	117
4	Minimally Proficient	34530	1596	1022	185	7449	1647	16907	1614	10613	13597	7841	7064	18796	10717
	Partially Proficient	12086	620	325	49	3921	693	4973	353	5519	5702	815	665	5528	1091
	Proficient	29351	1640	743	94	17250	1517	9282	479	18254	17544	1324	582	10688	1125
	Highly Proficient	10359	620	232	21	10585	477	1989	76	8490	7018	305	0	2211	84
5	Minimally Proficient	27336	1152	912	119	4581	1151	13367	1225	7999	10829	7551	5111	14785	8603
	Partially Proficient	19648	1019	592	66	6108	1151	8689	557	8420	9529	1429	971	9376	1633
	Proficient	27336	1551	715	75	14887	1416	9023	401	16841	15593	1020	324	9736	635
	Highly Proficient	11105	709	271	17	12597	708	2339	67	8841	7797	306	65	2524	60
6	Minimally Proficient	34707	1672	1144	144	6762	1674	17720	1352	10298	14340	7735	4424	18797	8097
	Partially Proficient	17777	994	534	60	6387	1116	7355	383	8650	9126	955	344	7803	768
	Proficient	26242	1536	737	60	17657	1202	7355	262	17712	16513	668	147	7803	338
	Highly Proficient	5079	361	127	8	7138	301	669	20	4119	3476	95	0	709	24
7	Minimally Proficient	31972	1481	1040	115	6682	1800	16294	1236	9940	13238	7211	3653	17052	6667
	Partially Proficient	17669	987	508	54	5940	960	7468	380	8284	8541	890	287	7656	616
	Proficient	29448	1705	774	69	17819	1080	9165	285	19052	17509	712	164	9048	313
	Highly Proficient	5048	359	121	7	6311	160	1018	19	4142	3416	89	0	696	16

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodation
8	Minimally Proficient	36423	1776	1180	125	7324	1828	17898	1239	12227	15964	7270	2878	18722	5884
	Partially Proficient	19039	1093	530	50	6958	1016	7623	314	9782	10083	770	234	7689	395
	Proficient	20695	1275	554	46	13550	975	6298	175	13857	11763	428	201	6018	174
	Highly Proficient	5795	410	145	9	8789	244	1326	17	4891	3781	86	33	1003	17
9	Minimally Proficient	29648	1378	1042	144	6015	1209	15268	791	9872	12178	5487	2650	15002	2716
	Partially Proficient	22436	1248	654	78	7784	1287	9421	337	10662	10554	972	523	8401	283
	Proficient	20033	1205	557	55	12383	1053	6497	143	13426	12178	417	279	5401	96
	Highly Proficient	7212	474	194	14	9553	390	1624	26	5923	5277	69	35	1200	16
10	Minimally Proficient	39638	1998	1335	137	9132	2073	19879	934	14603	16951	5241	2171	18011	2030
	Partially Proficient	12479	719	368	31	4408	429	4893	144	6936	6265	422	235	3973	83
	Proficient	14681	879	414	29	8817	715	4282	96	9857	8476	301	183	3443	60
	Highly Proficient	7340	439	184	10	9132	357	1529	24	5476	5159	60	52	1059	18
11	Minimally Proficient	34362	1566	1230	143	7889	1860	17756	784	12883	14342	4396	1404	14868	1496
	Partially Proficient	11022	600	339	33	4080	441	4439	144	6120	5215	300	176	3329	85
	Proficient	12318	733	360	33	6801	568	4162	93	8052	8149	200	140	2885	51
	Highly Proficient	7132	400	191	11	8161	284	1387	21	5153	4889	50	35	1110	17
							Mathematics								
3	Minimally Proficient	22076	10404	11685	1705	195	82	12976	1762	5006	572	5449	4828	12262	7453
	Partially Proficient	26491	13438	12583	1391	390	95	13787	1585	8343	709	2335	2587	12633	3947
	Proficient	24725	12571	12583	987	779	95	10138	837	11346	872	1362	1035	9661	1782
	Highly Proficient	15012	6936	8089	404	1071	47	4055	220	8676	572	584	172	2973	336
4	Minimally Proficient	25146	12356	13228	1976	211	94	14955	2046	5656	585	6570	5461	14348	8586
	Partially Proficient	22545	11504	11464	1257	351	90	11413	1306	7652	661	1950	2016	10669	2995
	Proficient	29482	14486	14550	1078	1053	129	11019	914	14305	966	1437	840	10301	1436
	Highly Proficient	8671	4261	4850	180	726	31	1968	131	5656	331	308	84	1840	117
5	Minimally Proficient	23144	10555	13047	1782	172	67	13404	2092	5364	493	6797	4311	12936	7504
	Partially Proficient	23144	11822	11307	1381	369	70	11489	1291	8047	583	2060	1502	10780	2425

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodation
	Proficient	25716	13511	12612	980	886	103	9957	890	12405	740	1133	588	9342	924
	Highly Proficient	12858	6333	6524	312	1009	39	3447	178	7376	426	309	131	2875	153
6	Minimally Proficient	31330	14411	16963	2408	353	87	17737	2556	7682	630	7510	3974	17304	7389
	Partially Proficient	20322	10293	10004	1045	376	81	9435	997	7682	488	1252	695	8829	1285
	Proficient	20322	10705	10004	772	753	60	7548	607	10353	549	674	248	6710	514
	Highly Proficient	11855	5764	6524	273	871	46	3019	173	7682	366	289	50	2472	121
7	Minimally Proficient	37641	17735	19923	2751	433	114	20686	2811	10366	736	7644	3687	19772	6814
	Partially Proficient	18821	9674	8716	887	377	68	8127	774	7775	460	809	331	7631	502
	Proficient	18002	9271	8716	621	565	44	6280	407	9395	442	450	124	5550	232
	Highly Proficient	7365	4031	3736	177	509	17	1847	81	4859	202	90	41	1387	63
8	Minimally Proficient	35628	16710	18940	2649	354	102	19482	2700	9560	719	7272	2743	18013	5694
	Partially Proficient	16067	8185	7862	883	311	42	7054	685	6541	337	761	252	6717	445
	Proficient	11876	6139	5718	547	354	32	4703	304	5787	264	338	126	3969	196
	Highly Proficient	6287	3069	3216	168	382	26	2015	114	3271	147	169	32	1832	67
Algebra I	Minimally Proficient	37180	16574	19801	2613	408	107	19995	3028	10096	507	6155	2848	17842	2588
	Partially Proficient	15698	8489	7583	856	336	73	7406	757	6514	280	742	412	6053	272
	Proficient	22308	11723	10533	901	961	65	8146	623	11399	400	519	412	6372	180
	Highly Proficient	7436	3638	3792	135	697	16	1851	89	4560	147	74	75	1593	28
Geometry	Minimally Proficient	29378	13904	15458	1977	377	100	15627	2193	8389	406	4149	1914	13576	1564
	Partially Proficient	17197	8913	8268	916	424	44	7661	863	6891	290	745	514	6266	230
	Proficient	20063	10339	9347	659	1013	62	6128	503	10786	336	372	371	5222	81
	Highly Proficient	5732	2496	2876	110	518	13	919	72	3595	127	53	57	783	11
Algebra II	Minimally Proficient	28014	13846	14443	1918	429	104	14343	1956	8843	410	2897	1014	12998	1055
	Partially Proficient	15225	8000	6921	735	390	46	6403	672	6502	234	429	382	4901	136
	Proficient	13398	7077	6319	479	683	35	4098	263	7803	234	179	216	2983	50
	Highly Proficient	4263	1846	2407	96	449	8	768	29	2861	98	36	50	426	6

Exhibit 4.5.2 Percentage of Students At Each Performance Level by Gender, Ethnicity, and Other Demographic Information.

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level													
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
					ELA										
3	Minimally Proficient	45	41	48	56	21	45	56	66	29	36	77	85	56	83
	Partially Proficient	14	14	14	15	13	14	14	14	14	14	8	7	15	8
	Proficient	29	31	27	24	37	32	23	18	37	33	11	6	23	8
	Highly Proficient	12	14	11	6	28	10	6	3	20	17	4	1	6	1
4	Minimally Proficient	40	36	44	53	19	38	51	64	25	31	77	85	51	82
	Partially Proficient	14	14	14	14	10	16	15	14	13	13	8	8	15	8
	Proficient	34	37	32	27	44	35	28	19	43	40	13	7	29	9
	Highly Proficient	12	14	10	6	27	11	6	3	20	16	3	0	6	1
5	Minimally Proficient	32	26	37	43	12	26	40	55	19	25	74	79	41	79
	Partially Proficient	23	23	24	24	16	26	26	25	20	22	14	15	26	15
	Proficient	32	35	29	27	39	32	27	18	40	36	10	5	27	6
	Highly Proficient	13	16	11	6	33	16	7	3	21	18	3	1	7	1
6	Minimally Proficient	41	37	45	53	18	39	53	67	25	33	81	90	53	88
	Partially Proficient	21	22	21	22	17	26	22	19	21	21	10	7	22	8
	Proficient	31	34	29	22	47	28	22	13	43	38	7	3	22	4
	Highly Proficient	6	8	5	3	19	7	2	1	10	8	1	0	2	0
7	Minimally Proficient	38	33	43	47	18	45	48	65	24	31	81	89	49	88
	Partially Proficient	21	22	21	22	16	24	22	20	20	20	10	7	22	8
	Proficient	35	38	32	28	48	27	27	15	46	41	8	4	26	4
	Highly Proficient	6	8	5	3	17	4	3	1	10	8	1	0	2	0
8	Minimally Proficient	44	39	49	55	20	45	54	71	30	38	85	86	56	91
	Partially Proficient	23	24	22	22	19	25	23	18	24	24	9	7	23	6
	Proficient	25	28	23	20	37	24	19	10	34	28	5	6	18	3
	Highly Proficient	7	9	6	4	24	6	4	1	12	9	1	1	3	0
9	Minimally Proficient	37	32	43	50	17	31	47	61	25	30	79	76	50	87
	Partially Proficient	28	29	27	27	22	33	29	26	27	26	14	15	28	9
	Proficient	25	28	23	19	35	27	20	11	34	30	6	8	18	3
	Highly Proficient	9	11	8	5	27	10	5	2	15	13	1	1	4	1
10	Minimally Proficient	54	50	58	66	29	58	65	78	40	46	87	83	68	93
	Partially Proficient	17	18	16	15	14	12	16	12	19	17	7	9	15	4
	Proficient	20	22	18	14	28	20	14	8	27	23	5	7	13	3

Grade	Performance Level	Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
		Percentage of Students in Each Grade and Subject At Each Performance Level													
	Highly Proficient	10	11	8	5	29	10	5	2	15	14	1	2	4	1
11	Minimally Proficient	53	47	58	66	29	59	64	76	40	44	88	80	67	91
	Partially Proficient	17	18	16	15	15	14	16	14	19	16	6	10	15	5
	Proficient	19	22	17	15	25	18	15	9	25	25	4	8	13	3
	Highly Proficient	11	12	9	5	30	9	5	2	16	15	1	2	5	1
Math															
3	Minimally Proficient	25	24	26	38	8	26	32	40	15	21	56	56	33	55
	Partially Proficient	30	31	28	31	16	30	34	36	25	26	24	30	34	29
	Proficient	28	29	28	22	32	30	25	19	34	32	14	12	26	13
	Highly Proficient	17	16	18	9	44	15	10	5	26	21	6	2	8	2
4	Minimally Proficient	29	29	30	44	9	27	38	47	17	23	64	65	39	65
	Partially Proficient	26	27	26	28	15	26	29	30	23	26	19	24	29	23
	Proficient	34	34	33	24	45	37	28	21	43	38	14	10	28	11
	Highly Proficient	10	10	11	4	31	9	5	3	17	13	3	1	5	1
5	Minimally Proficient	27	25	30	40	7	24	35	47	16	22	66	66	36	68
	Partially Proficient	27	28	26	31	15	25	30	29	24	26	20	23	30	22
	Proficient	30	32	29	22	36	37	26	20	37	33	11	9	26	8
	Highly Proficient	15	15	15	7	41	14	9	4	22	19	3	2	8	1
6	Minimally Proficient	37	35	39	53	15	32	47	59	23	31	78	80	49	79
	Partially Proficient	24	25	23	23	16	30	25	23	23	24	13	14	25	14
	Proficient	24	26	23	17	32	22	20	14	31	27	7	5	19	6
	Highly Proficient	14	14	15	6	37	17	8	4	23	18	3	1	7	1
7	Minimally Proficient	46	44	48	62	23	47	56	69	32	40	85	89	57	90
	Partially Proficient	23	24	21	20	20	28	22	19	24	25	9	8	22	7
	Proficient	22	23	21	14	30	18	17	10	29	24	5	3	16	3
	Highly Proficient	9	10	9	4	27	7	5	2	15	11	1	1	4	1
8	Minimally Proficient	51	49	53	63	25	51	58	71	38	49	86	87	59	89
	Partially Proficient	23	24	22	21	22	21	21	18	26	23	9	8	22	7
	Proficient	17	18	16	13	25	16	14	8	23	18	4	4	13	3
	Highly Proficient	9	9	9	4	27	13	6	3	13	10	2	1	6	1
Algebra I	Minimally Proficient	45	41	47	58	17	41	54	68	31	38	83	76	56	84
	Partially Proficient	19	21	18	19	14	28	20	17	20	21	10	11	19	9
	Proficient	27	29	25	20	40	25	22	14	35	30	4	11	20	6

Grade	Performance Level	Percentage of Students in Each Grade and Subject At Each Performance Level													
		Overall	Female	Male	African American	Asian	Alaskan/Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	SPED	LEP	FRL	Accommodation
	Highly Proficient	9	9	9	3	29	6	5	2	14	11	1	2	5	1
Geometry	Minimally Proficient	41	39	43	54	16	45	51	61	28	35	78	67	52	83
	Partially Proficient	24	25	23	25	18	20	25	24	23	25	14	18	24	12
	Proficient	28	29	26	18	43	28	20	14	36	29	7	13	20	4
	Highly Proficient	8	7	8	3	22	6	3	2	12	11	1	2	3	1
Algebra II	Minimally Proficient	46	45	48	60	22	54	56	67	34	42	81	61	61	85
	Partially Proficient	25	26	23	23	20	24	25	23	25	24	12	23	23	11
	Proficient	22	23	21	15	35	18	16	9	30	24	5	13	14	4
	Highly Proficient	7	6	8	3	23	4	3	1	11	10	1	3	2	0

Note: Alaskan = Alaskan Native; Hawaiian = Hawaiian Pacific Islander; SPED = Special Education; LEP = Limited English Proficiency; FRL=Free or Reduced Lunch.

4.6 RELIABILITY

Reliability refers to the consistency or precision of test scores and performance level classifications, and essentially addresses the question of how likely would a student be to achieve the same score, or be classified in the same performance level, across multiple administrations of equivalently constructed and administered test forms. As part of each test administration, the reliability of test scores and performance classifications is evaluated from a variety of perspectives. Evidence of the reliability of AzMERIT ELA and math scores are provided with respect to both classical and IRT indices of internal consistency of test scores, and decision accuracy and consistency of performance level classifications.[25]

Test score reliability is traditionally estimated using both classical and IRT approaches. Classical estimates of test reliability such as coefficient alpha, provide a general index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. The equations and formula for estimating reliability are presented in Appendix E.[26]

[25] Standard 2.2 – The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.
Standard 2.3 – For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

[26] Standard 2.19 – Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.

4.6.1 INTERNAL CONSISTENCY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Classical estimates of test reliability such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. Exhibit 4.6.1.1 shows the Cronbach's alpha internal consistency estimates for each of the AzMERIT ELA and math assessments. Internal consistency estimates are uniformly in the 0.9 range, consistent with most similar length achievement tests.

Exhibit 4.6.1.1 Internal Consistency Reliabilities for AzMERIT Scores

Grade	ELA		MATH	
	Reliability	Variance	Reliability	Variance
G3	0.89	999	0.91	1711
G4	0.89	1099	0.92	1518
G5	0.90	1154	0.92	1620
G6	0.90	1089	0.92	1685
G7	0.89	973	0.91	1263
G8	0.90	1016	0.90	1234
G9E / Algebra I	0.89	909	0.90	1126
G10E / Geometry	0.87	764	0.88	1216
G11E / Algebra II	0.89	927	0.85	1004

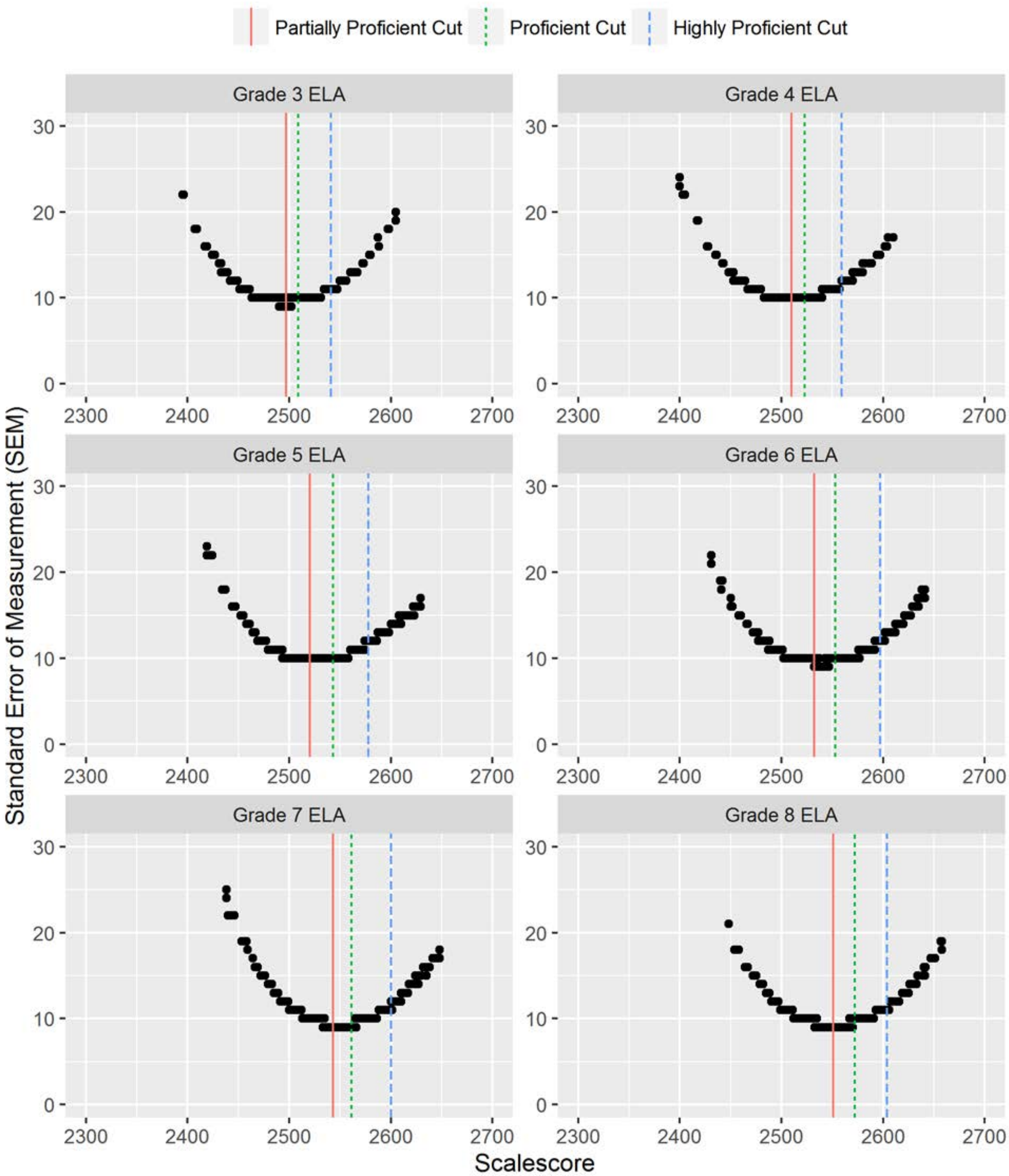
4.6.2 STANDARD ERROR OF MEASUREMENT

Because measurement error is conditional on test information, the precision of test scores varies with respect to the information value of the test at each location along the ability distribution. Precision of individual test scores is critically important to valid test score interpretation. Test scores are most precise in locations where test information is greatest. Because relatively little test information is targeted to measurement of very low and high performing students, the precision of test scores decreases near the tails of the ability distribution.

The figures in Exhibit 4.6.2.1 and Exhibit 4.6.2.2 present graphically the standard errors of measurement for the AzMERIT ELA and math assessments. Each figure also includes the location of the three AzMERIT performance standards. As the figures indicate, the AzMERIT test scores are most precise near the middle of the ability distribution, and especially near the Partially Proficient and Proficient performance standards.[27] Test scores near the tails of the ability distribution are somewhat less precise as expected. An SEM of .3 on the theta metric is consistent with an internal consistency of 0.9. The tables in Appendix I show the mean SEMs for students scoring in each of the performance levels on the AzMERIT reporting scale. While these tables also indicate that the AzMERIT test scores are somewhat more precise for test scores near the middle of the scale, they also show that test scores remain precise even for students in the lowest and highest performance level classifications.

[27] Standard 2.14 – When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Exhibit 4.6.2.1 Overall Standard Error of Measurement for ELA



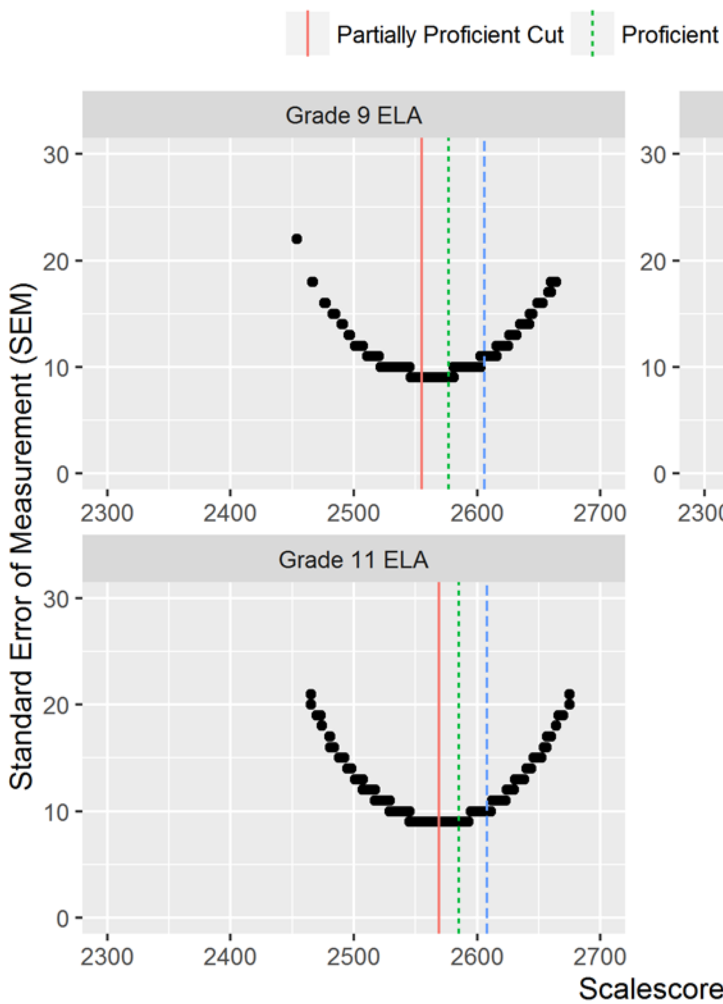
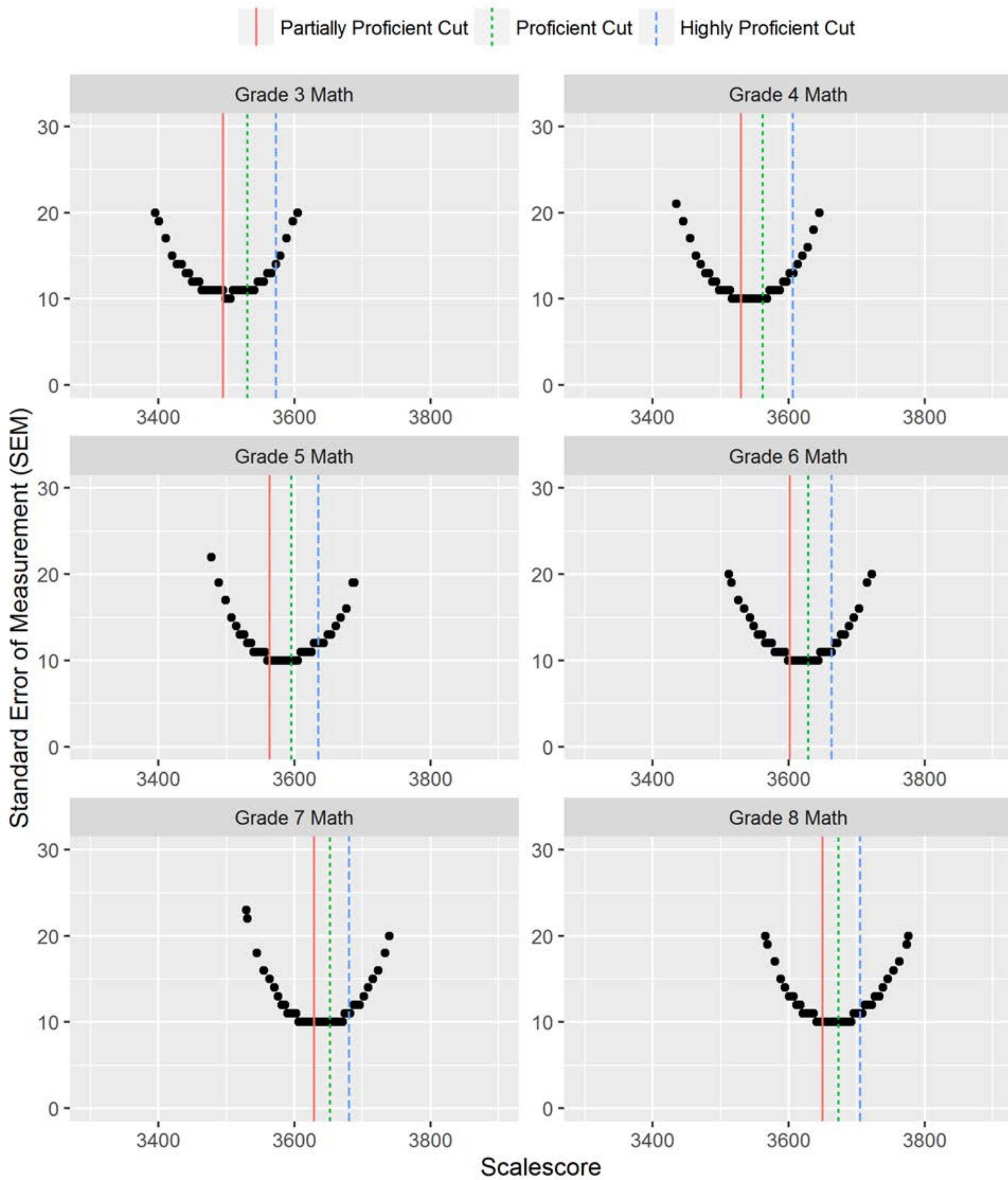
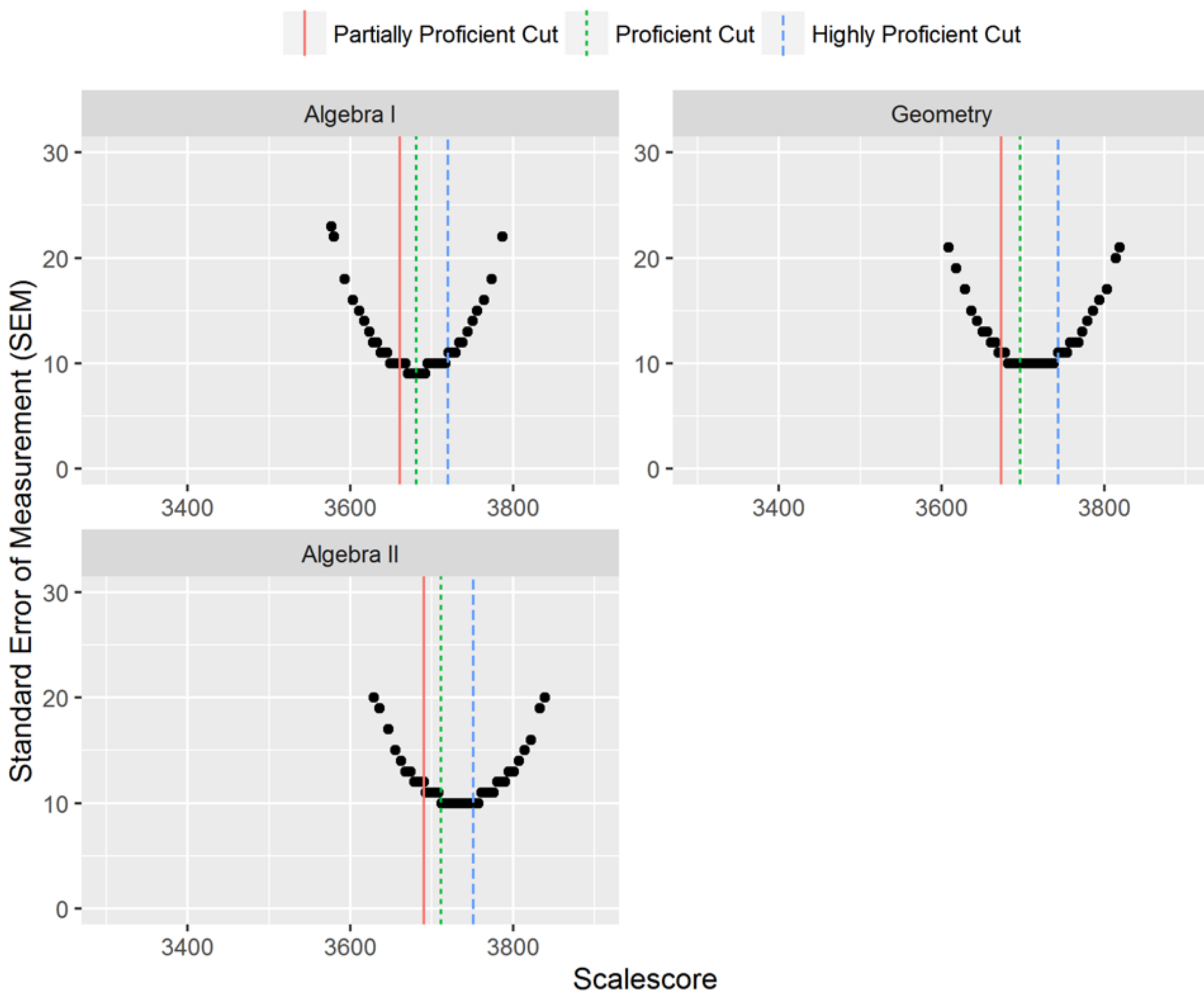

Partially Proficient Cut
Proficient Cut
Highly Proficient Cut
Grade 9 ELA
Grade 10 ELA
Grade 11 ELA
Standard Error of Measurement (SEM)
Scalescore
0
10
20
30
2300
2400
2500
2600
2700

Exhibit 4.6.2.2 Overall Standard Error of Measurement for Math





4.6.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index is computed to estimate the likelihood of consistent classification of students as specified in standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).[28] This index considers the consistency of classifications for the percentage of examinees that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications is typically estimated on test scores based on a single test form from a single test administration using the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995).

The classification index can be examined for classification accuracy and classification consistency. Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications

[28] Standard 2.16 – When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate, equivalently constructed test form—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test administrations.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with measurement error.

4.6.4 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution, we can estimate the above probabilities directly using the likelihood function. The likelihood function of θ given a student's item scores represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's estimated ability (theta) is below the cut score, the probability of *at or above* the cut score is an estimate of the chance that this student is misclassified as below the cut score, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

In Exhibit 4.6.4.1, accurate classifications occur when the decision made on the basis of the true score agrees with the decision made on the basis of the form actually taken. Misclassifications, false positives and false negatives, occur when students' true score classifications are different from students' observed scores (e.g., a student whose true score results in a classification as Proficient, but whose observed score results in an incorrect classification as Partially Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of students falsely above the cut score; N_{00} represents the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

Exhibit 4.6.4.1 Classification Accuracy

		Classification on a Form Actually Taken	
		At or Above the Cut Score	**Below the Cut Score**
Classification on True Score	**At or Above the Cut Score**	N_{11} (Truly above the cut)	N_{10} (False negative)
	Below the Cut Score	N_{01} (False positive)	N_{00} (Truly below the cut)

4.6.5 CLASSIFICATION CONSISTENCY

As shown in Exhibit 4.6.5.1, consistent classification occurs when two forms agree on the classification of a student as either *at and above* or *below* the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit 4.6.5.1 Classification Consistency

		Classification on the 2nd Form Taken	
		Above the Cut Score	**Below the Cut Score**
Classification on the 1st Form Taken	**At or Above the Cut Score**	N_{11} (Consistently above the cut)	N_{10} (Inconsistent)
	Below the Cut Score	N_{01} (Inconsistent)	N_{00} (Consistently below the cut)

4.6.6 CLASSIFICATION ACCURACY AND CONSISTENCY ESTIMATES

Exhibit 4.6.6.1 presents the classification accuracy and consistency indexes for spring 2016 administration of AzMERIT. Exhibit 4.6.6.2 presents the classification accuracy and consistency indexes for each of the identified subgroups: gender (females and males), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with special education, free or reduced lunch, and accommodations). Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency index assumes two test scores, both of which include measurement error, while the accuracy index assumes only a single test score plus the true score, which does not include measurement error.

Exhibit 4.6.6.1 Classification Accuracy and Consistency Estimates for Performance Standards Overall

	Accuracy			Consistency		
Grade	**Partially Proficient**	**Proficient**	**Highly Proficient**	**Partially Proficient**	**Proficient**	**Highly Proficient**
			ELA			
3	0.91	0.91	0.95	0.88	0.88	0.93
4	0.92	0.91	0.95	0.88	0.88	0.93
5	0.92	0.92	0.94	0.89	0.88	0.92
6	0.92	0.92	0.97	0.89	0.89	0.96
7	0.92	0.91	0.96	0.88	0.88	0.95
8	0.92	0.92	0.96	0.88	0.89	0.95
9	0.91	0.91	0.96	0.88	0.88	0.95
10	0.90	0.92	0.96	0.87	0.89	0.95
11	0.91	0.93	0.96	0.88	0.90	0.95
			MATH			
3	0.94	0.92	0.94	0.92	0.89	0.92
4	0.93	0.92	0.95	0.91	0.89	0.94
5	0.93	0.93	0.95	0.91	0.90	0.93
6	0.93	0.93	0.95	0.90	0.90	0.94
7	0.92	0.93	0.96	0.89	0.90	0.95
8	0.91	0.93	0.97	0.88	0.91	0.96
Algebra I	0.91	0.93	0.97	0.87	0.90	0.96
Geometry	0.89	0.93	0.98	0.85	0.90	0.97
Algebra II	0.88	0.93	0.98	0.83	0.90	0.97

Exhibit 4.6.6.2 Classification Accuracy and Consistency Estimates for Performance Standards across Subgroups

		Accuracy			Consistency		
Grade	Subgroup	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
				ELA			
G3E	Overall	0.91	0.91	0.95	0.88	0.88	0.93
	Female	0.91	0.91	0.95	0.87	0.88	0.93
	Male	0.91	0.92	0.96	0.88	0.89	0.94
	African American	0.90	0.92	0.97	0.87	0.89	0.95
	Hispanic/ Latino	0.90	0.92	0.97	0.87	0.89	0.96
	Asian	0.92	0.92	0.91	0.90	0.88	0.88
	White	0.92	0.91	0.93	0.89	0.87	0.90
	Native Hawaiian/Pacific Islander	0.91	0.90	0.96	0.88	0.87	0.94
	American Indian or Alaskan	0.90	0.93	0.98	0.87	0.90	0.97
	Multiple Ethnicities	0.91	0.91	0.94	0.88	0.87	0.91
	Limited English Proficiency	0.94	0.96	0.99	0.91	0.95	0.99
	Special Education	0.94	0.96	0.98	0.92	0.94	0.98
	Free/Reduced Lunch	0.90	0.92	0.97	0.87	0.89	0.96
	Accommodations	0.93	0.96	0.99	0.91	0.94	0.99
G4E	Overall	0.92	0.91	0.95	0.88	0.88	0.93
	Female	0.92	0.91	0.94	0.88	0.88	0.92
	Male	0.92	0.91	0.96	0.88	0.88	0.94
	African American	0.91	0.92	0.97	0.88	0.89	0.96
	Hispanic/ Latino	0.91	0.91	0.97	0.87	0.88	0.96
	Asian	0.94	0.91	0.91	0.92	0.88	0.88
	White	0.93	0.91	0.92	0.90	0.87	0.89
	Native Hawaiian/Pacific Islander	0.91	0.90	0.95	0.87	0.86	0.93
	American Indian or Alaskan	0.90	0.93	0.98	0.86	0.90	0.97
	Multiple Ethnicities	0.92	0.90	0.93	0.88	0.87	0.91
	Limited English Proficiency	0.94	0.97	1.00	0.91	0.95	0.99
	Special Education	0.94	0.96	0.99	0.92	0.94	0.98
	Free/Reduced Lunch	0.91	0.91	0.97	0.87	0.88	0.96
	Accommodations	0.94	0.96	1.00	0.91	0.95	0.99
G5E	Overall	0.92	0.92	0.94	0.89	0.88	0.92
	Female	0.93	0.91	0.93	0.90	0.88	0.91
	Male	0.92	0.92	0.95	0.89	0.89	0.94
	African American	0.91	0.92	0.97	0.88	0.89	0.95
	Hispanic/ Latino	0.91	0.92	0.96	0.88	0.88	0.95
	Asian	0.95	0.92	0.90	0.93	0.89	0.86
	White	0.94	0.91	0.92	0.92	0.87	0.89
	Native Hawaiian/Pacific Islander	0.92	0.91	0.92	0.89	0.88	0.90
	American Indian or Alaskan	0.90	0.93	0.98	0.86	0.91	0.98
	Multiple Ethnicities	0.93	0.91	0.93	0.90	0.88	0.90
	Limited English Proficiency	0.92	0.97	1.00	0.89	0.96	0.99
	Special Education	0.94	0.97	0.99	0.91	0.95	0.98
	Free/Reduced Lunch	0.91	0.92	0.96	0.88	0.89	0.95
	Accommodations	0.93	0.97	1.00	0.90	0.96	0.99
	Overall	0.92	0.92	0.97	0.89	0.89	0.96
	Female	0.92	0.92	0.97	0.89	0.89	0.95
	Male	0.92	0.92	0.98	0.89	0.90	0.97

		Accuracy			Consistency		
Grade	Subgroup	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
G6E	African American	0.91	0.93	0.99	0.88	0.91	0.98
	Hispanic/ Latino	0.91	0.93	0.99	0.87	0.90	0.98
	Asian	0.94	0.92	0.94	0.92	0.89	0.91
	White	0.93	0.91	0.95	0.90	0.87	0.93
	Native Hawaiian/Pacific Islander	0.91	0.91	0.96	0.87	0.88	0.95
	American Indian or Alaskan	0.91	0.95	0.99	0.88	0.93	0.99
	Multiple Ethnicities	0.92	0.91	0.97	0.89	0.88	0.95
	Limited English Proficiency	0.95	0.98	1.00	0.93	0.98	1.00
	Special Education	0.95	0.98	1.00	0.93	0.96	0.99
	Free/Reduced Lunch	0.91	0.93	0.99	0.87	0.90	0.98
	Accommodations	0.95	0.98	1.00	0.92	0.98	1.00
G7E	Overall	0.92	0.91	0.96	0.88	0.88	0.95
	Female	0.91	0.91	0.96	0.88	0.87	0.94
	Male	0.92	0.92	0.97	0.88	0.89	0.96
	African American	0.91	0.91	0.98	0.87	0.88	0.97
	Hispanic/ Latino	0.90	0.92	0.98	0.87	0.89	0.97
	Asian	0.94	0.90	0.94	0.92	0.87	0.91
	White	0.93	0.90	0.94	0.90	0.86	0.92
	Native Hawaiian/Pacific Islander	0.90	0.90	0.98	0.86	0.87	0.97
	American Indian or Alaskan	0.90	0.94	0.99	0.86	0.92	0.99
	Multiple Ethnicities	0.93	0.90	0.96	0.90	0.87	0.94
	Limited English Proficiency	0.95	0.98	1.00	0.93	0.97	1.00
	Special Education	0.94	0.97	1.00	0.91	0.96	0.99
	Free/Reduced Lunch	0.90	0.92	0.98	0.87	0.89	0.97
	Accommodations	0.94	0.98	1.00	0.92	0.97	1.00
G8E	Overall	0.92	0.92	0.96	0.88	0.89	0.95
	Female	0.91	0.92	0.96	0.88	0.88	0.94
	Male	0.92	0.93	0.97	0.89	0.90	0.96
	African American	0.91	0.93	0.98	0.88	0.91	0.97
	Hispanic/ Latino	0.91	0.93	0.98	0.87	0.91	0.97
	Asian	0.94	0.91	0.92	0.91	0.88	0.89
	White	0.92	0.91	0.94	0.89	0.87	0.92
	Native Hawaiian/Pacific Islander	0.93	0.90	0.96	0.89	0.87	0.95
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.99
	Multiple Ethnicities	0.92	0.91	0.96	0.88	0.88	0.94
	Limited English Proficiency	0.96	0.98	0.99	0.94	0.97	0.99
	Special Education	0.96	0.98	1.00	0.94	0.97	0.99
	Free/Reduced Lunch	0.91	0.93	0.98	0.87	0.91	0.97
	Accommodations	0.96	0.99	1.00	0.95	0.98	1.00
G9E	Overall	0.91	0.91	0.96	0.88	0.88	0.95
	Female	0.91	0.91	0.96	0.87	0.87	0.94
	Male	0.91	0.92	0.97	0.88	0.89	0.95
	African American	0.91	0.93	0.98	0.87	0.90	0.97
	Hispanic/ Latino	0.90	0.93	0.98	0.86	0.90	0.97
	Asian	0.93	0.90	0.93	0.91	0.86	0.90
	White	0.92	0.90	0.94	0.89	0.86	0.92
	Native Hawaiian/Pacific Islander	0.90	0.90	0.97	0.87	0.86	0.96

Grade	Subgroup	Accuracy: Partially Proficient	Accuracy: Proficient	Accuracy: Highly Proficient	Consistency: Partially Proficient	Consistency: Proficient	Consistency: Highly Proficient
	American Indian or Alaskan	0.89	0.95	0.99	0.85	0.93	0.98
	Multiple Ethnicities	0.92	0.91	0.96	0.89	0.87	0.94
	Limited English Proficiency	0.93	0.96	0.99	0.90	0.94	0.98
	Special Education	0.94	0.97	0.99	0.91	0.96	0.99
	Free/Reduced Lunch	0.90	0.93	0.98	0.87	0.90	0.97
	Accommodations	0.95	0.98	1.00	0.93	0.97	1.00
G10E	Overall	0.90	0.92	0.96	0.87	0.89	0.95
	Female	0.90	0.92	0.96	0.86	0.89	0.94
	Male	0.91	0.93	0.97	0.88	0.90	0.96
	African American	0.91	0.94	0.98	0.88	0.91	0.97
	Hispanic/ Latino	0.91	0.94	0.98	0.87	0.91	0.97
	Asian	0.91	0.91	0.93	0.88	0.87	0.90
	White	0.90	0.90	0.95	0.86	0.87	0.93
	Native Hawaiian/Pacific Islander	0.91	0.93	0.97	0.88	0.90	0.95
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.99
	Multiple Ethnicities	0.91	0.92	0.95	0.87	0.89	0.94
	Limited English Proficiency	0.94	0.96	0.99	0.91	0.95	0.98
	Special Education	0.95	0.98	0.99	0.93	0.97	0.99
	Free/Reduced Lunch	0.91	0.94	0.98	0.88	0.92	0.97
	Accommodations	0.96	0.99	1.00	0.95	0.98	0.99
G11E	Overall	0.91	0.93	0.96	0.88	0.90	0.95
	Female	0.91	0.92	0.96	0.87	0.89	0.94
	Male	0.92	0.94	0.97	0.89	0.91	0.95
	African American	0.92	0.95	0.98	0.89	0.93	0.97
	Hispanic/ Latino	0.92	0.94	0.98	0.89	0.92	0.97
	Asian	0.92	0.91	0.93	0.89	0.88	0.90
	White	0.91	0.91	0.94	0.87	0.88	0.92
	Native Hawaiian/Pacific Islander	0.93	0.95	0.96	0.91	0.93	0.95
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.99
	Multiple Ethnicities	0.92	0.92	0.95	0.89	0.89	0.93
	Limited English Proficiency	0.94	0.96	0.99	0.91	0.95	0.98
	Special Education	0.96	0.98	0.99	0.95	0.97	0.99
	Free/Reduced Lunch	0.92	0.95	0.98	0.89	0.93	0.97
	Accommodations	0.97	0.98	1.00	0.96	0.98	0.99
MATH							
G3M	Overall	0.94	0.92	0.94	0.92	0.89	0.92
	Female	0.94	0.92	0.94	0.92	0.89	0.92
	Male	0.95	0.93	0.94	0.92	0.90	0.92
	African American	0.93	0.93	0.96	0.91	0.90	0.95
	Hispanic/ Latino	0.93	0.92	0.96	0.90	0.89	0.94
	Asian	0.97	0.94	0.90	0.96	0.91	0.86
	White	0.96	0.92	0.92	0.94	0.89	0.88
	Native Hawaiian/Pacific Islander	0.94	0.92	0.93	0.91	0.89	0.91
	American Indian or Alaskan	0.91	0.93	0.97	0.88	0.90	0.96
	Multiple Ethnicities	0.95	0.92	0.93	0.93	0.89	0.90
	Limited English Proficiency	0.92	0.95	0.99	0.88	0.93	0.98

		Accuracy			Consistency		
Grade	Subgroup	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
	Special Education	0.93	0.96	0.98	0.91	0.94	0.97
	Free/Reduced Lunch	0.93	0.92	0.96	0.90	0.89	0.94
	Accommodations	0.92	0.95	0.99	0.89	0.93	0.98
G4M	Overall	0.93	0.92	0.95	0.91	0.89	0.94
	Female	0.93	0.92	0.96	0.90	0.89	0.94
	Male	0.94	0.92	0.95	0.91	0.90	0.94
	African American	0.92	0.93	0.98	0.89	0.91	0.97
	Hispanic/ Latino	0.92	0.92	0.97	0.89	0.90	0.96
	Asian	0.97	0.93	0.90	0.96	0.90	0.87
	White	0.95	0.92	0.93	0.93	0.89	0.90
	Native Hawaiian/Pacific Islander	0.95	0.91	0.96	0.92	0.88	0.94
	American Indian or Alaskan	0.91	0.93	0.98	0.87	0.91	0.98
	Multiple Ethnicities	0.94	0.92	0.94	0.91	0.89	0.92
	Limited English Proficiency	0.91	0.96	0.99	0.87	0.95	0.99
	Special Education	0.93	0.96	0.99	0.90	0.95	0.98
	Free/Reduced Lunch	0.92	0.92	0.97	0.89	0.90	0.96
	Accommodations	0.92	0.96	0.99	0.88	0.95	0.99
G5M	Overall	0.93	0.93	0.95	0.91	0.90	0.93
	Female	0.93	0.92	0.95	0.90	0.89	0.93
	Male	0.93	0.93	0.95	0.91	0.90	0.93
	African American	0.92	0.93	0.97	0.88	0.91	0.96
	Hispanic/ Latino	0.92	0.93	0.96	0.89	0.90	0.95
	Asian	0.97	0.94	0.92	0.95	0.91	0.88
	White	0.95	0.92	0.93	0.93	0.89	0.91
	Native Hawaiian/Pacific Islander	0.93	0.94	0.93	0.91	0.91	0.91
	American Indian or Alaskan	0.91	0.93	0.98	0.87	0.91	0.97
	Multiple Ethnicities	0.94	0.92	0.94	0.92	0.89	0.92
	Limited English Proficiency	0.91	0.96	0.99	0.88	0.95	0.99
	Special Education	0.93	0.97	0.99	0.90	0.95	0.98
	Free/Reduced Lunch	0.92	0.93	0.97	0.89	0.90	0.95
	Accommodations	0.92	0.97	0.99	0.88	0.95	0.99
G6M	Overall	0.93	0.93	0.95	0.90	0.90	0.94
	Female	0.93	0.93	0.95	0.90	0.90	0.93
	Male	0.93	0.93	0.96	0.90	0.91	0.94
	African American	0.92	0.95	0.97	0.89	0.92	0.96
	Hispanic/ Latino	0.92	0.94	0.97	0.89	0.91	0.96
	Asian	0.96	0.93	0.93	0.94	0.91	0.90
	White	0.94	0.92	0.93	0.92	0.89	0.91
	Native Hawaiian/Pacific Islander	0.92	0.93	0.95	0.89	0.90	0.93
	American Indian or Alaskan	0.91	0.95	0.98	0.88	0.92	0.97
	Multiple Ethnicities	0.94	0.92	0.95	0.91	0.89	0.93
	Limited English Proficiency	0.94	0.97	0.99	0.91	0.96	0.99
	Special Education	0.95	0.97	0.99	0.93	0.96	0.99
	Free/Reduced Lunch	0.92	0.94	0.97	0.89	0.91	0.96
	Accommodations	0.94	0.97	0.99	0.92	0.96	0.99
	Overall	0.92	0.93	0.96	0.89	0.90	0.95
	Female	0.92	0.92	0.96	0.88	0.90	0.95
	Male	0.92	0.93	0.96	0.89	0.91	0.95
	African American	0.92	0.95	0.98	0.88	0.93	0.97

		Accuracy			Consistency		
Grade	Subgroup	Partially Proficient	Proficient	Highly Proficient	Partially Proficient	Proficient	Highly Proficient
G7M	Hispanic/ Latino	0.91	0.94	0.98	0.88	0.92	0.96
	Asian	0.94	0.91	0.94	0.92	0.88	0.91
	White	0.92	0.91	0.94	0.89	0.88	0.92
	Native Hawaiian/Pacific Islander	0.90	0.93	0.98	0.86	0.91	0.97
	American Indian or Alaskan	0.92	0.96	0.99	0.89	0.94	0.98
	Multiple Ethnicities	0.92	0.92	0.96	0.88	0.89	0.94
	Limited English Proficiency	0.95	0.98	1.00	0.93	0.98	0.99
	Special Education	0.96	0.98	0.99	0.94	0.97	0.99
	FRL	0.91	0.94	0.98	0.88	0.92	0.97
	Accommodations	0.96	0.99	1.00	0.94	0.98	0.99
G8M	Overall	0.91	0.93	0.97	0.88	0.91	0.96
	Female	0.91	0.93	0.97	0.87	0.90	0.95
	Male	0.92	0.94	0.97	0.88	0.91	0.96
	African American	0.91	0.95	0.98	0.88	0.93	0.97
	Hispanic/ Latino	0.91	0.94	0.98	0.88	0.92	0.97
	Asian	0.93	0.92	0.93	0.90	0.88	0.91
	White	0.91	0.92	0.95	0.88	0.88	0.94
	Native Hawaiian/Pacific Islander	0.91	0.93	0.95	0.88	0.90	0.94
	American Indian or Alaskan	0.92	0.96	0.99	0.88	0.94	0.98
	Multiple Ethnicities	0.91	0.94	0.96	0.87	0.91	0.95
	Limited English Proficiency	0.95	0.98	0.99	0.93	0.97	0.99
	Special Education	0.95	0.98	0.99	0.93	0.97	0.99
	Free/Reduced Lunch	0.91	0.94	0.98	0.87	0.92	0.97
	Accommodations	0.96	0.99	1.00	0.94	0.98	0.99
Algebra I	Overall	0.91	0.93	0.97	0.87	0.90	0.96
	Female	0.91	0.92	0.97	0.87	0.90	0.96
	Male	0.91	0.94	0.97	0.87	0.91	0.96
	African American	0.90	0.94	0.99	0.86	0.91	0.98
	Hispanic/ Latino	0.90	0.94	0.98	0.86	0.91	0.97
	Asian	0.94	0.93	0.93	0.92	0.90	0.90
	White	0.92	0.92	0.95	0.88	0.89	0.94
	Native Hawaiian/Pacific Islander	0.88	0.91	0.98	0.84	0.88	0.97
	American Indian or Alaskan	0.90	0.95	0.99	0.86	0.94	0.99
	Multiple Ethnicities	0.91	0.92	0.97	0.87	0.89	0.95
	Limited English Proficiency	0.92	0.96	0.99	0.88	0.94	0.98
	Special Education	0.93	0.97	1.00	0.89	0.96	0.99
	Free/Reduced Lunch	0.90	0.94	0.98	0.86	0.92	0.98
	Accommodations	0.93	0.98	1.00	0.90	0.97	0.99
Geometry	Overall	0.89	0.93	0.98	0.85	0.90	0.97
	Female	0.89	0.92	0.98	0.85	0.89	0.97
	Male	0.90	0.93	0.98	0.86	0.91	0.97
	African American	0.88	0.94	0.99	0.84	0.91	0.99
	Hispanic/ Latino	0.88	0.94	0.99	0.83	0.91	0.99
	Asian	0.94	0.92	0.94	0.91	0.89	0.92
	White	0.91	0.92	0.97	0.87	0.89	0.95
	Native Hawaiian/Pacific Islander	0.90	0.93	0.97	0.86	0.91	0.96
	American Indian or Alaskan	0.87	0.94	0.99	0.82	0.92	0.99

Grade	Subgroup	Accuracy Partially Proficient	Accuracy Proficient	Accuracy Highly Proficient	Consistency Partially Proficient	Consistency Proficient	Consistency Highly Proficient
	Multiple Ethnicities	0.89	0.93	0.96	0.85	0.90	0.95
	Limited English Proficiency	0.89	0.95	0.99	0.84	0.93	0.99
	Special Education	0.90	0.97	1.00	0.86	0.96	0.99
	Free/Reduced Lunch	0.88	0.94	0.99	0.83	0.91	0.98
	Accommodations	0.90	0.98	1.00	0.86	0.97	1.00
Algebra II	Overall	0.88	0.93	0.98	0.83	0.90	0.97
	Female	0.87	0.93	0.98	0.83	0.89	0.97
	Male	0.88	0.93	0.98	0.84	0.91	0.97
	African American	0.88	0.94	0.99	0.83	0.92	0.99
	Hispanic/ Latino	0.87	0.94	0.99	0.82	0.91	0.98
	Asian	0.91	0.93	0.95	0.88	0.89	0.93
	White	0.89	0.92	0.97	0.85	0.88	0.96
	Native Hawaiian/Pacific Islander	0.86	0.96	0.99	0.82	0.93	0.98
	American Indian or Alaskan	0.87	0.96	1.00	0.82	0.93	0.99
	Multiple Ethnicities	0.88	0.94	0.97	0.83	0.91	0.96
	Limited English Proficiency	0.88	0.94	0.99	0.83	0.91	0.98
	Special Education	0.90	0.97	1.00	0.86	0.96	0.99
	Free/Reduced Lunch	0.87	0.94	0.99	0.82	0.92	0.99
	Accommodations	0.91	0.98	1.00	0.87	0.97	1.00

4.6.7 RELIABILITY FOR SUB-GROUPS IN THE POPULATION

Exhibit 4.6.7.1 and 4.6.7.2 shows the mean reliability for each of the identified subgroups: gender (females and males), ethnicity (African American, Asian, Native Hawaiian/Pacific Islander, Hispanic/Latino, American Indian or Alaskan, White, Multiple Ethnicities), and special groups (limited English proficient students, and students with IEPs (Special Education)[29] , free or reduced lunch, and accommodations). As the Exhibit indicates, internal consistency reliabilities are generally stable across subgroups, indicating that the AzMERIT assessments measure a common underlying achievement dimension across all subgroups, and that test scores are similarly precise across demographic subgroups. For subgroups where a the reliability coefficients are attenuated, there is a corresponding decrease in the subgroup variance relative to the overall student population, indicating that attenuation of reliability in subgroups is due to restriction of range.

[29] Standard 2.11 – Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

Exhibit 4.6.7.1 Internal Consistency Reliability by Subgroup– ELA

Grade	Statistic	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodations
							ELA								
3	Reliability	0.89	0.89	0.89	0.88	0.89	0.88	0.87	0.83	0.89	0.89	0.86	0.77	0.87	0.80
	Variance	999	982	999	866	1094	917	823	636	997	1023	827	487	817	552
4	Reliability	0.89	0.89	0.89	0.89	0.89	0.89	0.88	0.85	0.88	0.88	0.87	0.78	0.88	0.81
	Variance	1099	1080	1095	1014	1111	1000	948	749	1022	1002	918	532	932	632
5	Reliability	0.90	0.90	0.90	0.88	0.90	0.90	0.88	0.85	0.89	0.89	0.87	0.81	0.89	0.81
	Variance	1154	1111	1150	1064	1114	994	1020	798	1059	1109	909	608	1011	642
6	Reliability	0.90	0.90	0.90	0.88	0.90	0.90	0.88	0.85	0.89	0.90	0.83	0.74	0.88	0.75
	Variance	1089	1080	1073	919	1123	1060	903	712	1018	1057	697	449	896	469
7	Reliability	0.89	0.89	0.90	0.89	0.89	0.88	0.88	0.85	0.88	0.89	0.83	0.76	0.88	0.77
	Variance	973	916	999	898	999	799	844	676	901	933	676	502	823	509
8	Reliability	0.90	0.90	0.90	0.89	0.90	0.90	0.89	0.85	0.89	0.90	0.82	0.86	0.89	0.76
	Variance	1016	964	1028	941	1049	970	883	650	956	966	632	761	858	475
9	Reliability	0.89	0.88	0.89	0.88	0.89	0.88	0.88	0.84	0.89	0.89	0.83	0.87	0.87	0.77
	Variance	909	851	928	834	920	789	789	606	873	897	643	806	780	484
10	Reliability	0.87	0.87	0.88	0.86	0.88	0.89	0.85	0.80	0.87	0.88	0.80	0.84	0.85	0.74
	Variance	764	722	787	678	865	890	630	471	759	802	523	614	631	400
11	Reliability	0.89	0.56	0.90	0.88	0.90	0.89	0.87	0.83	0.90	0.90	0.81	0.86	0.87	0.78
	Variance	927	871	946	798	1003	860	761	563	929	1016	611	733	759	529

Exhibit 4.6.7.2 Internal Consistency Reliability by Subgroup – Math

Grade	Statistic	Overall	Female	Male	African American	Asian	Alaskan/ Hawaiian	Hispanic/ Latino	American Indian	White	Multiple Ethnicities	Special Education	Limited English Proficiency	Free/Reduced Lunch	Accommodations
							Mathematics								
3	Reliability	0.91	0.91	0.92	0.92	0.88	0.91	0.91	0.90	0.89	0.91	0.92	0.90	0.91	0.91
	Variance	1711	1571	1844	1788	1468	1570	1621	1391	1506	1731	1954	1384	1612	1674
4	Reliability	0.92	0.91	0.92	0.91	0.89	0.91	0.91	0.90	0.90	0.91	0.91	0.87	0.91	0.89
	Variance	1518	1419	1613	1431	1405	1440	1345	1169	1385	1382	1466	994	1343	1171

5	Reliability	0.92	0.92	0.92	0.91	0.91	0.91	0.91	0.90	0.91	0.92	0.89	0.87	0.91	0.87
	Variance	1620	1515	1716	1450	1507	1379	1462	1234	1497	1551	1367	1058	1428	1135
6	Reliability	0.92	0.92	0.92	0.91	0.92	0.92	0.91	0.89	0.92	0.92	0.88	0.85	0.91	0.86
	Variance	1685	1564	1792	1493	1788	1735	1490	1257	1546	1623	1373	1068	1485	1186
7	Reliability	0.91	0.90	0.91	0.89	0.91	0.90	0.89	0.86	0.90	0.91	0.83	0.79	0.89	0.79
	Variance	1263	1187	1331	1069	1356	1194	1075	866	1229	1265	819	645	1073	672
8	Reliability	0.90	0.90	0.90	0.88	0.91	0.91	0.89	0.86	0.90	0.90	0.81	0.81	0.88	0.78
	Variance	1234	1170	1288	1072	1422	1328	1138	929	1189	1239	827	790	1101	737
AlgI	Reliability	0.90	0.89	0.90	0.86	0.91	0.86	0.87	0.82	0.91	0.90	0.75	0.84	0.86	0.73
	Variance	1126	1045	1198	822	1339	781	896	665	1198	1064	533	802	861	522
Geo	Reliability	0.88	0.88	0.89	0.83	0.92	0.89	0.83	0.78	0.90	0.89	0.72	0.81	0.83	0.64
	Variance	1216	1140	1290	906	1508	1227	877	705	1324	1215	646	858	898	543
AlgII	Reliability	0.85	0.84	0.86	0.77	0.91	0.82	0.78	0.69	0.87	0.87	0.65	0.80	0.76	0.56
	Variance	1004	926	1084	720	1433	863	734	545	1120	1118	551	810	690	464

4.6.8 SUBSCALE RELIABILITY

Coefficient alpha estimates of internal consistency reliability associated with the subscales for the 2016 operational forms are presented in Exhibit 4.6.8.1-4.6.8.6. As indicated in the Exhibits, subscale reliabilities are generally moderate in magnitude, as expected for subscales of the length observed in AzMERIT.

Exhibit 4.6.8.1 Subscale Reliabilities – ELA Grades 3-11

	Reading Standards for Informational Text	Reading Standards for Literature	Writing & Language
Grade 3	0.74	0.73	0.76
Grade 4	0.76	0.77	0.70
Grade 5	0.80	0.71	0.74
Grade 6	0.82	0.71	0.73
Grade 7	0.77	0.72	0.72
Grade 8	0.79	0.72	0.76
Grade 9	0.75	0.74	0.74
Grade 10	0.74	0.65	0.74
Grade 11	0.78	0.70	0.75

Exhibit 4.6.8.2 Subscale Reliabilities – Math Grades 3-5

	Numbers & Operations-Fractions	Measurement & Data and Geometry	Operations & Algebraic Thinking, and Numbers & Operations-Base Ten
Grade 3	0.64	0.74	0.82
Grade 4	0.70	0.63	0.86
Grade 5	0.76	0.77	0.83

Exhibit 4.6.8.3 Subscale Reliabilities – Math Grades 6 & 7

	Expressions & Equations	The Number System	Ratio and Proportional Relationships	Geometry, and Statistics & Probability
Grade 6	0.78	0.77	0.72	0.58
Grade 7	0.68	0.68	0.67	0.73

Exhibit 4.6.8.4 Subscale Reliabilities – Math Grades 8

	Expressions & Equations	Functions	Geometry	Statistics & Probability and the Number System
Grade 8	0.79	0.59	0.59	0.57

Exhibit 4.6.8.5 Subscale Reliabilities – Algebra I & II

	Algebra	Functions	Statistics
Algebra I	0.80	0.75	0.62
Algebra II	0.67	0.51	0.63

Exhibit 4.6.8.6 Subscale Reliabilities – Geometry

	Circles, Geometric Measurement, and Geometric Properties with Equations	Congruence	Modeling with Geometry	Similarity, Right Triangles & Trigonometry
Geometry	0.58	0.62	0.62	0.67

4.7 SUBSCALE INTERCORRELATIONS

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Exhibits 4.7.1-4.7.6. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability.[30] The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y , r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y.

When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underling construct.

[30] Standard 1.21 – When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.

Exhibit 4.7.1 Subscale Intercorrelations and Reliability Estimates – ELA Grades 3 to 11

Grade	Subscale	Observed Correlation		Disattenuated Correlation	
		Informational Text	Literature	Informational Text	Literature
3	Literature	0.71		0.97	
	Writing & Language	0.64	0.63	0.86	0.84
4	Literature	0.74		0.97	
	Writing & Language	0.64	0.63	0.87	0.86
5	Literature	0.72		0.96	
	Writing & Language	0.68	0.62	0.93	0.86
6	Literature	0.73		0.96	
	Writing & Language	0.67	0.61	0.93	0.85
7	Literature	0.72		0.97	
	Writing & Language	0.66	0.65	0.92	0.91
8	Literature	0.71		0.94	
	Writing & Language	0.68	0.62	0.92	0.83
9	Literature	0.71		0.95	
	Writing & Language	0.65	0.64	0.87	0.86
10	Literature	0.66		0.95	
	Writing & Language	0.63	0.58	0.91	0.83
11	Literature	0.69		0.94	
	Writing & Language	0.65	0.62	0.90	0.85

Exhibit 4.7.2 Subscale Intercorrelations– Math Grade 3 to 5

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		NF	MDG	NF	MDG
3	MDG	0.68		0.99	
	OAT_NBT	0.72	0.77	0.98	0.98
4	MDG	0.69		1.00	
	OAT_NBT	0.77	0.76	0.99	0.99
5	MDG	0.73		0.95	
	OAT_NBT	0.77	0.77	0.96	0.96

Note: NF = Numbers and Operations-Fractions; MDG = Measurement, Data & Geometry; OAT_NBT = Operations and Algebraic Thinking, and Numbers in Base Ten.

Exhibit 4.7.3 Subscale Intercorrelations– Math Grade 6 & 7

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	NS	RP	EE	NS	RP
6	**NS**	0.78			1.00		
	RP	0.78	0.78		1.00	1.00	
	GSP	0.68	0.68	0.68	1.00	1.00	1.00
7	**NS**	0.70			1.00		
	RP	0.72	0.71		1.00	1.00	
	GSP	0.69	0.71	0.71	0.99	1.00	1.00

Note: EE = Expressions and Equations; NS = Number System; RP = Ratio and Proportional Relationships; GSP = Geometry, Statistics and Probability.

Exhibit 4.7.4 Subscale Intercorrelations– Math Grade 8

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		EE	F	G	EE	F	G
8	Functions (F)	0.67			0.99		
	Geometry(G)	0.71	0.59		1.00	1.00	
	SPNS	0.71	0.60	0.62	1.00	1.00	1.00

Note: EE = Expressions and Equations; F = Functions; G = Geometry; SPNS = Statistics and Probability and the Number System.

Exhibit 4.7.5 Subscale Intercorrelations and Reliability Estimates – Algebra I & Algebra II

Grade	Subscale	Observed Correlations		Disattenuated Correlations	
		Algebra	Functions	Algebra	Functions
Algebra I	Functions	0.76		0.98	
	Statistics	0.72	0.72	1.00	1.00
Algebra II	Functions	0.67		1.00	
	Statistics	0.66	0.61	1.00	1.00

Exhibit 4.7.6 Subscale Intercorrelations and Reliability Estimates – Geometry

Grade	Subscale	Observed Correlations			Disattenuated Correlations		
		CGM_GPE	C	MG	CGM_GPE	C	MG
Geometry	Congruence(C)	0.69			1.00		
	Modeling with Geometry (MG)	0.64	0.60		1.00	0.97	
	Similarity, Right Triangles and Trigonometry (SRTT)	0.71	0.70	0.62	1.00	1.00	0.96

Note: CGM_GPE = Circles, Geometric Measurement and Geometric Properties with Equations;

4.8 RATER EFFECTS

For grades in which statistical models were constructed for machine scoring of essay responses, Measurement, Inc. (MI) hand-scored over 4,100 responses, with each response double scored and any discrepant scores routed for a final resolution score. At each grade, students responded to one of six randomly selected writing tasks. Appendix J shows the rater agreement for each of the writing prompts per grade administered on the AzMERIT. The rater agreement reports show percentages of exact agreement (Equal), adjacent scores (Adj. Low or Adj. High) and nonadjacent scores (Low or High). The tables also identify mismatched scores when there is a difference involving nonscorable condition codes (Mismatch CC), or a nonscorable/scorable mix (MM CC/Score). Exhibit 4.8.1 provides a summary of those results, showing the mean exact agreement rate for dimension scores across grades. Generally exact agreement rates ranged from 65%-70%, with little variability across the essay prompts.

Exhibit 4.8.1 Mean Exact Agreement Rates for Online Essay Responses.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 11	
Dimension	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Purpose/Organization	64	2	72	2	70	1	63	2	70	3	65	1
Evidence/Elaboration	63	2	76	2	69	3	66	3	72	3	62	1
Conventions	70	2	66	1	73	1	67	3	73	4	74	2

For all grades, approximately 15% of essay responses for paper-based test administrations were double-scored. As the tables in Appendix J show, agreement rates for scores assigned to hand written responses were higher than achieved for responses made online. One possible reason for this effect is that paper scoring was conducted following scoring of online responses so that by the time of paper scoring, readers were better synced.

5. ITEM DEVELOPMENT & TEST CONSTRUCTION

The AzMERIT assessments are rigorously examined in accordance to the guidelines provided in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). The Elementary and Secondary Education Act (ESEA) legislation also describes the evidence based on these standards that is necessary to validate assessments for their intended purposes.

The AzMERIT assessments were designed to measure student progress toward achievement of the Arizona College and Career Ready Standards (ACCRS). Although the validity of AzMERIT test score interpretations are evaluated along several dimensions, as a criterion-referenced system of tests, the meaning of test scores are critically evaluated by the degree to which test content was aligned with the ACCRS.[31]

Alignment of content standards is achieved through a rigorous test development process that proceeds from the content standards and refers back to those standards in a highly iterative test development process that included the state department of education, test developers, and educator and stakeholder committees. Items used to develop the spring 2016 operational test forms were drawn mainly from the AIRCore pool of items developed to align with the Common Core State Standards. These items were also reviewed by Arizona content experts and educators prior to field-testing in spring 2015 and subsequent operational test administration in spring 2016. Only items that were found to align well with the ACCRS were used. To supplement the AIRCore pool of items, a few previously developed Arizona items that also aligned to the ACCRS were used.

In addition to ensuring that test items are aligned with their intended content standards, each assessment is intended to measure a representative sample of the knowledge and skills identified in the standards. Test blueprints specify the range and depth with which each of the content strands and standards that are covered in each test administration. Thus, the test specification blueprints represent a policy document specifying the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determined how student

[31] Standard 1.11 – When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

achievement of the ACCRS was evaluated, alignment of test blueprints with the content standards was critical. The ELA and math blueprints are also provided as an attachment in Appendix A.

With the desired alignment of test blueprints to ACCRS, alignment of test forms to the learning standards becomes a mechanical, although sometimes difficult, task of developing test forms that meet blueprint. Developing test forms is difficult because test blueprints could be highly complex, specifying not only the range of items and points for each strand and standard, but also cross-cutting criteria such as distribution across item types, depth of knowledge, writing genre, and so on. And in addition to meeting complex blueprint requirements, test developers worked to meet psychometric goals so that alternate test forms measure equivalently across the range of student ability.

In addition to a review-intensive item development process and form construction process that ensures test forms meet complex blueprint specifications, Student Achievement Partners reviewed the AzMERIT English Language Arts and Math tests to determine how well they are aligned to the Arizona College and Career Ready Standards. This review was based on the criteria recently published in Criteria for Procuring and Evaluating High-Quality Assessments (Council of Chief State School Officers, 2014).

5.1 ITEM DEVELOPMENT PROCESS[32]

The content development process for AzMERIT is managed by AIR's Item Tracking System (ITS), which acts as a content development and management tool, item bank, and publication system supporting both paper and online publication. This item development workflow leads items from inception, through a series of content, fairness, graphic, and other reviews to final publication. The system captures the outcomes and rationales at each review and maintains previous drafts of each item. The workflow management ensures that each item receives each review in the designated sequence, and that the review is conducted (or recorded in the case of committee review) by an authorized person. As items travel through Arizona's extensive review process, every version of every item is archived, along with each comment received in any review. Reviewers have immediate access to all older versions, providing version control throughout development.

ITS allows remote Internet access by item writers and reviewers while ensuring security with individualized passwords for all users, limited access for external users, and strong encryption of all information. Upon publication, ITS tracks the item's use on test forms. After items are used, ITS stores the resulting statistics, including exposure statistics, classical item statistics, and statistics based on item response theory (IRT).

The AzMERIT item development process is predicated on a high level of interaction between test developers at AIR and the Arizona Department of Education, as well as with Arizona educators and stakeholders. AIR's ITS manages item content throughout the entire life cycle of an item, from inception, through series of agreed-upon item review levels culminating in operational pool approval. It also manages item content beyond the operational life of the item, including migration of items for use in practice tests or other training materials. ITS ensures that every item follows through the entire sequence of development and provides Arizona and AIR management on-demand reports of the content and status of the inventory of items. Each item is directed through a sequence of reviews and sign-offs by AIR and ADE staff before it is locked for field test or operational administration.

[32] Standard 4.7 – The procedures used to develop, review, and tryout items and to select items from the item pool should be documented.

The ITS is integrated with the item display engine used by the AzMERIT online test delivery system. This feature, combined with a “web approval” process, allows the display of online items to be “locked” well before test forms are constructed and ensures that only approved items are administered to Arizona students.

5.1.1 ITEM WRITING

Test development experts use item specifications to guide the item development process. [33] These item specifications, developed by content experts at AIR and ADE, strategically guide the item development process. They are detailed documents that specify content limits, model tasks, and response types for a particular standard. Item writers use these specifications while developing items to make the best use of the available item types.

The item specifications were developed using a vertical alignment for each standard, wherein the suggested task demands and cognitive complexity of items build upon those of the previous grade level, just as the standards themselves do.

Additionally, the item specifications provide models for item writers. The models include item samples that target different Depth of Knowledge (DOK) and difficulty levels. These item models also annotate the information in order to communicate the intent of the standard and DOK and to clarify for the writer how to manipulate the item difficulty while keeping the cognitive demands the same.

Detailed item specifications include the following:

- Content Limits: This section delineates the specific content measured by the standard and the extent to which the content is different across grade levels. For example, in grade 3, fraction denominators are limited to 2, 3, 4, 6, and 8.
- Acceptable Response Mechanisms: This section identifies the various ways in which students may respond to a prompt—e.g., multiple choice, graphic response, proposition response, equation response, multi-select.
- Depth of Knowledge: The task demands of each standard can be classified as DOK 1, DOK 2, DOK 3 and/or DOK 4.
- Task Demands: In this section, the standards are broken down into specific task demands aligned to the standard. In addition, each task demand is assigned appropriate response mechanism, DOK, and practice clusters specifically relevant to that particular task demand.
- Examples and Sample Items: In this section, sample items are delineated along with their corresponding expected difficulties (easy, medium, and hard.) Notes for modifying the difficulty of each task demand are detailed with suggestions for the item writer. The suggestions for adapting the difficulty based on the task demands are research-based and have been reviewed by both content experts and a cognitive psychologist.

Item writers consistently followed the item specifications during the item development process. During each level of review, items were compared to the item specifications to ensure their alignment to the standard, grade-level appropriateness, and adherence to the content limits set forth in the item specifications.

[33] Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

Within each grade or course, all items are aligned according to DOK. Depth of Knowledge refers to the cognitive complexity of the item and the cognitive demands on the student. Based on work done by Webb (2002), there are four levels of DOK:

- DOK 1—Recall. Students recall basic mathematical ideas, perform basic arithmetic operations using established algorithms, and identify examples of general math principles.
- DOK 2—Skill/Concept. Students apply their basic knowledge (DOK 1) and extend their thinking to problem solve, identify relationships, and draw conclusions.
- DOK 3—Strategic Thinking. Students go beyond basic problem solving (e.g., word problems) to extend their thinking to nonroutine problem solving, hypothesize, and critique arguments or problem solving strategies.
- DOK 4—Extended Thinking. At this highest level, students engage in extended problem-solving activities, which require integration of multiple standards. For example, students may engage in a performance task that includes a common stimulus and four to six associated items related to the stimulus.

Depending upon the subject area and grade or course assessment, the percentage of items and score points aligned to DOK 1, DOK 2, DOK 3, and DOK 4 vary. The percentage of test items aligned to each DOK level for each assessment is indicated in the test specifications document. Although the exact number of items on each form may vary, the test specifications ensure that students are administered a substantial proportion of items that assess higher-order thinking skills.

ELA

ELA item development often begins with development of reading passages. AzMERIT passages represent a variety of genres and topics. AIR's content experts develop informational texts from multiple content areas, such as history, science, and technical subjects. Literary texts represent authentic pieces from multiple genres, including stories, poetry, and drama. The ratio of informational to literary texts increases at each grade band with a greater percentage of informational texts in the upper grades. The AzMERIT utilizes both single passages as well as passage sets in which students are asked to synthesize information across texts.

To ensure that all passages align to the correct grade level and provide sufficient complexity for close analytical reading, test developers adhere to detailed passage specifications. Content experts use passage complexity worksheets—based on the passage specifications—to perform an in-depth analysis of each passage. The passage specifications call for a close examination of both quantitative measures, such as word counts and Lexile readabilities, as well as qualitative measures, such as passage structure and levels of meaning, all of which are defined as important measures of text complexity.

AzMERIT's ELA assessments include extended writing tasks that provide students with meaningful contexts in which to construct their responses. Each writing-prompt presents students with a variety of stimuli (usually at least two to three per task) that serve as a springboard for an informed piece of writing. Students are given research articles, charts and graphs, and narratives to serve as the basis for their written response. Students can then use this information, along with their own reasoning, to formulate an essay that is not only a clear and coherent expression of their own thinking but that is also grounded in research and evidence. Each student is administered a single informative/explanatory or opinion/argumentative writing essay.

Informative/explanatory writing is focused on conveying information accurately. Informative writing seeks to enlighten the reader about processes or procedures, phenomena, states of affairs, and terminology. To produce this kind of writing, students draw from what they already know as well as from primary and secondary sources. Students develop a controlling idea and a primary focus as they relate facts, details, and examples.

Opinion/argumentative prompts ask students to analyze primary and secondary sources, make sound judgments, and present their opinions and arguments in a coherent way that weaves personal opinion with evidence from the texts. The stimuli present opposing points of view about a topic so that students have enough information to take a stand. The stimuli are followed by a prompt that asks students to write an opinion/argumentative essay. The students are required to synthesize information across the passages to write the essay and must cite specific information from the passages to support the ideas they present.

Writing prompts present students with two or three passage stimuli on a single topic from science, technical subjects, or social studies. The reading level of the stimulus does not exceed the easy Lexile range for the grade level to enable the students to attend to the content of the passages and not struggle over unfamiliar language and non-content-related vocabulary. Moreover, this helps ensure students are assessed on their writing skills and not their reading abilities.

The stimulus is followed by a prompt that asks students to write a short essay about the topic. The students are required to synthesize information across the passages to write the essay and must cite specific information from the passages to support their main ideas. For example, the prompt might require students to describe the steps in a process or describe problems that need to be solved.

MATH

Calculators are not allowed for assessments at grades 3–6, while students participating in high school assessments are allowed continual access to specific calculator functions. For the grades 7 and 8 assessments, where calculator usage is allowable for some item types, the test items are grouped into two segments, administered separately to students: calculator and no-calculator. The construct of the items dictate which section they are to be assessed in.

5.1.2 MACHINE-SCORED CONSTRUCTED-RESPONSE ITEM DEVELOPMENT TOOLS

AzMERIT includes a number of machine-scored constructed response (MSCR) items which leverage a sophisticated system that allows for a large variety of item types expecting varied student responses to be developed, and scored efficiently and economically.

Machine-Scored Constructed-Response (MSCR) item development tools put the power of both item and rubric creation into the hands of item writers, and allow reviewers to score possible responses to ensure the rubric is enacted correctly. For example, when administered a graphic-response item, students can respond by drawing, moving, arranging, or selecting graphic regions. The scoring rubric allows for each answer to be scored using scoring logic created by the item writer. Test developers have flexibility in identifying features of student responses to score, which go beyond simple features (e.g., whether the correct object is put in the correct place) but can involve abstraction. For example, if a student is asked to design an experiment, the rubric can discern whether the objects representing the experimental variable actually vary across conditions or cover the range of inquiry, among other capabilities. These concepts are abstracted and many different responses may reflect those abstract features. This ability enables machine rubrics to “justify” the partial credit assigned in terms of the skills that particular response features exemplify.

In addition, throughout the item development and review process, test developers can mimic the many different possible student responses, and review how the rubric is applied to those responses. Test developers can test the scoring rubric and make corrections to the scoring logic at each step.

When creating equation items, test developers have access to the Equation Editor tool. Student responses can be simple numeric responses or complex equations or even sets of equations. This tool allows for multiple answers and the development of multistep items. Test developers can customize the equation palette to show the appropriate functions. Just as the key pad is customizable, the answer spaces are as well. Additional answer spaces can be added as needed by the item writer. The scoring rubric allows for each answer to be scored using scoring logic created by the item writer.

Such tools are integrated into the ITS, providing test developers the power and flexibility to use technology to create sophisticated AzMERIT items.

5.1.3 ITEM TYPES

AzMERIT includes a wide variety of item types that are designed around a broad and growing catalog of response mechanisms. In addition to selected response items, which include traditional multiple choice and more advanced multi-select and two-part items, AzMERIT tests utilize items with the following response mechanisms:

- Graphic Response, which includes any item to which students respond by drawing, moving, arranging, or selecting graphic regions.
- Hot Text, in which students select or rearrange sentences or phrases in a passage.
- Equation Response, in which students respond by entering an equation or number.
- Word Builder, in which students respond by entering a single number or word.
- Proposition Response, in which students respond in one English language sentences or more, which may be scored by our proposition-scoring engine, human scored, or a mixture of both.
- Essay Response, in which the student response is a longer written response.

AzMERIT items use technology to measure deeper knowledge and application of knowledge in a more open ended way and to machine score many such items. All MSCR items administered in AzMERIT are accessible. There could be occasions where it is necessary to sacrifice accessibility for some population to measure a critical standard, but test development staff would need to carefully consider the measurement benefit before developing that item.

Where possible, MSCR items were rendered for administration on paper test forms, using the gridded response field in the scannable answer documents. Where equation and graphic response items could not be rendered to accept a gridded response on paper forms, responses were hand-scored. For other MSCR items that could not readily be rendered for paper test administration, the item was replaced by another item measuring the same content standards.

The graphic-response mechanism supports most of the typical technology-enhanced item types, including sorting, matching, hot-spot, and drag-and-drop. In addition, it supports items where students actually draw a machine-scorable response and respond by constructing complex, open-ended diagrams, as well as many other possibilities. Because they are uniformly derived from a single response mechanism, the manipulations and interactions are consistent across these technology-enhanced item types, eliminating one possible source of construct-irrelevant variance.

Hot-text items are effectively selected-response items, though in some cases the number of potential selections is quite large. These machine-scored items can have multiple correct answers and allow for very flexible student responses.

The equation response mechanism asks students to enter one or more numbers or equations using a palette of symbols. Test developers can specify which symbols are available on an item-by-item basis, or the Department can choose to have the palette remain consistent across all of the items within a grade level.

The availability of tools organized around response mechanisms creates a very flexible capability for test developers to create authentic, challenging tasks.

5.2 ITEM REVIEW

This section describes the multi-step item review process that items travel through from inception, to several rounds of test developer, Department of Education, and educator review, to field testing and final review prior to inclusion on operational test forms.[34] Items used to develop the spring 2016 operational test forms were mainly drawn from the AIRCore pool of items developed to align with the Common Core State Standards. These items were also reviewed for alignment to the Arizona College and Career Ready Standards (ACCRS) by Arizona content experts and educators prior to field-testing in spring 2015 and subsequent operational test administration in spring 2016. In subsequent years, test forms will be constructed using items developed directly with Arizona, meaning ADE and Arizona educator committees and parent/community committees act as reviewers throughout the item development cycle.

The item review procedures used to develop and review AzMERIT test items are designed to ensure item accuracy and alignment with the intended ACCRS. Following a standard item review process, item reviews proceed initially through a series of internal reviews before items are eligible for review by ADE content experts. Most of AIR's content staff members, who are responsible for conducting internal reviews, are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passes through four internal review steps before it is eligible for review by ADE. Those steps include

- Preliminary review, conducted by a group of AIR content area experts
- Content Review 1, performed by an AIR content specialist
- Edit, in which a copyeditor checks the item for correct grammar/usage
- Senior Content Review, by the lead content expert.

At every stage of the item review process, beginning with preliminary review, AIR's test developers analyze each item to ensure that

- The item is well-aligned with the intended content standard
- The item conforms to the item specifications for the target being assessed
- The item is based on a quality idea (i.e. it assesses something worthwhile in a reasonable way);
- The item is properly aligned to a depth of knowledge (DOK) level;
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter, and takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward
- Any accompanying graphic and stimulus materials are actually necessary to answer the question
- The item stem is clear, concise, and succinct, meaning it contains enough information to know what is being asked, is stated positively (and does not rely on negatives such as no, not, none, never, unless absolutely necessary), and it ends with a question
- For selected response items, the set of response options is succinct; parallel in structure, grammar, length, and content; sufficiently distinct from one another; and all plausible, but with only one correct option
- There is no obvious or subtle cluing within the item

[34] Standard 4.8 – The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.

- The score points for constructed-response items are clearly defined; and
- For machine-scored constructed-response (MSCR) items, the items score as intended at each score point in the rubric.

Based on their review of each item, the test developer can accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to the Department for their review. At this stage, items may be further revised based on any edits or changes requested by Department, or rejected outright. Items passing through the Department review level then have to pass through a stakeholder review in which a committee of educators review each item's accuracy, alignment to the intended standard and DOK level, as well as item fairness and language sensitivity. Thus, all items considered for inclusion in the AzMERIT item pools were initially reviewed by an educator committee which checked to ensure that each item and associated stimulus materials was:

- aligned to the Arizona content standards
- appropriate for the grade level
- accurate
- presented online in a way that is clear and appropriate
- free from bias, sensitive issues, controversial language, stereotyping, and statements that reflect negatively on race, ethnicity, gender, culture, region, disability, or other social and economic conditions and characteristics.

Items successfully passing through this committee review process were then presented to a parent/community review committee to ensure that test content met community standards. Items successfully passing through all review levels were then field tested to ensure that the items behaved as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is therefore an important step in constructing valid and equivalent operational test forms.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning analyses).

Items flagged for review based on their statistical performance must pass a three-stage review to be included in the final item pool from which operational forms were created. In the first stage of this review, a team of psychometricians reviewed all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct and there are no other obvious problems with the items.

ADE content staff then re-evaluated flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, ADE determined that certain flagged items must be rejected, or deemed the item eligible for inclusion in operational test administrations.

5.3 FIELD TESTING

To establish a pool of items for constructing future AzMERIT test forms, newly developed test items were embedded in the spring 2015 AzMERIT test forms for field-testing. Embedding field-test items in operational assessments yields item parameter estimates that capture all the contextual effects that contribute to item difficulty in operational test administrations. A number of factors that may influence item difficulty in the context of operational test administrations may be less relevant in stand-alone field-test contexts. For example, in a high-stakes test, such as high school end-of-course (EOC) exams where test performance may impact student grades, students may be motivated to expend greater effort to achieve maximum performance. Conversely, the high-stakes assessments may also be more likely to elicit anxiety in some students, thus impairing their performance on the tests. Even when assessments are low stakes for students, schools often work to convey to students the importance of statewide assessments in ways that are likely not done for independent field tests. While the impact of contextual factors may not be great, embedded field testing ensures that all aspects of the operational testing context influencing item difficulty are incorporated into the resulting item parameter estimates.

Embedded field-testing is especially useful in the context of a pre-equating model for scoring and reporting test results. Because the test administration context remains the same between the embedded field test (EFT) and subsequent operational test administration, item parameter estimates are more stable over time than they may be when obtained through stand-alone field-testing.

A potential drawback of the EFT approach is the increased assessment burden placed on students and schools. For this reason, AzMERIT utilizes EFT designs for purposes of item bank maintenance. Arizona uses AIR's online field-test engine for computer-administered tests, which, when combined with Arizona's large student population, serves to greatly reduce the number of EFT slots necessary to replenish and even grow the item banks for the Arizona assessments.

The field test engine randomly samples field test items for each individual test administration, essentially creating thousands of unique EFT forms. This sampling approach to embedding field-test items results in several important outcomes:[35]

- Reduction in the number of embedded field-test items that each student must respond to and more efficient "spiraling" of items, which reduces clustering of item responses, resulting in more precise parameter estimates
- More generalizable item statistics because they are not based on items appearing in a single position
- A truly representative sample of respondents for each item

The embedded field testing algorithm actually consists of two different algorithms – one for identifying which field test items will be administered to which student (the distribution algorithm), and one for selecting the position on the test for each item administered the student (the positioning algorithm). When a student starts a test, the system randomly selects a pre-determined number of item groups, stopping when it has selected item groups containing at least the minimum number of field test items designated for administration to each student. This randomization ensure that a) each item is seen by a representative sample of Arizona students, and b) every item is as likely as every other item to appear in a class or school, minimizing clustering effects.

[35] Standard 4.9 – When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.

In addition, a fixed block of field test items was also embedded in paper AzMERIT test forms.

Following the spring 2016 test administrations, the free calibration was performed on the operational items on each of the ELA and mathematics tests. Then the free calibrated item parameters were linked back to the 2015 spring scale using the mean-mean equating method. The tables in Appendix K present the linking constant, post-equated parameters and item drifts for each test. The field test item calibration was conducted by anchoring on the post-equated operational item parameters for all of the ELA and mathematics tests. However, only the ELA spring 2016 operational tests were scored using the post-equated item parameters.

5.4 ITEM STATISTICS

Following the close of test administration windows, AIR psychometrics staff worked to analyze field test data in preparation for item data review meetings and promotion of high quality test items to operational item pools.[36] Analysis of field test items includes classical item statistics as well as the IRT item calibrations. Classical item statistics are designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit bias across subsgroups (DIF analyses). The IRT item analyses allow examination of the fit of items to the measurement model and provide the statistical foundation for operational form construction and test scoring and reporting. Items are flagged if analyses indicate resulting values are out of range. Flagged items are reviewed by AIR and ADE psychometric and content staff for possible miskey or scoring errors. Items that pass through AIR and ADE statistical review are accepted for future operational use. Appendix L provides the slide presentation used to train reviewers for item data review. The training is designed to ensure that all reviewers understand how items are evaluated and that they are interpreting item statistics correctly.

5.4.1 CLASSICAL STATISTICS

Classical item analyses ensured that the field test items function as intended with respect to the AzMERIT's underlying scales. AIR's analysis program computed the required item and test statistics for each multiple-choice and constructed-response (CR) item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item. Key statistics that are computed and examined include item difficulty, item discrimination, and distractor analysis.

Items that are either extremely difficult or extremely easy are flagged for review but not necessarily rejected if they align with the test and content specifications. For multiple-choice items, the proportion of examinees in the sample selecting the correct answer (p-values) is computed, as well as those selecting the incorrect responses. For constructed-response items, item difficulty is calculated both as the item's mean score and as the average proportion correct (analogous to p-value and indicating the ratio of an item's mean score divided by the number of points possible). Items are flagged for reviews if the p-value was less than .25 or greater than .95.

The item discrimination index indicated the extent to which each item differentiated between those examinees who possessed the skills being measured and those who do not. In general, the higher the value, the better the

[36] Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

item was able to differentiate between high- and low- achieving students. The discrimination index for multiple-choice items was calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we computed the mean total number correct for student scoring within each of the possible score categories. Items were flagged for subsequent reviews if the biserial correlation for the keyed (correct) response is less than .25.

Distractor analysis for the multiple-choice items was used to identify items that had marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items were flagged for subsequent reviews if the biserial correlation for the distractor response is greater than .05. In addition, items are flagged if the proportion of students responding to a distractor exceeds the proportion selecting the keyed response. Although non-modal response keys are typically observed with difficult items, in combination with poor item discrimination it may indicate a miskeyed item.

5.4.2 IRT STATS

Rasch and Masters' Partial Credit Model are used to estimate the item response theory (IRT) model parameters for dichotomously and polytomously scored items, respectively. The Winsteps output showing the item statistics resulting from the free (unanchored) estimation of parameters for items in the operational tests were reviewed, as well as the Winsteps-generated item and persons maps. Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are flagged if Infit or Outfit values are less than 0.7 or greater than 1.3.

5.4.3 ANALYSIS OF DIFFERENTIAL ITEM FUNCTIONING

Differential item functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because sometimes it is a clue that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in low-income areas are less likely to offer geometry classes, students at those schools might perform more poorly on geometry items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but the curriculum. However, DIF can indicate bias, so all field- tested items were evaluated for DIF, and all items exhibiting DIF were flagged for further examination by AIR and ADE staff to make a final decision about whether the item should be excluded from the pool of potential items given its performance in field testing potential items.

AIR conducts DIF analysis on all field-tested items to detect potential item bias across major ethnic and gender groups. In Arizona, DIF is investigated among the following group comparisons (reference group/ focal group):

- Male/ Female
- Hispanic, Latino or Spanish origin/ Non-Hispanic
- White/ Black, African American, or Negro

- White/ American Indian or Alaskan Native
- White/ Asian
- White/ Native Hawaiian or Other Pacific Islander
- White/ Multiple ethnicities selected
- Special Education
- Limited English Proficiency
- Free or Reduced Lunch
- Accommodations

AIR uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. Because students within a district, school, and classroom are more similar that would be expected in a simple random sample of students statewide, the information provided by students within a school is not independent, so that standard errors based on the assumption of simple random samples are underestimated. We compute design consistent standard errors that reflect the clustered nature of educational systems. While clustering is mitigated through random administration of large numbers of embedded field test items, design effects in student samples are rarely reduced to the level of a simple random sample.

The ability distribution is divided into a configurable number of intervals to compute the Mantel-Haenszel (MH) chi-square DIF statistics. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta for the MC items; the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the CR items.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed below. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, or female), or negative DIF (i.e., –A, –B, or –C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the “C” category for any group. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. DIF classification rules are presented in Exhibit 5.4.3.1. Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used when evaluating DIF classifications for items where focal or modal groups are less than 200 students (Mazor, Clauser, & Hambleton, 1992, Camilli & Shepard, 1994, Muniz, Hambleton, & Xing, 2001, Sireci & Rios, 2013).

Exhibit 5.4.3.1 DIF Classification Rules

Item Type	Category	Rule
Dichotomous Items	C	$MH\chi^2$ is significant and $\lvert \hat{\Delta}_{MH} \rvert \geq 1.5$
	B	$MH\chi^2$ is significant and $\lvert \hat{\Delta}_{MH} \rvert < 1.5$
	A	$MH\chi^2$ is not significant.
Polytomous Items	C	$MH\chi^2$ is significant and $\lvert SMD \rvert / \lvert SD \rvert \geq .25$.
	B	$MH\chi^2$ is significant and $\lvert SMD \rvert / \lvert SD \rvert < .25$.
	A	$MH\chi^2$ is not significant.

5.5 TEST CONSTRUCTION

The process for constructing fixed-form operational tests begins after field testing and review of item performance. Once an operational item pool is established, AIR content specialists begin the process of constructing test forms. Operational passages and items qualified for operational forms are those that met all of the criteria established by the Department in terms of content, fairness review, and data characteristics.

5.5.1 OPERATIONAL FORM CONSTRUCTION

Each AzMERIT form is built to exactly match the detailed test blueprint, and match the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the depth of knowledge with which it covered, the type of items that measure the constructs, and every other content-relevant aspect of the test. The statistical targets ensure that students receive scores of similar precision, regardless of which form of the test they receive.[37]

AIR's test developers used the FormBuilder software to help construct operational forms. FormBuilder interfaces with AIR's Item Tracking System (ITS) to extract test information and interactively create test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement Curves (SEMCs) as test developers built a test map. This helps content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism.

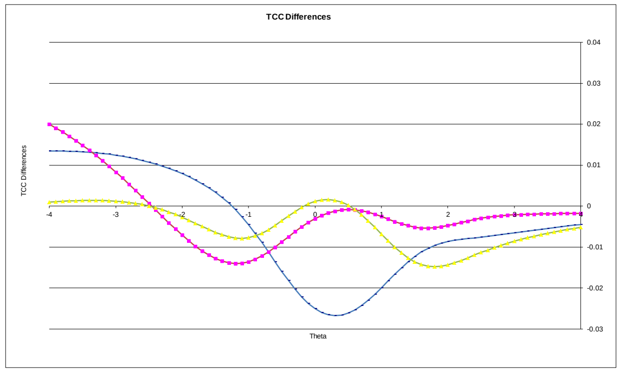
Immediately upon generation of a test form, the FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint were satisfied. In addition, the FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. The summary report also flags items with low biserial correlations, as well as very easy and very difficult items. Although items in the operational pool have passed through data review, construction of fixed form assessments allow another opportunity to ensure that poorly performing items are not included in operational test forms.

The FormBuilder also plotted the distribution of item difficulties, both classical and IRT indices, to both flag extremely easy or difficult items and to ensure that the distribution of item difficulties was consistent across test forms.

As test developers built forms, FormBuilder generated TCCs and SEMCs were plotted using a different color trace line for each prototype form. At this point, the test developer can see the exact difficulty relationship between the target and reference forms. Exhibit 5.5.1.1 shows a sample graph of TCC differences. There are several important things to note when examining TCC differences. First, differences in TCCs can occur at specific locations in the TCCs are across a range of abilities. These differences reflect different emphases in test information across forms at these ability levels. If the difficulty and error structure for the target forms is virtually identical to the reference form, as in the sample TCC and SEM curves, then the item selection process concludes with multiple, parallel test forms. Once the goal of parallel forms is achieved, the information is entered into ITS, which tracks item usage and generates bookmaps (test maps) for use in scoring, forms development, and other processes.

[37] Standard 4.12 – Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

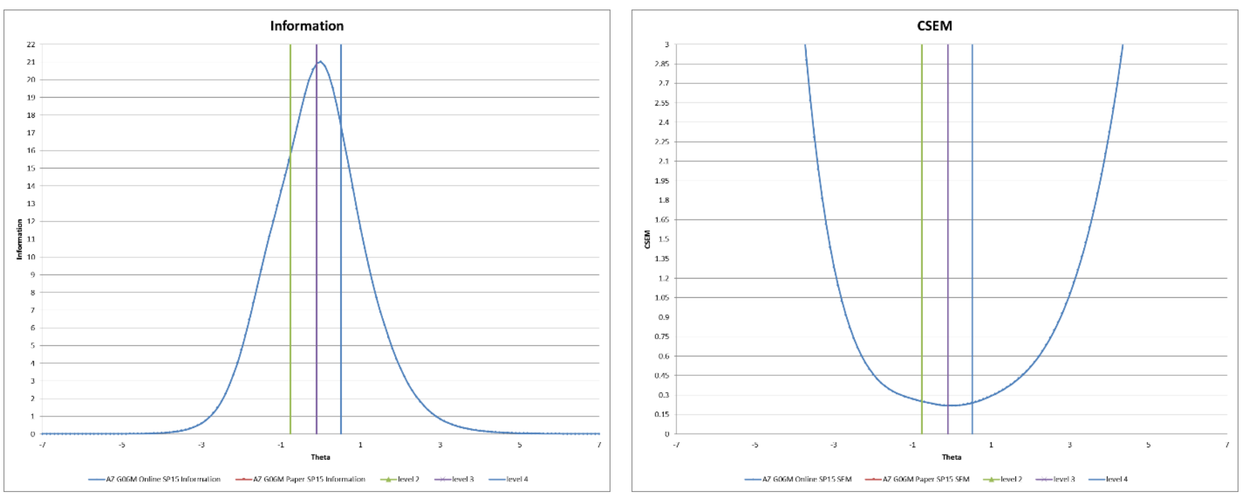
Exhibit 5.5.1.1 Test Characteristics Curve Differences



The reference form for each assessment is the operational test form administered in spring 2015. As illustrated in Exhibit 5.5.1.2, by evaluating test characteristics in reference to the base year forms, students are administered tests each year that are equivalent in difficulty across the range of ability. The Test Characteristic Curve (TCC) and SEM graphs that were used to evaluate the spring 2016 operational test forms are presented in Appendix M.

In addition, although paper test forms were developed to be as nearly identical to the online forms, there were some items that could not readily be rendered for paper test administration. In those instances, replacement items were identified and TCCs and SEMCs were evaluated to ensure equivalence between online and paper test forms.

Exhibit 5.5.1.2 Test Information and Standard Errors Relative to Performance Standards



5.5.2 ASSEMBLING TEST FORMS

The mechanical features of a test—arrangement, directions and production—are just as important as the quality of the items. Many factors directly affect a student's ability to demonstrate proficiency on the assessment, while others relate to the ability to score the assessment accurately and efficiently. Still others affect the inferences made from the test results.

When the test developer reviews a test form for content, in addition to making sure all the benchmark/indicator item requirements are met, he or she also makes sure that the items on the form do not cue each other – that one item does not present material that indicates the answer to another item. This is important to ensure that a student's response on any particular test item is unaffected by, and is statistically independent of, a response to any other test item. This is called "local independence." Independence is most commonly violated when there is a hint in one item about the answer to another item. In that case, a student's true ability on the second item is not being assessed.

Test Developers begin the form construction process by first identifying the pool of items from which forms are built. This pool of items resides at a locked operational status in the Item Tracking System. Each item contains a historical record that clearly demonstrates it has survived the full review process from internal development through client, committee, and statistical data review.

Upon identifying and reviewing the eligible pool of items, a test developer then considers the limitations of the pool, if any. For example, there might be a shortage of high depth of knowledge (DOK 3) items at a particular benchmark. The test developer will review and select from among these items first to ensure that the constraints of the blueprint are met.

Once the items and passages for the form are selected and matched against the blueprint, the test developer reviews the form for a variety of additional content considerations, including the following:

- The items are sequentially ordered.
- Each item of the same type is presented in a consistent manner.
- The listing of the options for the multiple-choice items is consistent.
- The answer options are lettered with A, B, C, and D.
- All graphics are consistently presented.
- All tables and charts have titles and are consistently formatted.
- The number of the answer choice letters is approximately equal across the form.
- The answer key was checked by the initial reviewer and one additional independent reviewer.
- All stimuli have items associated with them.
- The topics of items, passages or stimuli are not too similar to one another.
- There are no errors in spelling, grammar or accuracy of graphics.
- The wording, layout and appearance of the item matches how the item was field-tested.
- There is gender and ethnic balance.
- The passage sets do not start with or end with a constructed response item.
- Each item and the form are checked against the appropriate style guide.
- The directions are consistent across items and were accurate.
- All copyrighted materials have up-to-date permissions agreements.
- Word counts are within documented ranges.

After completing the initial build of the form, the test developer hands it off to another content specialist, who conducts a final review of the criteria listed above. If the test specialist reviewer finds any issues, the form is sent back for revisions. If the form meets blueprint and complies with all specified criteria, the test developer sends it to the psychometric team for review. When the psychometric team approves the form, the test developer forwards the form evaluation workbooks to ADE for review and approval.

6. TEST ADMINISTRATION

6.1 ELIGIBILITY

Arizona public school students in Grade 3 and above were required to participate in AzMERIT testing.[38] Additionally, any student enrolled in a private school or Bureau of Indian Education school and any students who are home schooled had the option to participate as well. Students enrolled in Grades 3 – 8 took English Language Arts (ELA) and Math at the grade level in which they were enrolled. Students, in any grade, who are enrolled in high school level English language arts courses (Freshman English, Sophomore English, Junior English, or their equivalents) or high school level math courses (Algebra I, Geometry, Algebra II, or their equivalents) took the respective End-of-Course (EOC) test.

Students with significant cognitive disabilities and whose current Individualized Education Program (IEP) designates them eligible for the alternate assessment for ELA and Math were excluded from AzMERIT and instead took the Multi-State Alternate Assessment (MSAA).

6.2 ADMINISTRATION PROCEDURES

Key personnel involved with AzMERIT administration include the District Test Coordinators, School Test Coordinators, and Test Administrators who proctor the test. For information about the roles and responsibilities of testing staff, see below.

A secure browser developed by AIR was required to access the computer-based AzMERIT tests. The secure browser provided a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to desktop functionalities, such as the Internet and email. Other measures that protect the integrity and security of the online test are presented in “Test Security Procedures” below.

Prior to each test administration, statewide District Test Coordinator training sessions were conducted to provide information regarding both the paper and computer-based test administrations. The training also provided an overview of the Test Delivery System (TDS), Online Reporting System (ORS), and Test Information Distribution Engine (TIDE). Recorded training sessions and narrated training videos were posted online. The Test Coordinator Manual and Test Administration Directions were shipped to every testing district. Additionally, test administrators were required to complete the online TA Certification Course before administering a computer-based test.[39] District Test Coordinators and School Test Coordinators were responsible ensuring that all test administration personnel (paper and computer-based) were properly trained using the various resources prior to the start of testing.

---

[38] Standard 7.2 – The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

[39] Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.
Standard 12.16 – Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

Manuals and guides on test administrations are available on the AzMERIT Portal.[40] The Test Administrator User Guide was designed to familiarize Test Administrators with the Test Delivery System and contained tips and screenshots throughout the text. The guide provides enough how-to information to enable TAs to access and navigate the Test Delivery System. The user guide provides the following information:

- Steps to take prior to accessing the system and logging in
- Navigating the TA interface application
- The Student Interface, used by students for computer-based testing
- Training sites available for Test Administrators and students
- Secure browsers and keyboard shortcut keys

The *AzMERIT Test Coordinator's Manual* provides information about policies and procedures for AzMERIT Test Coordinators. This manual is updated prior to each test administration and includes test administration policies and guidance for Test Coordinators before, during, and after the window.

The *AzMERIT Test Administration Directions, End-of-Course* and the *AzMERIT Test Administration Directions, Grades 3-8* provide information about policies and procedures for the AzMERIT, both computer-based and paper-based versions. The *Test Administration Directions*, which is updated prior to each test administration, includes test administration information, guidance, and directions.

The *AzMERIT Test Administration Directions* provide easy-to-follow instructions for the online testing environment, such as creating online testing sessions, monitoring online sessions, verifying student information, assigning test accommodations, starting and pausing test sessions.[41] Similar guidance is provided for the paper testing environment, including instructions for the paper testing session, monitoring sessions, verifying student information, and assigning test accommodations. Additional instructions for administering tests to students using Braille accommodated test booklets are provided in the *Supplemental Instructions for Braille* documents.

District and school personnel involved with AzMERIT test administration played an important role in ensuring the validity of the assessment by maintaining both standardized administration conditions and test security.

District Test Coordinators were responsible for coordinating testing at the district level. District Test Coordinators were ultimately accountable for ensuring that testing was conducted in accordance with the test security and other policies and procedures established by the Arizona Department of Education. They ensured that the Test Administrators in each school were appropriately trained and aware of policies and procedures, and that they were trained to use the reporting system.

Districts may also identify School Test Coordinators. School Test Coordinators may assist in the identification and training of Test Administrators. They may also create testing schedules and procedures for the school. If the school administered AzMERIT online, the School Test Coordinators may work with Technology Coordinators to ensure that the necessary secure browsers were installed and any other technical issues were resolved. During the testing

[40] Standard 7.13 – Supporting documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to the appropriate people in a timely manner.
[41] Standard 4.15 – The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.

window, School Test Coordinators needed to monitor testing progress, ensure that all students participate as appropriate, and handle testing incidents as necessary.

Test Administrators (TA) were responsible for reviewing necessary manuals and user guides to prepare the testing environment and ensuring that students did not have unapproved books, notes, or electronic devices out during testing. TAs were required to administer AzMERIT tests following the directions found in the *AzMERIT Test Administration Directions.* [42] Any deviation in test administration must be reported by TAs to the School Test Coordinator, who reports it to the District Test Coordinator. The District Test Coordinator then reports it to ADE.

Test Administrators who administered computer-based AzMERIT tests conducted a training test session using the AzMERIT Sample Tests. Test Administrators were required to pass a qualifying test before they were eligible to administer the AzMERIT online. [43]

Test Administrators must also ensure that only resources that were allowed for specific tests were available and no additional resources were being used during the test. No calculators were permitted in AzMERIT Math tests for grades 3-6. Scientific calculators were permitted in AzMERIT Math Part 1 for grades 7 and 8. Graphing calculators were permitted in AzMERIT Math EOC Parts 1 and 2 (Algebra I, Geometry, and Algebra II). Online calculators were provided as embedded tools within the appropriate computer-based test parts. Handheld calculators could be provided to students during the appropriate test sessions. Calculator guidance was provided in both the *AzMERIT Test Coordinator's Manual* and the *AzMERIT Test Administration Directions*. The online calculators were made publicly available on the AzMERIT Portal, as well as made securely available in a secure browser for paper-based test students to access, if needed. Providing a calculator with prohibited functionality or in the incorrect test session is cause for test invalidation.

For the computer-based ELA Reading tests, headphones or earbuds were required. There were no technical specifications for headphones or earbuds. The equipment was to be checked to ensure it worked with the computer or device the students would use for the assessment prior to the first day of testing. A sound test was also built in to the computer-based assessment and students were asked to verify that headphones and earbuds were working prior to entering the test.

For the paper-based AzMERIT tests, Test Administrators needed to ensure that students used No. 2 pencils to record their responses. School Test Coordinators provided TAs with the materials needed to administer each test session. Secure materials were delivered or picked up immediately before the beginning of each test session. During math testing and when responding to the writing prompt, students were permitted to use the scratch paper as a workspace. After testing, TAs needed to return the testing materials to the School Test Coordinator.

The School Test Coordinator and Test Administrators worked together to determine the most appropriate testing option(s) and testing environment and the average time needed to complete each test. The appropriate protocols were established to maintain a quiet testing environment throughout the testing session. TAs also needed to ensure that adequate time was available to start computers, load secure browsers, and log in students for computer-based tests and pass out and collect test booklets and materials for paper-based tests.

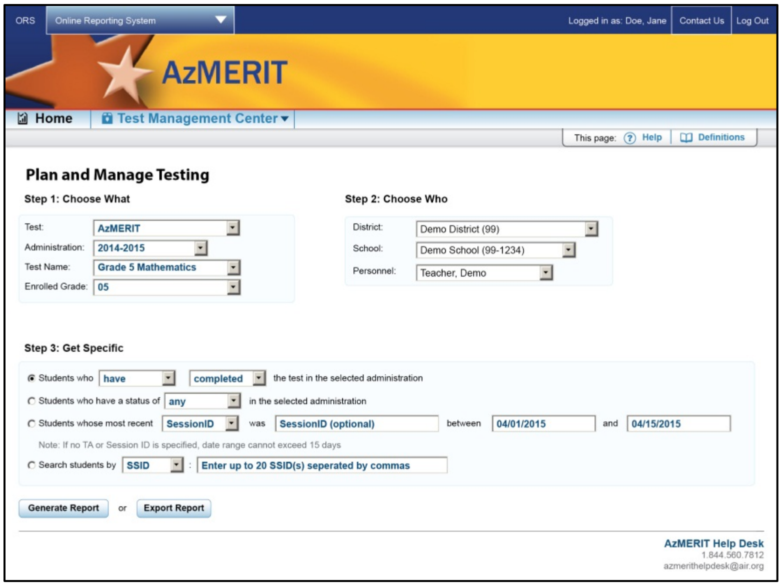
---

[42] Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

[43] Standard 12.16 – Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

MANAGING TESTING

To help schools manage their test schedule, allocate testing resources, and prioritize testing, the AzMERIT online reporting system, which is described in detail in Chapter 6, offered participation reports for online testers. Within the online reporting system, educators can generate up-to-the-minute reports showing students' test status. In addition, users can set testing schedules, monitor testing progress across schools, and track students' participation based on their performance on previous tests.



6.3 TESTING CONDITIONS, TOOLS, AND ACCOMMODATIONS

This section summarizes the testing conditions, tools, and accommodations that are available to AzMERIT testers, as described in the *Testing Conditions, Tools, and Accommodations Guidance* manual that is available each administration. Test tools and accommodation requirements are designed to ensure that test content is accessible for all students.

6.3.1 UNIVERSAL TEST ADMINISTRATION CONDITIONS

Test administrators are required to provide students with an appropriate testing location that is comfortable and free from distractions. Universal test administration conditions are specific testing situations and environments that may be offered to any student in order to provide a more comfortable and distraction-free testing environment.[44] Universal test administration conditions are available for both paper-based test (PBT) and computer-based testing (CBT) modes. Universal test administration conditions include:

[44] Standard 3.4 – Test takers should receive comparable treatment during the test administration and scoring process.
Standard 4.5 – If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.

- Testing in a small group, testing one-on-one, testing in a separate location or in a study carrel,
- Being seated in a specific location within the testing room or being seated at special furniture,
- Having the test administered by a familiar test administrator,
- Using a special pencil or pencil grip,
- Using a place holder,
- Using devices that allow the student to see the test: glasses, contacts, magnification, and special lighting,
- Using different color choices or reverse contrast (for CBT) or color overlays (for PBT),
- Using devices that allow the student to hear the test directions: hearing aids and amplification,
- Wearing noise buffers after the scripted directions have been read,
- Signing the scripted directions,
- Having the scripted directions repeated (at student request),
- Having questions about the scripted directions or the directions that students read on their own answered,
- Reading the test quietly to himself/herself as long as other students are not disrupted, and
- Extended time. (Testing session must be competed in the same school day it was started. No student is expected to need more than twice the estimated testing time.)

While some of the items listed as universal test administration conditions might be included in a student's individualized education plan as an accommodation, for AzMERIT testing purposes these are not considered testing accommodations and are available to any student who needs them not just to students with IEPs.

Standard 6.4 – The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

6.3.2 UNIVERSAL TESTING TOOLS FOR COMPUTER BASED TESTERS

The AzMERIT computer-based testing platform offers numerous testing tools. All tools are available in the AzMERIT Sample Tests, which are available to test administrators and students prior to each test administration. Test administrators are encouraged to ensure that students who will participate in the computer-based AzMERIT take the AzMERIT Sample Tests and familiarize themselves with the available tools.

Exhibit 6.3.2.1 summarizes the Universal Test Tools are available to all students in all AzMERIT tests; these features cannot be disabled by test administrators.

Exhibit 6.3.2.1 Universal Testing Tools for CBT Available to All Students

Universal Test Tool	Description
Area Boundaries	Allows student to click anywhere on the selected response text or button for multiple choice options.
Expand/Collapse Passage	Expand a passage for easier readability. Expanded passages can also be collapsed.
Help	View the on-screen *Test Instructions and Help*.
Highlighter	Highlight text in a passage or item.
Line Reader	Allows student to track the line he or she is reading.
Mark (Flag) for Review	Mark an item for review so that it can be easily found later.
Notes/Comments	Allows student to open an on-screen notepad and take notes or make comments. In ELA, notes are available globally and available throughout the session. In math, comments are attached to a specific test item and available throughout the session.
Pause and Restart	Allows the session to be paused at any time and restarted and taken over a one day period. For test security purposes, visibility on past items is not allowed when paused longer than 20 minutes.
Review Test	Allows student to review the test before ending it.
Strikethrough	Cross out answer options for multiple-choice and multi-select items.
System Settings	Adjust audio (volume) during the test.
Text-to-Speech for Instructions	Listen to test instructions.
Tutorial	View a short video about each item type and how to respond.
Writing Tools	Editing tools (cut, copy, and paste) and basic text formatting tools (bold, underline, and italic) for extended response items.
Zoom In/Zoom Out	Enlarge the font and images in the test. Undo zoom in and return the font and images in the test to original size.

6.3.3 SUBJECT AREA TOOLS FOR CBT AND PBT

AzMERIT testing requires specific subject area tools or resources for certain portions of AzMERIT. The required tools are described in Exhibit 6.3.3.1.

Exhibit 6.3.3.1 Subject Area Tools/Resources Available to All Students

Tool	Applicable Subject Area	Description of Tool
Dictionary/Thesaurus	Writing	CBT – Students have access to the dictionary/thesaurus tool. Students may opt to use a published, paper dictionary or thesaurus instead of using this tool. PBT – Schools must make published, paper dictionaries and thesauruses available to students. Students with a visual impairment may use an electronic dictionary and thesaurus with other features turned-off.
Writing Guide	Writing	CBT – Students have access to the writing guide tool. PBT – The writing guide is included within the test booklet.
Scratch Paper	Writing and Math	CBT – Schools must provide scratch paper (plain, lined, or graph) to students PBT – Schools must provide scratch paper (plain, lined, or graph) to students
Calculator Grades 7-8 (Part 1 only): scientific calculators are acceptable EOC (entire test): graphing calculators are acceptable	Math	CBT – Students have access to the calculator tool when calculator use is permitted. Students may opt to use an acceptable handheld calculator instead of this tool when calculator use is permitted. PBT – Students may use an acceptable handheld calculator when calculator use is permitted. Schools should provide students with an appropriate handheld calculator.

6.3.4 ACCOMMODATIONS

Accommodations are provisions made in how a student accesses and demonstrates learning that do not substantially change the instructional level, the content, or the performance criteria. Accommodations can be changes in the presentation, response, setting, and timing/scheduling of educational activities. Testing accommodations provide more equitable access during assessment but do not alter the validity of the assessment, score interpretation, reliability, or security of the assessment. For a student with disabilities, accommodations are intended to reduce or even eliminate the effects of the student's disability. For an English Language Learner or a Fluent English Proficient Year 1 or Year 2 student, accommodations are intended to allow the student the opportunity to demonstrate content knowledge even though the student may not be functioning at grade level in English.

Research indicates that more accommodations are not necessarily better. Providing students with accommodations that are not truly needed may have a negative effect on performance. There should be a direct connection between a student's disability, special education need, or language need and the accommodation(s) provided to the student during educational activities, including assessment. Test administrators are instructed to make accommodation decisions based on individual needs, and to select accommodations that reduce the effect of the disability or limited English proficiency. Selected accommodations should be provided routinely for classroom instruction and classroom assessment during the school year in order to be used for standardized assessments. Therefore, no accommodation may be put in place for an AzMERIT test that is not already used regularly in the classroom.

Testing accommodations may not violate the construct of a test item. Testing accommodations may not provide verbal or other clues or suggestions that hint at or give away the correct response to the student. Therefore, it is not permissible to simplify, paraphrase, explain, or eliminate any test item, writing prompt, or answer option. The accommodations available to students while testing on AzMERIT are generally limited to those listed in *AzMERIT Testing Conditions, Tools and Accommodations Guidance* manual, and summarized in this section.[45] Arizona takes care to ensure allowable testing accommodations do not alter the validity, score interpretation, reliability, or security of AzMERIT. If a student's individualized education plan calls for a testing accommodation that is not listed, test administrators are instructed to contact ADE for guidance.

Allowable accommodations are described below.[46]

ACCOMMODATIONS FOR STUDENTS WITH AN INJURY

Students with an injury, such as a broken hand or arm, that would make it difficult to participate in AzMERIT may use, as appropriate, any of the universal test administration conditions and any of the following accommodations described in Exhibit 6.3.4.1. There are no specific CBT tools to support these accommodations.

Exhibit 6.3.4.1 Accomodations for Students with an Injury

Accommodation	Description
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
Assistive Technology	Use of assistive technology for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.
Rest/Breaks	Student may take breaks during testing sessions to rest.

[45] Standard 3.10 – When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.
[46] Standard 3.9 – Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

ACCOMMODATIONS FOR ENGLISH LANGUAGE LEARNER (ELL) AND FEP STUDENTS

Students who are not proficient in English, as determined by the Arizona English Language Learner Assessment (AZELLA), may use, as appropriate, any of the universal test administration conditions and any of the following accommodations. This includes English Language Learner (ELL) students and students withdrawn from English language services at parent request. Reclassified Fluent English Proficient (FEP) students are monitored for two school years. These FEP Year 1 and FEP Year 2 students also may use, as appropriate, any of the universal test administration conditions and any of the following accommodations.

The *upon student request* accommodations are required to be administered in a setting that does not disturb other students such as in a one-on-one or very small group setting.

Exhibit 6.3.4.2 summarizes accommodations that may be provided for ELL and FEP students.

Exhibit 6.3.4.2 Allowable Accommodations for ELL and FEP Students

Accommodation	Description of Use
Read Aloud Test Content	CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test. PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test upon student request. Reading aloud the content of the Reading portion of the ELA test is prohibited.
Rest/Breaks	Student may take breaks during testing sessions to rest.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own upon student request.
Translate Directions	Exact oral translation, in the student's native language, of the scripted directions or the directions that students read on their own upon student request. Translations that paraphrase, simplify, or clarify directions are not permitted. Written translations are not permitted. Translation of test content is not permitted.
Translation Dictionary	Provide a word-for-word published, paper translation dictionary. Students with a visual impairment may use an electronic word-for-word translation dictionary with other features turned-off.

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

Students with disabilities may use any of the universal test administration conditions and any of the accommodations described in Exhibit 6.3.4.3, as designated in their IEP or 504 plan.

Exhibit 6.3.4.3 Allowable Accommodations for Students with Disabilities

Accommodation	Description of Use
Abacus	Students with a visual impairment may use an abacus without restrictions for any AzMERIT math test.
Adult Transcription	An adult marks selected response items on CBT test form or PBT test booklet based on student answers provided orally or using gestures. An adult transfers student responses produced using Assistive Technology on CBT test form or PBT test booklet.
ASL and Closed Caption	CBT – Available for the listening items on the Reading ELA test.
Assistive Technology	Use of assistive technology, including Braille writer, for the writing response and/or other open response items. Internet access, spell-check, grammar-check, and predict-ahead functions must be turned off. An adult must transfer the student's responses exactly as written to the CBT test form or PBT test booklet. Any print copy must be shredded. Any electronic copy must be deleted. This accommodation also requires Adult Transcription.
Braille Test Booklet	Provide a paper Braille test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.
Large Print Test Booklet	CBT – Either increase default zoom settings and student participates in CBT or provide a PBT Large Print test booklet. A PBT Large Print test booklet requires Adult Transcription on a regular size paper test booklet. PBT – Provide a Large Print test booklet. This accommodation also requires Adult Transcription on a regular size paper test booklet.
Paper Test Booklet	CBT – Provide a regular size paper test booklet for a student at a school administering the CBT. If a paper test booklet is ordered as an accommodation for a student at a CBT school, the student must use the paper test booklet and may not participate in computer-based testing.
Read Aloud Test Content	CBT – Accommodated Text-to-Speech for test content may be provided for the writing portion of the ELA test and the math test. PBT – Read aloud, in English, any of the test content in the writing portion of the ELA test and the math test. Reading aloud the content of the Reading portion of the ELA test is prohibited.
Rest/Breaks	Student may take breaks during testing sessions to rest.
Sign Test Content	Sign any of the content of the Writing portion of the ELA test. Sign any of the content of the Math test. Signing the content of the Reading portion of the ELA test is prohibited.
Simplified Directions	Provide verbal directions in simplified English for the scripted directions or the directions that students read on their own.

6.4 SYSTEM SECURITY

6.4.1 SECURE SYSTEM DESIGN

AIR has developed a custom single sign-on application that is made available in Arizona's secure portal. This application is used to support access to AIR's system in accordance with the Arizona's user ID and password policy. Authorized users can log in to Arizona's single sign-on using their current user IDs and passwords and can be redirected to AIR's portal, where they have access to AIR's secure applications such as the Test Information Distribution Engine (TIDE), the test delivery system (TDS), and online reporting system (ORS). Nightly backups protect the data. The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or they will need to rerun the backup. The system can withstand failure of almost any component with little or no interruption of service.

AIR's hosting provider, Rackspace, has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely. Rackspace partners with 9 different network providers, providing multiple, redundant data routes. Every installation is served by multiple servers, any one of which can take over for an individual test upon failure of another.

AIR's architecture ensures data are recoverable at all times. Each disk array is internally redundant, with multiple disks containing each data element. Immediate recovery from failure of any individual disk is performed by accessing the redundant data on another disk. AIR maintains support and maintenance agreements through our hosting provider for all of the hardware used by our systems.

6.4.2 SYSTEM SECURITY COMPONENTS

AIR has built-in security controls in all of its data stores and transmissions.[47] Unique user identification is a requirement for all systems and interfaces. All of AIR's systems encrypt data at rest and in transit.

---

[47] Standard 6.16 – Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores and pertinent ancillary information.
Standard 8.6 – Test data maintained or transmitted in data files, including all personally identifiable information (not just results), should be adequately protected from improper access, use, or disclosure, including by reasonable physical, technical, and administrative protections as appropriate to the particular data set and its risks, and in compliance with applicable legal requirements. Use of facsimile transmission, computer networks, data banks, or other electronic data-processing or transmittal systems should be restricted to situations in which confidentiality can be reasonably assured. Users should develop and/or follow policies, consistent with any legal requirements, for whether and how test takers may review and correct personal information.

PHYSICAL SECURITY

AzMERIT data resides on servers at Rackspace, AIR's hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. All access is keycard controlled, and sensitive areas require biometric scanning.

Secure data are processed at AIR facilities and are accessed from AIR machines. AIR's servers are in a secure, climate-controlled location with access codes required for entry. Access to our servers is limited to our network engineers, all of whom, like all AIR employees, have undergone rigorous background checks.

Staff, at both AIR and Rackspace, receive formal training in security procedures to ensure that they know the procedures and implement them properly. AIR and Rackspace protect data from accidental loss through redundant storage, backup procedures, and secure off-site storage.

NETWORK SECURITY

Hardware firewalls and intrusion detection systems protect our networks from intrusion. They are installed and configured to prevent access for services other than hypertext transfer protocol secure (HTTPS) for our secure sites.

AIR's systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts.

SOFTWARE SECURITY

All of AIR's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with Arizona's privacy laws, the Family Educational Rights and Privacy Act (FERPA), and other federal laws.

AIR's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. Different states interpret the FERPA differently, and our system is designed to support these interpretations flexibly. AIR has worked with the ADE to maintain data security according to their specifications.

AIR maintains logs of key activities and indicators, including data backup, server response time, user accounts, system events and security, and load test results. In addition, AIR runs automated functional tests of our test delivery system every morning, and logs from these runs are available for at least one week from the time of the run.

AIR psychometricians monitor the quality and performance of test administrations statewide through a series of quality assurance (QA) reports. The QA reports provide information on item behavior and also provide a forensics analysis report. The forensics analysis report is described more completely in Section 6.6 on Data Forensics.

6.5 TEST SECURITY

Maintaining a secure test environment is critical to ensure that scores represent what students know and are able to do. Because AzMERIT was administered both as a paper-based and a computer-based assessment, test security procedures must guard against item exposure, cheating on the part of test administrators or students, or other security problems for both testing modes.

The test security procedures involve the following:

- Procedures to ensure security of test materials
- Procedures to investigate test irregularities

Test Administrators are trained on test security procedures and both test security policies and procedures are clearly presented with the *AzMERIT Test Administration Directions.*[48]

Security of Test Materials

All test items, test materials, and student-level testing information are secure documents and must be appropriately handled. Secure handling protects the integrity, validity, and confidentiality of assessment questions, prompts, and student results. Any deviation in test administration must be reported to ensure the validity of the assessment results. Mishandling of test administration puts student information at risk and disadvantages the student. Failure to honor security severely jeopardizes district and state accountability requirements and the accuracy of student data.

The security of all test materials must be maintained before, during, and after test administration. Under no circumstances were students permitted to assist in preparing secure materials before testing or in organizing and returning materials after testing. After any administration, initial or make-up, secure materials (e.g., test booklets, test tickets, used scratch paper) were required to be returned immediately to the School Test Coordinator and placed in locked storage. Secure materials were never to be left unsecured and were not to remain in classrooms or be taken off the school's campus overnight. Secure materials were never to be destroyed (e.g., shredded, thrown in the trash), except for soiled documents. In addition, any monitoring software that would allow test content on student workstations to be viewed or recorded on another computer or device during testing needed to be turned off.

It is unethical and shall be viewed as a violation of test security for any person to:

- capture images of any part of the test via any electronic device;
- duplicate in any way any part of the test;
- examine, read, or review the content of any portion of the test;
- disclose or allow to be disclosed the content of any portion of the test before, during, or after test administration;
- discuss any AzMERIT test item before, during, or after test administration;
- allow students access to any test content prior to testing;

[48] Standard 6.7 – Test users have the responsibility of protecting the security of test materials at all times. Standard 7.9 – If test security is critical to the interpretation of test scores, the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session.

- provide any reference sheets to students during the Math test administration;
- allow students to share information during test administration;
- allow students to use scratch paper during the ELA Reading test;
- read any parts of the test to students except as indicated in the Test Administration Directions or as part of an accommodation;
- influence students' responses by making any kind of gestures (for example, pointing to items, holding up fingers to signify item numbers or answer options) while students are taking the test;
- instruct students to go back and reread/redo responses after they have finished their test since this instruction may only be given before the students take the test;
- review students' responses;
- read or review students' scratch paper; or
- participate in, direct, aid, counsel, assist in, encourage, or fail to report any violations of these test administration security procedures.

Additional security violations for paper-based testing include:

- Reading or reviewing any test booklet during or after testing,
- Changing any student response in test booklet,
- Erasing any students response in test booklet,
- Erasing any stray marks in test booklet,
- Failing to return all test booklets and other test materials.

Test Administrators and Proctors may not assist students in answering questions. Test Administrators and Proctors may not translate, reword, or explain any test content. No test content may ever be discussed before, during, or after test administration.

All regular test booklets and special documents (large print and Braille) test materials are secure documents and must be protected from loss, theft, and reproduction in any medium. A unique identification number and a bar code were printed on the front cover of all test booklets. Schools were expected to maintain test security by using the security numbers to account for all secure test materials before, during, and after test administration until the time they were returned to the contractor.

To access the computer-based AzMERIT tests, a secure Internet browser was required. The secure browser provides a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to the desktop (Internet, email, and other files or programs installed on school machines). The secure browser did not display the IP address or other URL for the site. Users could not access other applications from within the secure browser, even if they knew the keystroke sequences. The "back" and "forward" browser options were not available, except as allowed in the testing environment as testing navigation tools. Students were not able to print from the secure browsers. During testing, the desktop was locked down, and students were required to "Pause" (to save the test for another session) or "Submit" a test in order to exit the secure browser. The secure browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. See the *Test Administrator User Guide* for further details.

Throughout the testing window, test administrators were to report any test incidents (e.g., disruptive students, loss of Internet connectivity) to the School Test Coordinator immediately. A test incident could include testing that was interrupted for an extended period of time due to a local technical malfunction or severe weather. School Test Coordinators notified District Test Coordinators of any test irregularities that were reported. District Test Coordinators were responsible for submitting requests for test invalidations to the Department of Education via

AIR's Test Information Distribution Engine, or TIDE. The Department of Education made the final decision on whether to approve the requested test invalidation. District Test Coordinators could track the status and final decisions of requested test invalidations in TIDE.

6.6 DATA FORENSICS PROGRAM

The validity of test score interpretation depends critically on the integrity of the test administrations on which those scores are based. Any irregularities in the administration of assessments can therefore cast doubt on the validity of the inferences based on those test scores. Multiple facets ensure that tests are administered properly which include clear test administration policies, effective test administrator training, and tools to identify possible irregularities in test administrations.

For online administrations, quality assurance (QA) reports are generated during and after the test windows. These are geared toward detection of testing irregularities that may indicate possible cheating, aggregating unusual responses at the student level to detect possible group-level testing anomalies.

Online test administration allows Arizona's testing contractor to track information that was not possible to track in the context of the paper-and-pencil tests. This information includes not only item responses but also item response changes, latencies between item responses and changes, number of revisits to an item or items, test start and end times, scores in each opportunity in the current year, scores in the previous year, and other selected information in the system (e.g., accommodations) as requested by the state. AIR's test delivery system (TDS) captures all of this information.

Unlike with paper assessments where data analysis must await the close of test window and processing of answer documents, AIR's TDS allows AIR psychometricians and state assessment staff to monitor testing anomalies throughout each test administration window, following the first operational administration. Following the base year, the analyses used to detect the testing anomalies can be run anytime within the testing window. Evidence evaluated included changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. Analyses are performed at student-level and summarized for each aggregate unit, including testing session, test administrator, and school.

6.6.1 CHANGES IN STUDENT PERFORMANCE

The report examines score changes between years using a regression model. The scores between the previous and current year assessments are compared, with the current-year score regressed on the test score from the previous year.

A large score gain or loss between grades is detected by examining the residuals for outliers. The residuals are computed as observed value minus predicted value. To detect unusual residuals, we compute the studentized *t* residuals. An unusual increase or decrease in student scores between opportunities is flagged when absolute studentized *t* residuals are greater than 3.

The number of students with a large score gain or loss is aggregated for a testing session, test administrator, and school. Unusual changes in an aggregate performance between administrations and/or years is flagged based on the average studentized *t* residuals in an aggregate unit g (e.g., a testing session or a test administrator). For each aggregate unit, a critical *t* value is computed and flagged when absolute *t* was greater than 3,

$$t = \frac{Average\ residuals}{\sqrt{\frac{s^2}{n_g} + \frac{\sum_{j=1}^{n_g} var(e_i)}{n_g^2}}},$$

where *s* = standard deviation of residuals in an aggregate unit; n_g is number of students in the aggregate unit g (e.g., testing session or test administrator); and $var(e_i) = \sigma^2(1 - h_{ii})$. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

If the aggregate unit size is 1–5 students, the aggregate unit was flagged if the percentage of flagged students was greater than 50%. The aggregate unit size for the score change is based on the number of students included in the within- or between-year regression analyses in the aggregate unit.

6.6.2 ITEM RESPONSE LATENCY

The online environment also allows item response latency to be captured as the item page time (the time each item page is presented) in milliseconds. Discrete items appear one item on the screen at a time. However, for stimulus-based items selected as part of an item group, all items associated with the stimulus are selected and loaded as a group. For each student, the total time taken to complete the test is computed by summing up the page time for all items and item groups.

As expected, the item response time was shorter than the average time if students have prior knowledge of test items. An example of unusual item response time would be a test record for an individual who scores very well on the test even though the average time spent for each item was far less than that required of students statewide. If students already know the answers to the questions, the response time will be much shorter than the response time for those items where the student has no prior knowledge of the item content. Conversely, if a test administrator helps students by "coaching" them to change their responses during the test, the testing time could be longer than expected.

The average and the standard deviation of test-taking time are computed across all students for each opportunity. Students and aggregate units were flagged if the test-taking time was greater than |3| standard deviations of the state average. The state average and standard deviation was computed based on all students at the time the analysis was performed. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

6.6.3 INCONSISTENT ITEM RESPONSE PATTERN (PERSON FIT)

In Item Response Theory (IRT) models, person-fit measurement is used to identify examinees whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test-taker has prior knowledge of some test items (or is provided answers during the exam), the student will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. We note, however, that if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response latency index might flag such a student.

The person-fit index is based on all item responses. An unlikely response to a single test question may not result in a flagged person-fit index. Of course, not all unlikely patterns indicate cheating, as in the case of a student who is able to guess a significant number of correct answers. Therefore, the evidence of person-fit index should be evaluated along with other testing irregularities to determine possible testing irregularities. The number of flagged students is summarized for every testing session and test administrator.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985), Sotaridona, Pornell, and Vallejo (2003) define aberrant response patterns as a deviation from the expected item score model. Snijders (2001) showed that the distribution of l_z is asymptotically normal (i.e., with an increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the "asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05" (Snijders, 2001).

Sotaridona et al. (2003) report promising results of using l_z for systematic flagging of aberrant response patterns. Students with $|l_z|$ values greater than 3 are flagged. Aggregate units are flagged with $|t|$ greater than 3, where *t* is be calculated by

$$t = \frac{\bar{l}_z}{\sqrt{(s^2+1)/n_g}},$$

where *s* is the standard deviation of l_z values in an aggregate unit g and n_g is number of students in the aggregate unit. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit (e.g., test session, test administrator, school).

6.6.4 RESPONSE CHANGE AND RESPONSE SIMILARITY

Response Change in Paper-based Tests

Erasure patterns on paper-pencil tests are also examined for unusual patterns of response changes. For paper-based assessments, we use differences in mark density to infer student erasures, which is then used to identify instances where students may have changed an initial response from incorrect to correct, from incorrect to incorrect, or from correct to incorrect. A set of flagging rules is then used to identify an unusually large number of incorrect to correct erasures at the targeted level of analysis, whether student, testing group, or school. In the online environment, students may change their responses multiple times, and each of those response changes is recorded. Unlike with the mark discrimination analyses, there is no ambiguity about which response was selected or the order in which responses were made. The ease with which response changes can be made, and the accuracy of response capture (i.e., students no longer need to worry that an "erased" response might result in the detection of multiple marks that either cannot be resolved or do not correspond to the student's intended response) mean that students may now feel freer to change responses, even multiple times for a single item.

Response Pattern Similarity in Computer-based Tests

In fixed-form assessment environments, students may more readily copy from one another than would be possible in a computer adaptive test environment where students are seeing different sets of items in different sequences. To detect possible copying, it can be useful to examine student response records for patterns of excessive response similarity. While similarity in student responses to test questions may be an indicator of irregularities in test administration, response similarity does not always indicate a testing irregularity. For example, in schools with high levels of academic achievement, one would expect large numbers of students to respond correctly, and therefore similarly, to most items on the test. Nevertheless, patterns of similar responding can indicate testing irregularities, especially when students respond to items incorrectly in the same way. We employ an algorithm, following the model developed by Wesolowsky (2000), for detecting overly similar student responses to multiple-choice items to evaluate patterns of student responses in schools where test irregularities are suspected.

The basic unit of analysis for evaluating response similarity in fixed form assessments is the test session. For each pair of students in a session, we compute the probability of obtaining the same response for each item, including the likelihood of answering the item correctly, as well as selecting the same incorrect response option when answering an item incorrectly. The probability of two students answering an item correctly is conditioned on the average performance of other students in the school. The Bonferroni adjustment is used to correct for the large number of pairwise comparisons, reducing the likelihood of Type I (false positive) errors. A response similarity report identifies pairs of students with overly similar patterns of responding. Exhibit 6.6.4.1 provides sample output for the response similarity analysis. Each record indicates a pair of students flagged for overly similar patterns of responding. Access to a seating chart increases the power of this approach significantly, since students with overly similar response patterns who are known to have been seated in close proximity, obviously have greater opportunity to copy their responses. This method is also useful for detecting cheating rings, where the same students are identified across multiple flagged pairs. This is evident in Exhibit 6.6.4.1, where a common group of students are each flagged in multiple comparisons.

Exhibit 6.6.4.1 Sample Roster Flagging Student Pairs with Excessively Similar Responses

School	Testing Group	Subject	Class Size	Student1 Barcode	Student1 LastName	Student1 FirstName	Student2 Barcode	Student2 LastName	Student2 FirstName
SchoolA	Class1	Reading	18		Carter	Adam		Doe	Frank
SchoolA	Class1	Reading	18		Carter	Adam		Farmer	Fred
SchoolA	Class1	Reading	18		Carter	Adam		Miller	Steve
SchoolA	Class1	Reading	18		Carter	Adam		Smith	Cecil
SchoolA	Class1	Reading	18		Carter	Adam		Carter	Henry
SchoolA	Class1	Reading	18		Carter	Adam		Turner	Mark
SchoolA	Class1	Reading	18		Carter	Adam		Granger	Carl
SchoolA	Class1	Reading	18		Carter	Adam		Hall	Robert
SchoolA	Class1	Reading	18		Carter	Adam		Granger	Phillip
SchoolA	Class1	Reading	18		Doe	Frank		Farmer	Fred
SchoolA	Class1	Reading	18		Doe	Frank		Carter	Henry
SchoolA	Class1	Reading	18		Doe	Frank		Hall	Robert
SchoolA	Class1	Reading	18		Doe	Frank		Granger	Phillip
SchoolA	Class1	Reading	18		Farmer	Fred		Miller	Steve
SchoolA	Class1	Reading	18		Farmer	Fred		Smith	Cecil
SchoolA	Class1	Reading	18		Farmer	Fred		Carter	Henry
SchoolA	Class1	Reading	18		Farmer	Fred		Turner	Mark
SchoolA	Class1	Reading	18		Farmer	Fred		Granger	Carl
SchoolA	Class1	Reading	18		Farmer	Fred		Hall	Robert

School	Testing Group	Subject	Class Size	Student1 Barcode	Student1 LastName	Student1 FirstName	Student2 Barcode	Student2 LastName	Student2 FirstName
SchoolA	Class1	Reading	18		Farmer	Fred		Granger	Phillip
SchoolA	Class1	Reading	18		Miller	Steve		Smith	Cecil
SchoolA	Class1	Reading	18		Miller	Steve		Carter	Henry
SchoolA	Class1	Reading	18		Miller	Steve		Turner	Mark
SchoolA	Class1	Reading	18		Miller	Steve		Hall	Robert
SchoolA	Class1	Reading	18		Miller	Steve		Granger	Phillip

7. REPORTING AND INTERPRETING AZMERIT SCORES

A set of score reports is provided for each administration that summarizes student performance in each grade and content area. Score reports provide data on the performance of individual students and on the aggregated performance of students at various levels—such as state, districts, schools, and teachers. The test data are based on all students who participated in the AzMERIT assessment for the 2015-2016 school year.

The score reports include reliable and valid information describing student progress toward mastery of the state content standards. Arizona provides individual student score reports that are mailed directly to families, detailing student performance on overall tests and subscores. In addition, Arizona offers detailed individual and aggregate level data to educators via AIR's Online Reporting System (ORS), which provides score data for each AzMERIT test, both computer-based and paper-based. The ORS allows users to compare score data between individual students and the school, district, or overall state, and also provides information about performance on subscore categories.

7.1 APPROPRIATE USES FOR SCORES AND REPORTS

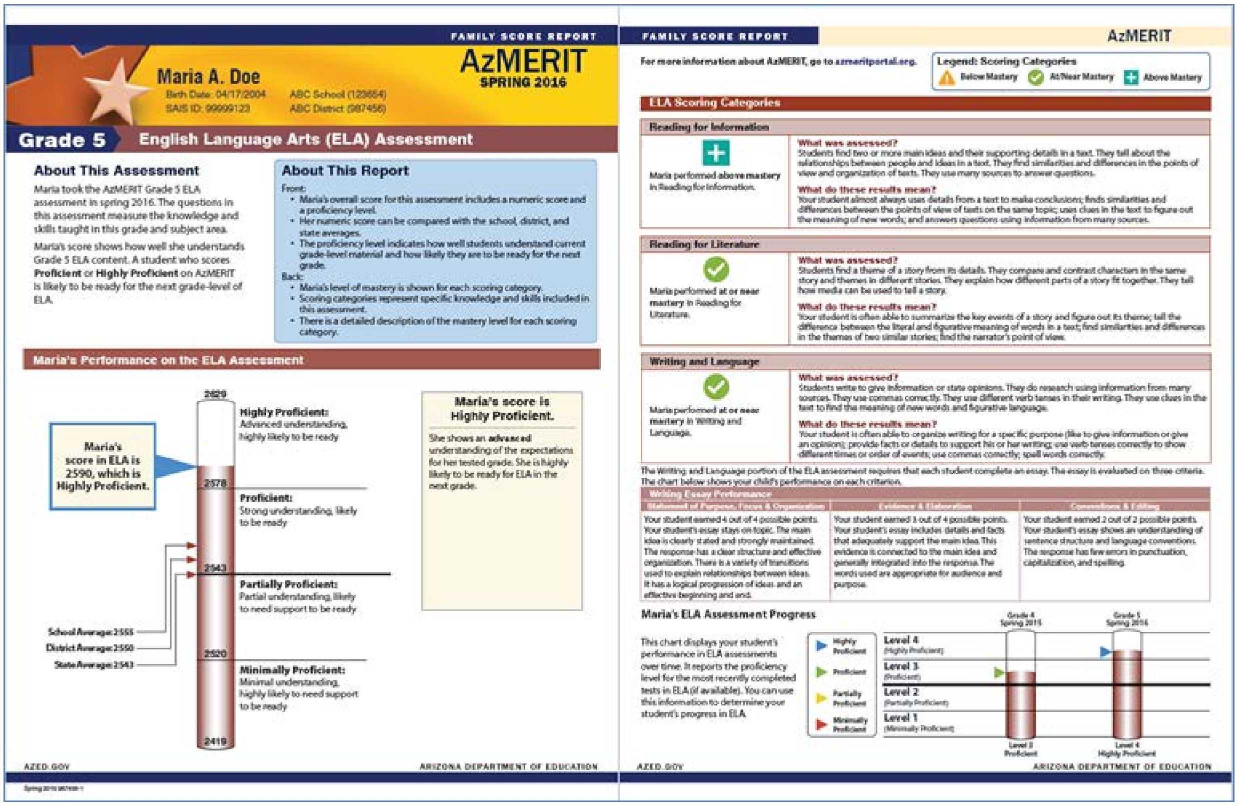
The state provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning and classroom instruction. All reporting systems for the AzMERIT, both paper and online, are designed with stakeholders, such as teachers, parents and students, who are not technical measurement experts, in mind and ensure that test results are used in ways that lead to valid inferences about student achievement and contribute to student learning.[49] For example, similar colors are used for groups of similar elements, such as performance levels, throughout the design. This design strategy guides the reader to compare like elements and avoid comparison of dissimilar elements.

Sample reports are available at azmeritportal.org. The sections below provide additional guidance for interpreting results.

[49] Standard 6.10 – When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. Standard 13.5 – Those responsible for the development and use of tests for evaluation or accountability purposes should take steps to promote accurate interpretations and appropriate uses for all groups for which results will be applied.

7.2 REPORTS PROVIDED

7.2.1 FAMILY REPORTS



Arizona provides full-color individual student reports to families of all AzMERIT testers. Reports are designed to be useful to families, and include:

- full color to aid readers' interpretation of the data;
- scale scores and performance level descriptors;
- scoring category performance, including descriptions of what was assessed and what results mean for each scoring category to guide parents and students in their understanding of student scores;
    - o A plus (+) symbol indicates that a student is performing above mastery in a particular scoring category
    - o A checkmark indicates that a student is performing at or near mastery within the scoring category, and
    - o The exclamation symbol indicates a student is performing below mastery in a scoring category
- rubric scores for the writing portion of the ELA test, including descriptions of what those rubric scores mean; and
- school, district, and state average scores for comparative purposes.

In addition, beginning with the spring 2016 administration, ADE provided reports that included longitudinal data as seen at the bottom of the second page of the report. This data is designed to allow parents to track student achievement over time.

7.2.2 ONLINE REPORTING SYSTEM FOR EDUCATORS

AzMERIT results are reported using AIR's Online Reporting System, which is designed to support educators as they evaluate the needs of their students and reflect on their own curricula and practice. Navigation in the system mirrors the instructional decision-making process, meaning the user can intuitively navigate in any of the three dimensions inherent in the data, helping the user answer three kinds of questions:

1. Who? The data can be displayed at levels of aggregation anywhere from the individual level for a specific student up to the entire state. Demographic breakdowns are immediately available at any level of aggregation.
2. What? The subject area data can be broken down in into finer or coarser "chunks" of content. Navigating this dimension allows the user to travel from subject to scoring category and back.
3. When? When data are available over time, the system allows the user to view a data trend over time or toggle to a fixed point in time.

Each navigational step changes the reporting display, providing richer context when interpreting a class's or individual student's performance. While the system contains many reports, the interface design encourages users to think about the substantive, educational questions to which they need answers and access information from that perspective. In addition, while finding and interpreting data from multiple online assessments can easily become overwhelming, the ORS minimizes information overload for educators and administrators by organizing score information in a conceptual framework that helps users quickly locate the right level of data, evaluate its impact, and identify the concrete actions they can take to help students improve.

The AzMERIT online system produces the following online score reports: individual student reports, and aggregate reports at the teacher, school, district, and state level. The AzMERIT online score reports are structured hierarchically. Upon selecting "Home" on the Welcome page, a user is taken to the Home Page Dashboard, which displays for all grades and content areas the number of students tested and the percent of students passing by grade and content area. Users who have access to multiple districts or schools are first required to select a single district or school. Once an aggregate unit is selected in this instance, the summary table of student performance for the selected entity displays. For more detailed information for a subject and a grade, the user must select that subject and grade.

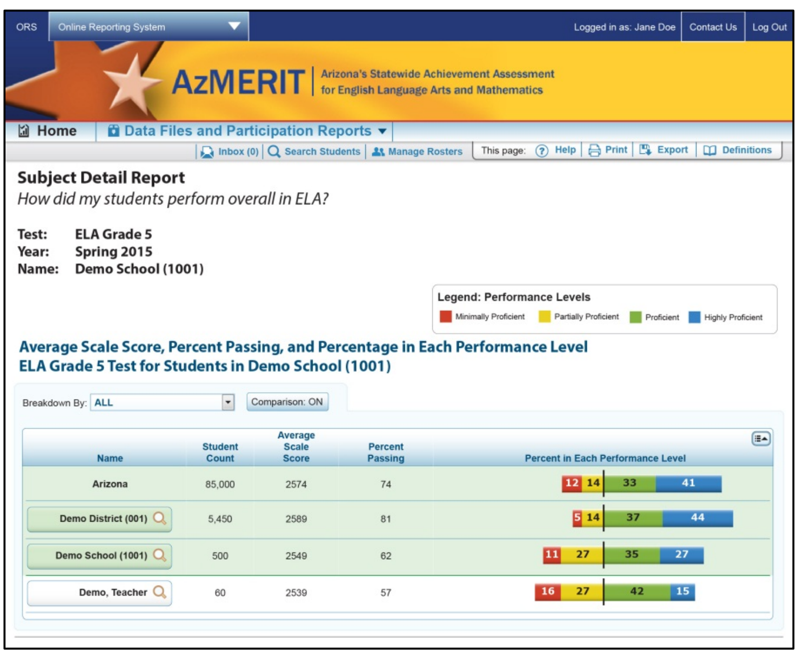
On each aggregate report, the summary report presents the results for the selected aggregate unit as well as the results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected on the school report page, the summary results of the state and the district the school belongs to are provided above the school summary results so that the school performance can be compared with the district and the state. If a teacher is selected, the summary results for state, district, and school are provided above the summary results for the teacher.

Exhibit 7.2.2.1 summarizes the types of online score reports available and the levels at which they can be viewed (e.g., student, roster, teacher, school, district).

Exhibit 7.2.2.1 AzMERIT Online Score Report Summary

Type of Report Page	Level of Aggregation	Description
Home Page Dashboard	District, school, and teacher	Summary of performance and participation (Number Tested and Percent Passing) across grades and subjects or course
Subject Detail	District	Average scale score, percent passing, and percent at each performance level for a district and each school within that district; ability to disaggregate data by subgroup
	School	Average scale score, percent passing, and percent at each performance level for a school and each teacher within that school; ability to disaggregate data by subgroup
	Teacher	Average scale score, percent passing, and percent at each performance level for a teacher and each class roster associated with that teacher; ability to disaggregate data by subgroup
Scoring Category Detail	District, school, teacher, and roster	Performance on the scoring category for a subject and a grade for all students and by subgroups; a relative strength and weakness indicator is also reported for each category
Student Roster	School, teacher, roster	List of students with performance on overall subject and scoring categories for a group of students associated with a school, teacher, or roster.
Individual Student Report	Student	Student performance for a selected subject; report includes performance on each scoring category, and performance on the writing essay dimensions, if applicable

SUBJECT DETAIL REPORTS

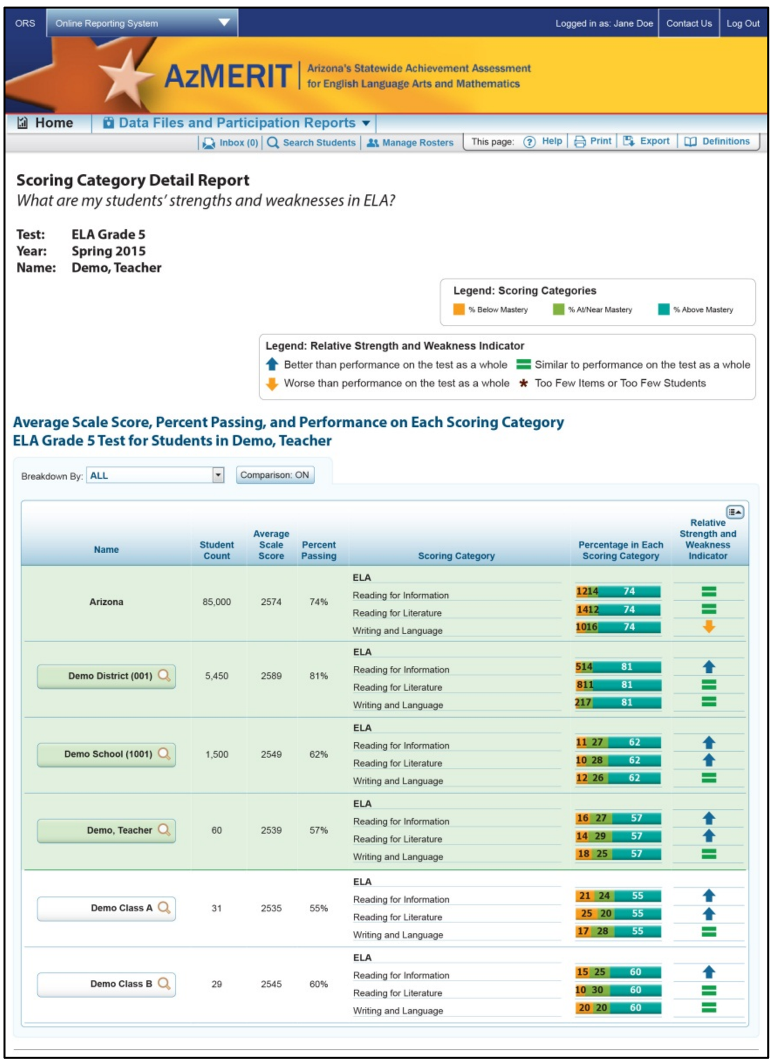


Aggregated subject reports show average performance for the state, districts, schools, teachers, and classes. Bar charts displays show the distribution of students' performance levels. These reports provide users with rosters of schools, teachers, and classes, allowing for simple comparisons across smaller groups.

The Subject Detail Report page shows the following data:

- **Student Count:** Number of students who have completed who completed the selected test
- **Average Scale Score:** Average scale score of students who completed the selected test
- **Percent Passing:** The percent of tested students reaching the proficient threshold on the selected test
- **Percent at Each Performance Level:** The distribution of students across each of the four performance levels

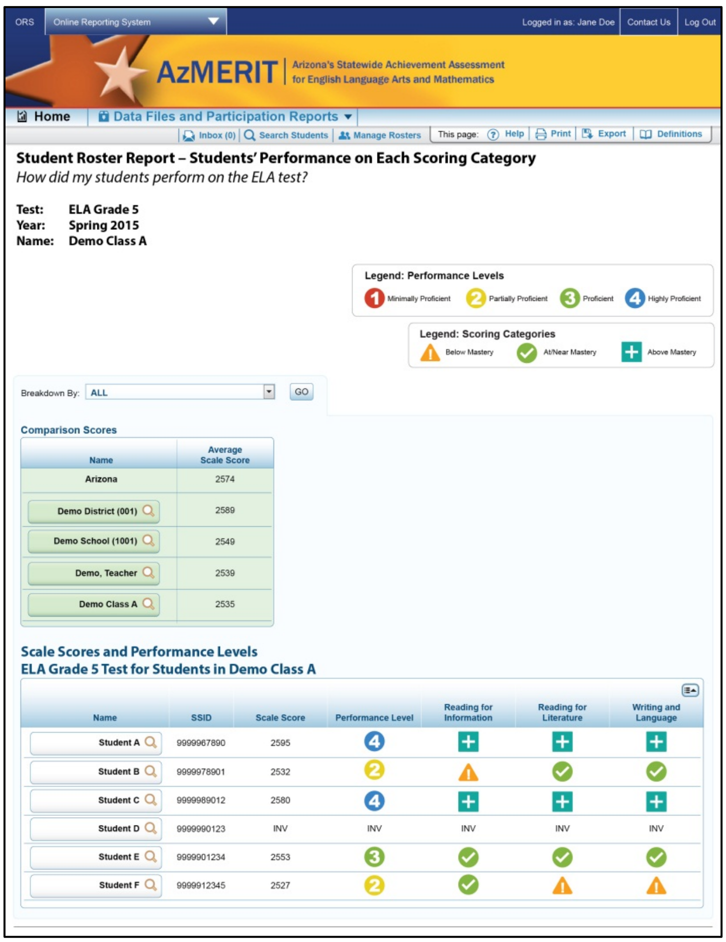
SCORING CATEGORY DETAIL REPORTS



Aggregated scoring category detail reports follow the layout of the subject detail reports, displaying the performance data for the state, districts, schools, teachers, and classes. In addition, these reports include a relative strength and weakness indicator for each category.

In addition to overall test scores, reporting category performance is reported as a strength and weakness indicator. The performance levels indicated on this report are relative to the test as a whole. Unlike performance levels provided at the subject level, these strengths and weaknesses do not imply proficiency. Instead, they show how a group of students' performance is distributed across the scoring categories relative to their overall subject performance on a test. For example, a group of students may have performed very well in a subject, but performed slightly lower in several scoring categories. Thus, the minus sign for a scoring category does not imply a lack of proficiency. Instead, it simply communicates that these students' performance on that scoring category was statistically lower than their performance across all other scoring categories put together. Although the students are doing well, an educator may want to focus instruction on these areas.

STUDENT ROSTER REPORTS



Student roster reports provide users with performance data for a group of students associated with a teacher or a school, as defined in TIDE. The report includes each student's unique state ID, overall subject score, and overall subject performance level. Using the exploration menu, a user can also view each student's scoring category performance for the selected test.

The table that appears on the Student Roster Report page shows the following data:

- **Scale score:** The score of each student who completed the test.

- **Performance level:** Represents levels of overall subject mastery with respect to the Arizona College and Career Ready Standards
- **Scoring Categories:** Represents levels of scoring category mastery with respect to the Arizona College and Career Ready Standards, characterizing achievement at "above," "at or near," or "below" mastery on each scoring category.

INDIVIDUAL STUDENT REPORTS

Individual Student Reports, which closely mirror the Family Reports, are also available through the Online Reporting System.

7.3 INTERPRETATION OF SCORES

Arizona provides a variety of resources for helping parents and educators understand and apply student performance results to improve student learning, including interpretive guides for navigating the online reporting system, and understanding paper family reports.[50] This section describes many of the measures presented in the paper and online score reports.

Performance levels represent levels of mastery with respect to the Arizona College and Career Ready Standards for a content area assessment. Performance levels are labeled as Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Performance standards are the points on the achievement scale that differentiate performance levels. Three performance standards are used to classify students into one of the four performance levels. Performance standards were recommended by panels of Arizona educators following the first administration of AzMERIT in 2015, and subsequently adopted by the Arizona Board of Education. Panelists engaged in a rigorous, technically sound standard setting process that is summarized in the Performance Standards section of this technical manual, and documented in detail in the 2015 standard setting technical report, available from ADE.

Performance Level Descriptors, or PLDs, define the content area knowledge, skills, and processes that examinees at a performance level are expected to possess. The descriptions of Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient performance are the public statements about what and how much Arizona educators want students to know and be able to do for each grade level and content area. The very detailed PLDs are summarized and included in score reports to provide context for the score and are designed to help parents understand what their students can and cannot do.

The student's performance in each content area assessment is summarized in an overall test score referred to as a scale score. The number of items a student answers correctly and the difficulty of the items presented are used to statistically transform theta scores to scale scores so that scores from different sets of items can be meaningfully compared. The scale score is used to determine how well students perform on each content area assessment. Scale scores can be used to measure how much students know and are able to do. Scale scores can also be used to compare student performance across administrations for the same grade and content area so that, for example, an average scale score of 2450 for grade 3 students in the 2015–2016 school year indicates the same level of

[50] Standard 12.18 – In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.

achievement as an average scale score of 2450 for grade 3 students in the 2016–2017 school year even though the test may include a slightly different set of items.

As described in Section 9 on Scaling and Equating, for the ELA assessment, the scale score reported can range from 2395 to 2675. For the math assessment, the scale score reported can range from 3395 to 3839. Overall scale scores for ELA and math are mapped into four performance levels using three performance standards (i.e., cut scores). The AzMERIT scale score ranges can be found in Exhibit 7.3.1.

Exhibit 7.3.1 AzMERIT Scale Score Ranges

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
		ELA		
Grade 3 ELA	2395-2496	2497-2508	2509-2540	2541-2605
Grade 4 ELA	2400-2509	2510-2522	2523-2558	2559-2610
Grade 5 ELA	2419-2519	2520-2542	2543-2577	2578-2629
Grade 6 ELA	2431-2531	2532-2552	2553-2596	2597-2641
Grade 7 ELA	2438-2542	2543-2560	2561-2599	2600-2648
Grade 8 ELA	2448-2550	2551-2571	2572-2603	2604-2658
Grade 9 ELA	2454-2554	2555-2576	2577-2605	2606-2664
Grade 10 ELA	2458-2566	2567-2580	2581-2605	2606-2668
Grade 11 ELA	2465-2568	2569-2584	2585-2607	2608-2675
		Math		
Grade 3 Math	3395-3494	3495-3530	3531-3572	3573-3605
Grade 4 Math	3435-3529	3530-3561	3562-3605	3606-3645
Grade 5 Math	3478-3562	3563-3594	3595-3634	3635-3688
Grade 6 Math	3512-3601	3602-3628	3629-3662	3663-3722
Grade 7 Math	3529-3628	3629-3651	3652-3679	3680-3739
Grade 8 Math	3566-3649	3650-3672	3673-3704	3705-3776
Algebra I	3577-3660	3661-3680	3681-3719	3720-3787
Geometry	3609-3672	3673-3696	3697-3742	3743-3819
Algebra II	3629-3689	3690-3710	3711-3750	3751-3839

ELA and math assessments are reported on a vertical scale. The item response theory (IRT) vertical scale was developed by embedding operational test items from the grade above in the embedded field test slot of each grade level assessment.

8. PERFORMANCE STANDARDS

In the summer of 2015, following the close of the first test administration windows, AIR convened panels of Arizona educators to recommend performance standards on each of the AzMERIT assessments. Details of the panels, procedures, and outcomes are documented in the “Recommending AzMERIT Performance Standards” technical report, which is available from ADE.[51] This section briefly describes the procedures used by educators to recommend standards, and resulting performance standards.

8.1 STANDARD SETTING PROCEDURES

Student achievement on the AzMERIT is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Interpretation of the AzMERIT test scores rests fundamentally on how test scores relate to performance standards that define the extent to which students have achieved the expectations defined in the Arizona College and Career Ready Standards. AzMERIT test scores are reported with respect to four proficiency levels, demarcating the degree to which Arizona students have achieved the learning expectations defined by the Arizona College and Career Ready Standards. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade level expectations for achievement of the Arizona College and Career Ready Standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standard for the AzMERIT assessments are therefore central to the validity of test score interpretations.

Following the first operational administration of the AzMERIT assessments in spring 2015, a standard setting workshop was conducted to recommend to the Arizona State Board of Education a set of performance standards for reporting student achievement of the Arizona College and Career Ready Standards. The workshop consisted of a series of standardized and rigorous procedures that Arizona educators, serving as standard setting panelists, followed to recommend performance standards. The workshops employed the Bookmark procedure, a widely used method in which standard setting panelists use their expert knowledge of the Arizona College and Career Ready Standards and student achievement to map the performance level descriptors adopted by the Arizona State Board of Education onto an ordered item book (OIB) based on the first operational test form administered to students in spring 2015.

Panelists were also provided with contextual information to help inform their primarily content driven cut score recommendations. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for the grade 11 ELA and Algebra II assessments, and Programme for International Student Assessment (PISA) performance standards for the grade 10 ELA and Geometry assessments. Panelists recommending performance standard for the grade 3-8 summative assessments were provided with the approximate location of relevant NAEP performance standards at grades 4 and 8, as well as interpolated values for grade 6. Panelists were provided with the approximate locations of the Smarter Balanced performance standards for the grade 3-8 and 11 assessments in ELA and math to provide additional context about the location of performance standards for statewide assessments. Additionally, panelists were provided the corresponding locations for the previous AIMS performance standards. Panelists were asked to consider the location of these benchmark locations when making

[51] Standard 5.21 – When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.
Standard 7.4 – Test documentation should summarize test development procedures, including descriptions and the results of the statistical analyses that were used in the development of the test, evidence of the reliability/precision of scores and the validity of their recommended interpretations, and the methods for establishing performance cut scores.

their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

Panelists were also provided with feedback about the vertical articulation of their recommended performance standards so that they could view how the locations of their recommended cut scores for each grade level assessment related to the cut score recommendations at the other grade levels. This approach allowed panelists to view their cut score recommendations as a coherent system of performance standards, and further reinforces the interpretation of test scores as indicating not only achievement of current grade level standards, but also preparedness to benefit from instruction in the subsequent grade level.

8.1.1 PERFORMANCE LEVEL DESCRIPTORS

Student achievement on the AzMERIT is classified into four performance levels: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. Performance level descriptors (PLDs) define the content area knowledge and skills that students at each performance level are expected to demonstrate. The standard setting panelists based their judgments about the location of the performance standards on the PLDs as well as the Arizona College and Career Readiness Standards. The AzMERIT PLDS describe four levels of achievement:

- Minimally Proficient
- Partially Proficient
- Proficient
- Highly Proficient

Prior to convening the standard setting workshops, AIR, in consultation with ADE, drafted PLDs for each test that described the range of achievement encompassed by each performance level on the test. The PLDs were designed to be clear, concrete, and reflect Arizona's expectations for proficiency based on the Arizona College and Career Ready Standards. Following a cycle of revisions to the draft PLDs, ADE invited Arizona educators to review PLDs for each of the assessments. Based on feedback from 166 educators, PLDs were further revised, and the resulting drafts were used by standard setting panelists. ADE considered any need for clarification or revision that arose throughout the standard setting process prior to publishing the final versions of the PLDs following the standard setting workshop. AzMERIT PLDs are available at azed.gov.

8.2 RECOMMENDED PERFORMANCE STANDARDS

Panelists were tasked with recommending three performance standards (Partially Proficient, Proficient, and Highly Proficient) that resulted in four performance levels (Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient). Exhibit 8.2.1 presents the performance standard associated with panelist-recommended OIB page numbers in logit value, as well as the percentage of students classified as meeting or exceeding each standard. Following the standard setting workshop, panelist recommendations were submitted to Arizona's State Board of Education; the Board formally adopted the standards in August 2015.

Exhibit 8.2.1 Final Recommended Performance Standards for AzMERIT

Performance Level	Partially Proficient		Proficient		Highly Proficient	
	Theta	% At or Above	Theta	% At or Above	Theta	% At or Above
			ELA			
3	-0.09	56	0.29	41	1.36	10

Performance Level	Partially Proficient		Proficient		Highly Proficient	
	Theta	% At or Above	Theta	% At or Above	Theta	% At or Above
4	0.14	57	0.6	39	1.8	5
5	-0.13	63	0.63	30	1.8	3
6	-0.12	61	0.58	34	2.03	4
7	-0.02	59	0.61	33	1.9	4
8	-0.06	60	0.64	33	1.72	6
9	-0.12	53	0.59	27	1.57	6
10	0.11	51	0.58	30	1.42	8
11	-0.02	46	0.52	26	1.27	8
Math						
3	-0.16	73	1.04	42	2.43	15
4	-0.31	71	0.76	42	2.2	10
5	-0.65	71	0.41	40	1.74	13
6	-0.48	62	0.41	32	1.55	11
7	-0.19	52	0.59	30	1.51	13
8	-0.69	57	0.09	32	1.15	13
Algebra I	-0.69	55	-0.03	32	1.27	9
Geometry	-1.37	53	-0.58	30	0.96	6
Algebra II	-1.49	53	-0.78	29	0.57	6

Exhibit 8.2.2 shows the percentage of student classified at each performance level in the initial year of AzMERIT administration, based on final panelist-recommended standards for the student population overall across grade levels and courses for the ELA and math assessments.

Exhibit 8.2.2 Percentage of Students at Each Performance Level based on Final Recommended Performance Standards

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
ELA				
3	44%	15%	31%	10%
4	43%	19%	33%	5%
5	37%	33%	27%	3%
6	39%	27%	30%	4%
7	41%	26%	29%	4%
8	40%	27%	26%	6%
9	47%	26%	21%	6%
10	49%	21%	22%	8%
11	54%	20%	17%	8%
Math				
3	27%	31%	27%	15%
4	29%	29%	32%	10%
5	29%	31%	27%	13%
6	38%	30%	21%	11%
7	48%	22%	18%	13%
8	43%	24%	20%	13%
Algebra I	45%	23%	23%	9%
Geometry	47%	24%	24%	6%
Algebra II	47%	24%	23%	6%

Exhibit 8.2.3 shows the percentage of students meeting the AzMERIT proficient standard for each assessment in the base year of 2015 (meaning they are categorized as Proficient or Highly Proficient), and the approximate percentage of Arizona students that would be expected to meet the ACT college ready standard, the percentage of Arizona students meeting the NAEP proficient standards at grades 4 and 8, and the expected proficient rate for the Smarter Balanced Assessments, system wide, based on the spring 2015 field test administration. As Exhibit 8.2.3 indicates, the performance standards recommended AzMERIT assessments are quite consistent with relevant ACT college ready, and the NAEP and Smarter Balanced proficient, benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Exhibit 8.2.3 Percentage of Students Meeting AzMERIT and Benchmark Proficient Standards

	Percent of Students Meeting Standard			
Grade/ Course	**AzMERIT Proficient**	**Arizona ACT College Ready**	**Arizona NAEP Proficient**	**Projected SBAC**
		ELA		
3	41%			38%
4	38%		28%	41%
5	30%			44%
6	34%			41%
7	33%			38%
8	32%		28%	41%
9	27%			
10	30%			
11	25%	34%		41%
		Math		
3	42%			39%
4	42%		42%	38%
5	40%			33%
6	32%			33%
7	31%			33%
8	33%		32%	32%
Algebra I	32%			
Geometry	30%			
Algebra II	29%	36%		33%

9. SCALING AND EQUATING

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. Traditional item response models assume a single underlying trait and assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^{n} P(z|\theta),$$

where *Z* represents the pattern of item responses, and θ represents a student's true proficiency.

Traditional item response models differ only in the form of the function *P*(*Z*). The one-parameter model (1PL; also known as the Rasch model), is used to calibrate AzMERIT items that are scored either right or wrong, and takes the form

$$P\left(x_j = 1|\theta_k, b_j\right) = \frac{1}{1+e^{\left(\theta_k - b_j\right)}} = P_{j1}(\theta_k).$$

The *b* parameter is often called the *location* or *difficulty* parameter—the greater the value of *b*, the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered response categories (i.e., partial credit items), AzMERIT items are calibrated using using the Rasch family Masters' (1982) partial credit model. Under Masters' model, the probability of a response in category *i* for an item with m_j categories can be written as

$$P\left(x_j = i|\theta_k, b_{j0} \dots b_{jm_j-1}\right) = \frac{e^{\sum_{v=0}^{i}\left(\theta_k - b_{jv}\right)}}{\sum_{g=0}^{m_j-1} e^{\sum_{v=0}^{g}\left(\theta_k - b_{jv}\right)}}.$$

Item parameters for the assessments were calibrated following the spring administration in 2015 and a vertical scale was established for reporting both ELA and math. In addition, a series of linking studies were performed to allow comparison of performance on the AzMERIT with other state and national scales. A mode comparability study was also completed to examine possible effects of test administration mode. These studies were completed prior to establishing performance standards in summer 2015 and subsequent scoring and reporting of AzMERIT results. AzMERIT ELA is reported on a scale ranging from 2395 to 2675 across the grade level and high school End of Course tests. AzMERIT math is reported on a scale ranged from 3395 to 3839 across grade level and high school End of Course tests.

9.1 ITEM RESPONSE THEORY PROCEDURES

The AzMERIT assessment was administered for the first time in the spring of 2015. Following test administration, item response theory (IRT) procedures were used to calibrate item parameter estimates and create the new

AzMERIT scale for scoring and reporting.[52] This section describes the procedures for calibration of operational item parameters. All calibration procedures are independently applied by AIR, ADE, and HumRRO, which acts as a third party quality assurance contractor.

Within AzMERIT, students are able to skip items in both the online and paper test platforms. While omitted items are scored as incorrect for purposes of ability estimation, all omitted responses are treated as not-administered for purposes of IRT analysis. All students who respond to at least one item within each test session are considered to have attempted a test. All attempted records are included in IRT analysis with the exclusion of students who had more than one record for the same test and records that are had been invalidated prior to scaling.

9.1.1 CALIBRATION OF AZMERIT ITEM BANKS

Winsteps was used to estimate Rasch and Masters' partial credit model item parameters for AzMERIT. Winsteps is publically available software from Mesa Press. Winsteps employs a joint maximum likelihood approach towards estimation (JMLE), which jointly estimates the person and item parameters. The Rasch model is fit to estimate student responses to dichotomous (0/1 point) items. Masters' (1982) partial credit model, an extension of the one parameter Rasch model, allows for graded responses and is fit to estimate responses for polytomous items.

In spring 2015, operational items for each test were freely calibrated establishing the new AzMERIT reference scales. Following the approval of final item parameter estimates for operational items, parameter estimates for the operational items were anchored to their new AzMERIT bank values and parameter estimates for field test and linking items were estimated under that constraint. This placed parameter estimates for all field test and external linking items on the same AzMERIT scale defined by the operational item parameters.

In spring 2016, pre-equated item parameters were used to score student test records for the math assessments. For ELA, since six new writing tasks were being administered in the ELA assessments at each grade, the ELA items were recalibrated, and the equating constant necessary to place the common items back to the reference scale was identified and applied to the recalibrated item parameters, placing all test items on the base year AzMERIT scale. Mean equating was used to compute the linking constant, and all operational reading items were included in the computation.

9.1.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

To identify the likelihood of a student's ability across the ability distribution, we begin by evaluating the likelihood of achieving a score point for an item given the underlying level of ability. Let X_i be a random variable taking a student's response on item i $(i = 1, \dots, N)$ with an outcome $x_i \in \{0,1, \dots, m_i\}$. Item i is a dichotomously scored item if $m_i = 1$, and polytomously scored item if $m_i > 1$. Based on Masters' (1982) partial credit model, the probability of getting a score of x_i on item i given ability θ can be written as

[52] Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

$$P(X_i = x_i|\theta) = \frac{\exp \sum_{k=0}^{x_i}(\theta - b_{ki})}{\sum_{l=0}^{m_i} \exp \sum_{k=0}^{l}(\theta - b_{ki})},$$

with the constraint that $\sum_{k=0}^{0}(\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i. Note that if item i is a dichotomously scored item, the partial credit model becomes the Rasch model and can be written as

$$P(X_i = 1|\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the difficulty parameter for item i.

LIKELIHOOD FUNCTION

The likelihood function of ability θ given responses to N items, $\boldsymbol{x} = \{x_i\}$, can be expressed as:

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{N} P(x_i|\theta).$$

The maximum likelihood estimate $\hat{\theta} = \arg\max_{\theta} L(\theta|\boldsymbol{x})$ or equivalently, $\hat{\theta} = \arg\max_{\theta} \ln L(\theta|\boldsymbol{x})$.

DERIVATIVES

Finding the maximum likelihood estimate requires an iterative method, such as Newton-Raphson iterations. Since the log-likelihood is a monotonic function of the likelihood, the following derivatives based on the log-likelihood function are used:

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^{N} \left[x_i - \sum_{x_i=0}^{m_i} x_i\, P(X_i = x_i|\theta) \right]$$

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = \sum_{i=1}^{N} \left[\sum_{x_i=0}^{m_i} x_i\, P(X_i = x_i|\theta) \right]^2 - \sum_{i=1}^{N} \sum_{x_i=0}^{m_i} x_i^2\, P(X_i = x_i|\theta)$$

The maximum likelihood estimates of θ is found via the following iterative routine:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{\partial \ln L(\hat{\theta}_t)}{\partial \hat{\theta}_t} \Big/ \frac{\partial^2 \ln L(\hat{\theta}_t)}{\partial \hat{\theta}_t^2}.$$

This iterative process repeats until the difference between $\hat{\theta}_t$ and $\hat{\theta}_{t+1}$ is less than a pre-specified threshold.

ESTIMATING ZERO AND PERFECT SCORES

In the event of zero or perfect scores, a procedure recommended by Berkson (as cited in Linacre, 2004) is implemented to add (or subtract) 0.5 to (from) the test score prior to estimating student ability. Thus, for students responding incorrectly to all items in a scale or subscale, students will be assigned a test score of 0.5. Conversely, for students responding correctly to all items in a scale or subscale, 0.5 will be subtracted from the test score.

9.2 ESTABLISHING A VERTICAL SCALE IN ELA AND MATH

To emphasize the acquisition of new knowledge and skills in the development of the vertical scale, operational items from each grade level assessment (g) were embedded in field test slots of the assessment in the grade below (g-1).[53] In this approach, the resulting linkage represents student achievement each year on the scale of the subsequent grade level assessment for which they are preparing to receive instruction. As such, the scale scores for each assessment can be interpreted as a pre-test score for measuring student acquisition of academic content in the subsequent grade level. While this approach risks administering to students 1-2 items measuring content that they may not yet have had the opportunity to learn, it provides a more sensitive measure of student growth than could be obtained by a linking design in the linkage represents continued growth on academic content assessed in the previous year's assessment.

9.2.1 LINKING ITEMS

Since the vertical scale essentially places each AzMERIT assessment on the scale for the assessment in the grade above, we can best assure comparability of test scores between the grades by establishing the linkage using all available operational test items. Thus, to link the grade 4 assessments to the grade 5 scales, all operational items in the grade 5 assessment were made available for administration in the grade 4 embedded field test (EFT) slots. Including all operational items in the vertical linking set ensures that the item set used to link to the target adjacent grade scale represents fully the measured construct in the target grade, allowing valid inferences to be made with respect to student baseline performance for achievement in the subsequent grade level.

Because the AzMERIT assessments of English language arts (ELA) in high school continue as end-of-course (EOC) or grade-level measures of student achievement of the Arizona College and Career Ready Standards (ACCRS), each assessment can be linked to the grade above using all available operational items.

However, AzMERIT assessments of high school math are composed of a set of EOC tests that are not as consistently associated with grade-level instruction and which measure specific subsets of the content domain. For example, while math coursework in high school follows a typical progression and it would therefore be possible to embed "grade 9" Algebra I EOC items in the grade 8 math assessment, embed the "grade 10" Geometry EOC items in the Algebra I EOC exam, and embed the "grade 11" Algebra II the Geometry exam, the constructs measured across the four exams vary considerably and have implications for the interpretation of growth, or lack thereof, across assessments. For example, it is not clear what the expectation for growth should be in a vertical scale established by embedding Geometry items in an Algebra I exam, since Geometry is not a focus of instruction in Algebra I courses. An alternative approach, and the one adopted by ADE, was to link the grade 8 math scale to both the Algebra I and Geometry EOC scales, since the grade 8 assessment includes items measuring both algebra

[53] Standard 5.0 – Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.
Standard 5.2 – The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

and geometry. Because Algebra II builds on the knowledge and skills assessed in Algebra I, all Algebra II items were used to link the Algebra I assessments to the Algebra II scale.

9.2.2 LINKING ANALYSIS

When feasible, it is desirable to establish linkages using both concurrent calibrations and chain-linking approaches to ensure that results are consistent across methods. An important advantage of chain linking approaches is that, because item response theory (IRT) calibrations proceed by establishing the within-grade scale, the achievement construct intended by the blueprint and enacted in the operational test form is preserved. Unfortunately, however, at each step in the linking chain, the linking error accumulates, so that linking constants for grades more distant from the reference grade are less precise than are linking constants for grades in closer proximity to the reference grade. Concurrent calibrations do not accrue linking error across grade levels, so that linking constants are similarly precise between all grade levels. However, the calibrations resulting from this approach measure the construct that is common across the linked assessments, which may be different from the intended achievement construct at each grade level, especially for subjects such as math where the assessed construct may change markedly across grade levels. Generally, both approaches tend to converge to produce vertical scales that operate similarly (Ito, Sykes, and Yao, 2008; Karkee, Lewis, Hoskens, Yao, and Haug, 2003), and we view convergence as evidence for the robustness of the vertical scale.

Final Linking Set

To facilitate the development of a vertical scale that will be sensitive to student growth over time, we first evaluated the performance of vertical linking items between the grade levels in which they were administered to identify any items that were more difficult for students in the intended grade than they were for students in the lower grade. For math, items that showed proportion correct scores lower in the intended grade than in the lower grade were dropped from the final vertical linking set. This resulted in dropping on average just over two items per linking set, with a maximum of six items dropped for the linkage between grade 6 and grade 7 math assessments.

For reading, the proportion correct values across grades were much closer, especially at the higher grade levels, so that elimination of all items where the proportion correct value in the lower grade exceeded the higher grade would result in dropping more items from the vertical linking set than would be desirable for executing a robust equating design. Thus, we modified the rule for reading to exclude from the vertical linking set those items which showed proportion correct values more than two standard errors beyond the average standard error for the total linking set (i.e., items that were reliably less difficult at the lower grade). This approach allowed us to identify a final set of linking items that would maximize detection of growth, while retaining sufficient items to establish a strong linkage between the grade level assessments.

Exhibit 9.2.2.1 Number of Items Dropped and Remaining in the Final Vertical Linking Set

	Math		ELA	
Linkage	**Dropped Items**	**Final VL Set**	**Dropped Items**	**Final VL Set**
G3→G4	1	44	1	42
G4→G5	0	45	3	46
G5→G6	1	46	0	47

G6→G7	6	41	5	39
G7→G8	3	47	2	46
G8 M→ Algebra I G8 ELA→G9 ELA	3	28	11	30
G8 M →Geometry G9 ELA→ G10 ELA	2	31	7	39
Algebra I→ Algebra II G10 ELA→ G11 ELA	2	32	10	35

CHAIN-LINKING

The chain linking approach proceeds from the within grade item parameters identified in the initial calibrations of the operational and embedded field test items. Because operational test items at each grade were administered in the EFT slots in the grade below, each item in the vertical linking set has two sets of item parameters: on-grade (g) and below-grade (g-1). The chain linking proceeds by identifying the linking constants necessary to place the below-grade item parameters on the on-grade scale for the items in the final vertical linking set. The linking constant for each grade was defined as the mean difference of the item difficulty estimates for the linking items between the linked grades. The chain linking began by placing the grade 3 item parameters on the grade 4 scale for both math and ELA and proceeded upwards. For math EOC assessments, the grade 8 math scale was linked to both the Algebra I and Geometry scales, and the Algebra I scale was linked to the Algebra II scale.

CONCURRENT CALIBRATION

A vertical scale for each subject area was also established by calibrating simultaneously all items in the final vertical linking set. As with the within grade calibrations, parameters were estimated using Winsteps. To compare results from the chain-linking and concurrent calibrations, the concurrent calibrations were placed on the grade 3 reference scale.

Exhibit 9.2.2.2 shows the vertical linking constants resulting from chain-linking the within grade scales as well as from concurrently calibrating items from across grade levels. The linking constants are applied to their respective within grade scale to place all item parameters on the grade 3 reference scale. To more directly examine the magnitude of gains across grade level assessments, Exhibit 9.2.2.3 shows the difference between linking constants between each of the grade levels assessed. Relative gains are also represented graphically in Exhibit 9.2.2.4 and Exhibit 9.2.2.5 for math and ELA, respectively, which plot the linking constants across grade level assessments. As the linking constants indicate, for math there is relatively large and steady growth across the grade level and end of course assessments. For the ELA assessments, the cross grade gains are more modest, and tend to diminish in the higher grade levels.

Exhibit 9.2.2.2 Vertical Linking Constants Resulting from Chain-Linking Within Grade Scales and Concurrent Calibration of Items Across Grades

	Vertical Linking Constants			
	Math		ELA	
Linkage	Chain-Linked	Concurrent	Chain-Linked	Concurrent
G3→G4	1.32	1.30	0.18	0.16
G4→G5	2.75	2.67	0.81	0.78
G5→G6	3.90	3.73	1.19	1.15
G6→G7	4.48	4.28	1.44	1.39

G7→G8	5.69	5.39	1.76	1.70
G8 M→ Algebra I G8 ELA→G9 ELA	6.07	5.76	1.97	1.88
G8 M →Geometry G9 ELA→ G10 ELA	7.15	6.86	2.12	1.98
Algebra I→ Algebra II G10 ELA→ G11 ELA	7.81	7.45	2.32	2.16

Exhibit 9.2.2.3 Linking Constant Differences Between Each of the Grade Level Scales

	Vertical Linking Constant Differences			
	Math		ELA	
Linkage	Chain-Linked	Concurrent	Chain-Linked	Concurrent
G3→G4	1.32	1.3	0.18	0.16
G4→G5	1.43	1.37	0.63	0.62
G5→G6	1.15	1.06	0.38	0.37
G6→G7	0.58	0.55	0.25	0.24
G7→G8	1.21	1.11	0.32	0.31
G8 M→ Algebra I G8 ELA→G9 ELA	0.38	0.37	0.21	0.18
G8 M →Geometry G9 ELA→ G10 ELA	1.08	1.10	0.15	0.10
Algebra I→ Algebra II G10 ELA→ G11 ELA	0.66	0.59	0.20	0.18

Linking constants resulting from the chain-linking and concurrent calibration approach are quite consistent, indicating that both approaches converge on a common growth scale. Although the linking constants derived from the concurrent calibration approach may be considered more precise, the chain-linking method preserves the within grade measurement construct, and was therefore selected as a preliminary vertical scale for the purpose of recommending performance standards. We note that ordered item books for the standard setting workshop were based on the within grade scales, so any modifications to the vertical scale will not impact the recommended performance standards.

The vertical linking constants also indicate much greater growth across grades and high school courses for math than is observed for ELA. In math, growth is on the order of about one standard deviation per year, with the exception of grade 6 to grade 7, which showed just over a half standard deviation gain. Similar half standard deviation gains were observed between grade 8 and Algebra I, which some students take concurrently, and between coursework in Algebra I and Algebra II. Gains in ELA are less pronounced, with somewhat larger gains in the elementary school years, with growth attenuating in the high school grades.

Exhibit 9.2.2.4 Vertical Linking Constants Estimated from Chain-Linking and Concurrent Calibrations: Math

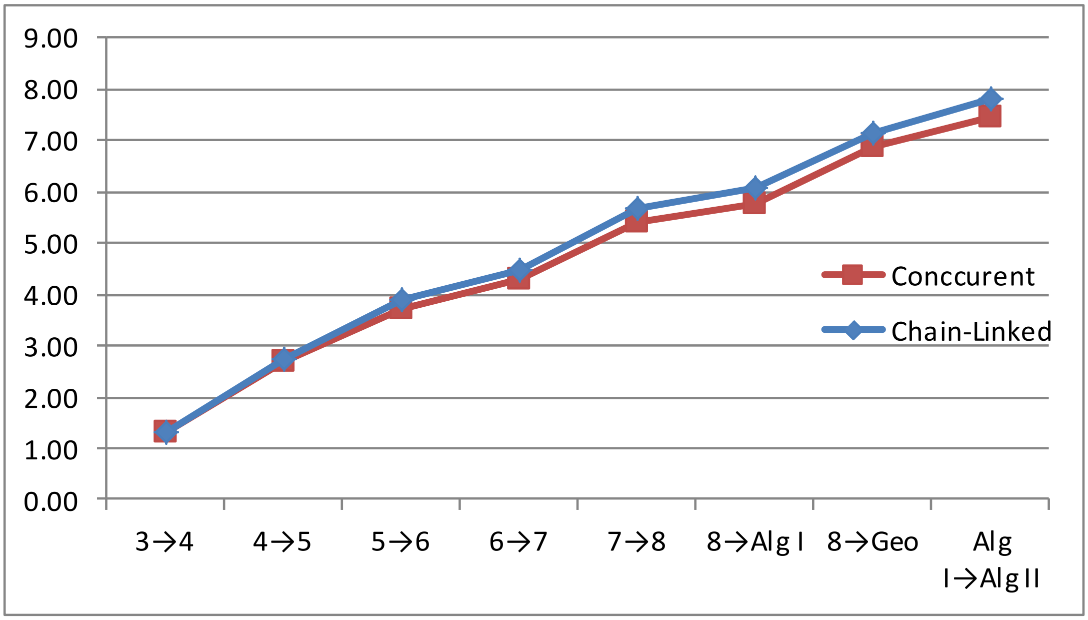
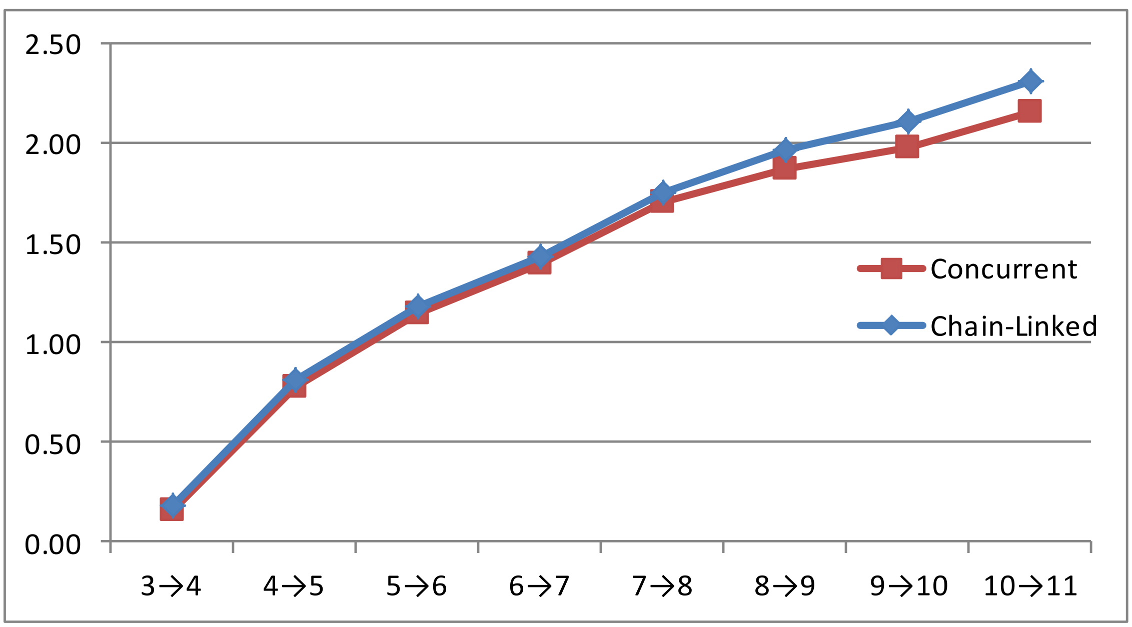


Exhibit 9.2.2.5 Vertical Linking Constants Estimated from Chain-Linking and Concurrent Calibrations: ELA



9.3 AZMERIT REPORTING SCALE (SCALE SCORES)

The AzMERIT assessments are reported on common scales within each subject (ELA and math). The IRT vertical scale scores (*SS*) are formed by linking each grade level assessment to the scale of the assessment in the grade level above. The vertical scale score is the linear transformation of the post-vertically scaled IRT ability estimate.[54]

$$SS = a * \theta_V + d$$

where $a = 30, d = 2500$ for ELA tests, and $a = 30, d = 3500$ for Math tests. $\theta_V = \theta + c$, where θ is the on-grade ability estimate and c is a vertical linking constant listed below for each of the tests, as described above. For reporting, the on-grade ability estimate is truncated at ± 3.5.

After transforming theta ability estimates to the vertical AzMERIT reporting scale, the observable scale scores nearest each of the performance standard cut scores are evaluated. If the observable scale score nearest the performance standard is below the cut score, the scale score is rounded up to be equal to the cut score. If the observable scale score nearest the performance standard is above the cut score, no special rounding rule is applied.

Overall scale scores for the AzMERIT are mapped into 4 performance levels per grade/course. The performance level designations are: Minimally Proficient, Partially Proficient, Proficient, and Highly Proficient. The performance level is evaluated using the rounded scale score.

Exhibit 9.3.1 shows the scale score ranges for the performance levels for each test.

Exhibit 9.3.1 Scale Score Ranges for Performance Levels

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
		ELA		
Grade 3 ELA	2395-2496	2497-2508	2509-2540	2541-2605
Grade 4 ELA	2400-2509	2510-2522	2523-2558	2559-2610
Grade 5 ELA	2419-2519	2520-2542	2543-2577	2578-2629
Grade 6 ELA	2431-2531	2532-2552	2553-2596	2597-2641
Grade 7 ELA	2438-2542	2543-2560	2561-2599	2600-2648
Grade 8 ELA	2448-2550	2551-2571	2572-2603	2604-2658
Grade 9 ELA	2454-2554	2555-2576	2577-2605	2606-2664
Grade 10 ELA	2458-2566	2567-2580	2581-2605	2606-2668
Grade 11 ELA	2465-2568	2569-2584	2585-2607	2608-2675
Grade 3 Math	3395-3494	3495-3530	3531-3572	3573-3605
Grade 4 Math	3435-3529	3530-3561	3562-3605	3606-3645
Grade 5 Math	3478-3562	3563-3594	3595-3634	3635-3688
Grade 6 Math	3512-3601	3602-3628	3629-3662	3663-3722

[54] Standard 5.2 – The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

Test	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient
Grade 7 Math	3529-3628	3629-3651	3652-3679	3680-3739
Grade 8 Math	3566-3649	3650-3672	3673-3704	3705-3776
Algebra I	3577-3660	3661-3680	3681-3719	3720-3787
Geometry	3609-3672	3673-3696	3697-3742	3743-3819
Algebra II	3629-3689	3690-3710	3711-3750	3751-3839

9.4 LINKING PAPER AND ONLINE TEST SCORES (MODE COMPARABILITY)

Prior to reporting test scores for the spring 2015 and spring 2016 administrations of AzMERIT, AIR and ADE performed mode comparability studies to evaluate differences in test performance attributable to the mode of test administration.[55]

9.4.1 MODE LINKING

A matched samples design (Way, Davis, and Fitzpatrick, 2006) to investigate mode comparability. A covariate regression approach was implemented to construct equivalent groups of students taking the AzMERIT assessments for both modes of test administration. For the spring 2015 mode investigation, the regression analysis identified for each student a predicted score on the paper AzMERIT assessment from previous year achievement on AIMS, covarying demographic variables that included gender, ethnicity, income level status, English language learner (ELL) status, and Individualized Education Program (IEP) in the development of the prediction equation. A nearest neighbor search procedure was then applied to the predicted AzMERIT scores to select the equivalent groups of students. This procedure resulted in the identification of two matched samples for each assessment to conduct the mode comparability study.

IRT parameter estimates were then calibrated independently for the matched online and paper test administration mode samples. The linking constant necessary to bring the matched sample paper item parameters onto the matched sample online scale was then computed. Mean-mean linking was taken as the difference between the average item difficulty estimates from the matched-sample paper calibration and the average item difficulty estimates from the matched-sample online item parameter estimates.

Mode linking constants were estimated again following the spring 2016 administration of AzMERIT. Three approaches were used to identify matched samples for these analyses. In the first approach, 2014 AIMS paper test scores were used to predict student performance on the spring 2016 paper tests, with the resulting prediction model then used to identify a matched sample of online test takers. This approach allowed all available paper records to be included in the analysis, but required constructing matched samples based on achievement scores estimated two years prior. To utilize a more recent and comparable test score, a second approach was used. In this approach, we identified students who were administered AzMERIT on paper in 2015, but who participated online in spring 2016. We then identified a matched sample of students, based on AzMERIT test scores, who took the paper version of AzMERIT in both 2015 and 2016. For students at grade 3, there were no previous test scores with

[55] Standard 5.13 – When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.

which to match student ability. We therefore used student performance on the multiple-choice items only on the spring 2016 AzMERIT math test to identify matched samples on the assumption that those items would be least susceptible to mode differences. To evaluate whether this approach yields results consistent with the other approaches, this approach was also applied to the grade 4 and grade 5 assessments.

Exhibit 9.4.1 presents the mode linking constants for the ELA assessments resulting from the matched sample analysis conducted on the spring 2015 administration of AzMERIT, as well as the linking constants resulting from each of the matched sample approaches used following the spring 2016 administration. In the grade 4 through 8 assessments, whether the matched samples are based on spring 2014 AIMS or spring 2015 AzMERIT, the obtained mode linking constants are generally small and equivalent across methods. For the high school end-of-course assessments, both approaches indicate that ELA assessments were somewhat more difficult online than on paper. The magnitude of those differences is greater when matching achievement based on 2014 AIMS than 2015 AzMERIT. We note that the R^2 for the prediction equation used to identify matched samples for ELA based on 2014 AIMS remained quite high even for the high school assessments, although matching based on spring 2015 AzMERIT achievement may nevertheless be more robust.

For grade 3 ELA, samples were matched based on student performance on the concurrently administered AzMERIT math multiple-choice items. To evaluate whether this approach yielded results consistent with the other two methods, we applied the same procedure in grades 4 and 5, where results indicated general convergence with the other methods, and indicating no effect for mode at grade 4 and a moderate mode effect at grade 5. When applied at grade 3, no effect for mode was identified.

We note that any effect of mode seems to interact with items, with some items easier when administered online, while others are more difficult. Thus, the effect of mode is likely to be form specific and vary across test administrations. And this seems to be the case when mode linking constants are compared between the 2015 and 2016 administrations of AzMERIT. As shown in Exhibit 9.4.1, in spring 2015, mode effects were observed in grades 3, 4, and 8, but were more moderate at the other grades. In spring 2016, however, mode effects were absent or moderate in grades 3 through 8, but appear in the high school end of course tests.

Exhibit 9.4.1 Mode Linking Constants for AzMERIT ELA Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3E	2015	0.13	-0.01	0.13	3.90
	2016 – Math MC Match	0.17	0.16	0.01	0.30
G4E	2015	-0.09	-0.19	0.11	3.30
	2016 – 2014 AIMS Match	0.21	0.19	0.02	0.60
	2016 – 2015 AzMERIT Match	0.21	0.18	0.03	0.90
	2016 – Math MC Match	0.21	0.21	0.00	0.00
G5E	2015	0.04	-0.02	0.06	1.80
	2016 – 2014 AIMS Match	0.02	-0.02	0.04	1.20
	2016 – 2015 AzMERIT Match	0.03	-0.02	0.05	1.50
	2016 – Math MC Match	0.04	-0.04	0.08	2.40
G6E	2015	0.07	-0.02	0.09	2.70
	2016 – 2014 AIMS Match	0.18	0.21	-0.03	-0.90
	2016 – 2015 AzMERIT Match	0.20	0.16	0.04	1.20
G7E	2015	-0.08	-0.16	0.08	2.40
	2016 – 2014 AIMS Match	0.19	0.12	0.07	2.10
	2016 – 2015 AzMERIT Match	0.12	0.05	0.07	2.10
G8E	2015	-0.04	-0.22	0.18	5.40
	2016 – 2014 AIMS Match	0.01	-0.01	0.02	0.60
	2016 – 2015 AzMERIT Match	0.00	-0.05	0.05	1.50
G9E	2015	0.13	0.09	0.04	1.20
	2016 – 2014 AIMS Match	0.07	-0.12	0.20	6.00
	2016 – 2015 AzMERIT Match	0.08	-0.16	0.24	7.20
G10E	2015	-0.03	-0.10	0.07	2.10
	2016 – 2014 AIMS Match	0.10	-0.10	0.20	6.00
	2016 – 2015 AzMERIT Match	0.09	-0.04	0.13	3.90
G11E	2015	0.12	0.15	-0.03	-0.90
	2016 – 2014 AIMS Match	0.16	-0.09	0.25	7.50
	2016 – 2015 AzMERIT Match	0.14	-0.04	0.18	5.40

Exhibit 9.4.2 presents the mode linking constants computed for the spring 2015 and spring 2016 administrations of the AzMERIT math assessments. As observed for ELA, in the grade 4 through 8, and Algebra I math assessments, whether the spring 2016 matched samples were based on spring 2014 AIMS or spring 2015 AzMERIT, the obtained mode linking constants are generally equivalent across methods. Effects of mode varied across grades, with the online form somewhat easier than paper at grade 4, somewhat more difficult at grade 7, and about the same at grades 5, 6, and 8. For the high school end-of-course assessments, both approaches indicate that math assessments were somewhat more difficult online than on paper. As with ELA, the magnitude of those differences was greater when matching achievement based on 2014 AIMS than 2015 AzMERIT. In this case we note that the R^2

for the prediction equation used to identify matched samples for math based on 2014 AIMS remained quite a bit lower (R^2≈.40) for the high school assessments compared to the lower grades (R^2≈.65), so that matching based on spring 2015 AzMERIT achievement are likely more robust.

For grade 3 math assessment, samples were matched based on student performance on the multiple-choice items on the assumption that those items would be relatively immune to mode effects. To evaluate whether this approach yielded results consistent with the other two methods, we applied the same procedure in grades 4 and 5, where results indicated general convergence with the other methods, indicating that items administered online were somewhat easier at grade 4 and no mode effect at grade 5. When applied at grade 3, a relatively large effect for mode was identified, indicating that items administered online were easier than on paper.

As with ELA, the identified mode effects varied across test administrations. The advantage of online over paper identified in 2016 was not observed in 2015. Likewise, observed effects of mode at grade 7 and for Algebra I and Algebra II in 2016 were not as pronounced in 2015, while effects of mode observed at grade 8 in 2015 were not observed in 2016. Thus, as with ELA, the effect of mode appears to be form specific and can be expected to vary across test administrations.

Exhibit 9.4.2 Mode Linking Constants for AzMERIT Math Assessments

Test	Matching Method	Mean_Online	Mean_Paper	Mode Linking	
				Theta Score Difference	Scale Score Difference
G3M	2015	-0.71	-0.77	0.06	1.80
	2016 – Math MC Match	-0.84	-0.57	-0.27	-8.10
G4M	2015	-0.40	-0.48	0.08	2.40
	2016 – 2014 AIMS Match	-0.43	-0.25	-0.17	-5.10
	2016 – 2015 AzMERIT Match	-0.57	-0.43	-0.14	-4.20
	2016 – Math MC Match	-0.41	-0.24	-0.17	-5.10
G5M	2015	-0.09	-0.09	-0.01	-0.30
	2016 – 2014 AIMS Match	-0.06	-0.02	-0.04	-1.20
	2016 – 2015 AzMERIT Match	-0.16	-0.12	-0.03	-0.90
	2016 – Math MC Match	-0.07	-0.06	0.00	0.00
G6M	2015	0.07	0.01	0.07	2.10
	2016 – 2014 AIMS Match	-0.01	0.04	-0.05	-1.50
	2016 – 2015 AzMERIT Match	-0.09	-0.06	-0.03	-0.90
G7M	2015	0.15	0.07	0.08	2.40
	2016 – 2014 AIMS Match	0.18	0.07	0.11	3.30
	2016 – 2015 AzMERIT Match	0.11	-0.03	0.14	4.20
G8M	2015	0.43	0.32	0.11	3.30
	2016 – 2014 AIMS Match	0.56	0.55	0.00	0.00
	2016 – 2015 AzMERIT Match	0.47	0.47	0.01	0.30
Alg I	2015	0.29	0.23	0.05	1.50
	2016 – 2014 AIMS Match	0.64	0.51	0.13	3.90
	2016 – 2015 AzMERIT Match	0.72	0.57	0.15	4.50

Geo	2015	1.12	0.99	0.13	3.90
	2016 – 2014 AIMS Match	1.34	1.15	0.20	6.00
	2016 – 2015 AzMERIT Match	1.19	1.03	0.16	4.80
Alg II	2015	1.45	1.36	0.09	2.70
	2016 – 2014 AIMS Match	1.45	1.17	0.28	8.40
	2016 – 2015 AzMERIT Match	1.06	0.91	0.15	4.50

9.4.2 SCHOOL PERFORMANCE

In a separate approach to evaluating mode comparability, ADE implemented an investigation based on the spring 2015 operational test administration statewide (Scott, 2015). In her study, Scott (2015) first identified which Arizona schools elected to administer AzMERIT online and which on paper, and then examined the two samples of schools for any differences in performance on the spring 2014 paper administration of AIMS. Having found no difference in mean 2014 performance between the two groups, there would be no expectation for performance differences on AzMERIT except as a function of test administration mode. Following the spring 2015 administration of AzMERIT, ADE examined the performance of schools participating online and on paper, and again found performance on the AzMERIT to be comparable between the two sets of schools.

9.5 LINKING THE AZMERIT TO OTHER SCALES FOR PERFORMANCE COMPARISON

9.5.1 ESTABLISHING LINKAGES TO AIMS, SAGE, SMARTER BALANCED, PISA

To facilitate comparisons of Arizona achievement to other national and international benchmarks, a number of external linking sets were embedded in the 2015 AzMERIT field test pools. Arizona identified the locations of performance standards of other assessments systems on the AzMERIT scale; this information was used to inform panelists recommending performance standards for the AzMERIT.[56] The location of performance standards from the following assessments were identified on the AzMERIT scale:

- Smarter Balanced, by linking to AIR Core items on the Smarter scale,
- PISA, by embedding PISA items in the grade 10 ELA, Algebra I and Geometry EOC assessments
- historical Arizona performance by embedding AIMS items to link to the AIMS scale, and
- Utah's SAGE via common items in the operational test form.

Subsequent to calibration of the AzMERIT operational items and establishment of the reference scale, parameter estimates for those items were anchored to their reference values and all items administered in the embedded

[56] Standard 5.23 – When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

field test (EFT) blocks were calibrated under that constraint, placing parameter estimates for all field test and external linking item sets on the same AzMERIT scale defined by the operational item parameters. All external linking items had two sets of item parameters: a) external scale, and b) AzMERIT scale. To identify the location of external scale performance standards on the AzMERIT scale, AIR identified the linking constants necessary to transform item parameters from the external reference scale to the AzMERIT scale. Where the external scale was calibrated using the Rasch model, such as with AIMS, mean-sigma equating was used to identify the location of external performance standards on the AzMERIT scale. For external scales calibrated using more general IRT models, Stocking-Lord equating was used to identify the location of external scale performance standards on the AzMERIT scale.

In the context of standard setting, this procedure enabled ADE to identify a location in the AzMERIT OIB that represented a level of difficulty similar to a particular level in the external scale. For example, after finding the linking constant necessary to put the Smarter Balanced item parameters on the AzMERIT scale, it was possible to provide standard setting panelists with the location in the OIB that represents the level of difficulty comparable to each performance standard on the Smarter Balanced assessment.

9.5.2 IDENTIFYING THE LOCATION OF THE ACT COLLEGE-READY CUT ON AZMERIT

To facilitate comparisons of Arizona achievement to other national and international benchmarks, the location of the ACT college ready cuts were identified on the AzMERIT scale and provide to panelists during performance standards workshops in 2015. In order to identify the location of the ACT college ready cuts for the AzMERIT End-of-Course assessments, a two-step approach was used to first identify the location of the ACT college ready benchmark on the AIMS scale, and then use the linkages between AIMS and AzMERIT to map the ACT college ready benchmark onto the AzMERIT scale(s). For this purpose, ADE provided ACT and AIMS scores for a recent cohort of students.

The sample used to investigate relationships between AzMERIT and ACT was based on records for students who took grade 11 ELA and Algebra II tests in spring 2015 and the ACT at an appropriate time for graduating in 2016. From among the full set of spring 2015 grade 11 ELA and Algebra II test takers, there are 58888 (93%) and 32945 (56%) grade 11 students, respectively. These records represent the target sample for the analyses reported in this study.

Because a large number of students did not take the ACT and the two subgroups differed systematically across demographic and achievement variables, the imputing approach is often employed to handle missing data in the analysis of the relationship between the AZMERIT scores and subsequent performance on the ACT. However, previous studies for Minnesota and Ohio show that imputing or deleting the missing records did not impact the linkage identified between their graduation tests and the ACT test. In this study, the regression model that links the AzMERIT scale score to the ACT scale score was built based on the merged ACT and AzMERIT records. Then the validation set approach (training and testing set split) was used to estimate the variability of the model fit and to estimate the error rate of the model when applied in new previously unseen data.

ELA Examinees with missing ACT or AzMERIT scale scores were removed from the merged dataset. The ACT reading scale score for the remaining 25977 students were regressed onto the applicable grade 11 ELA scale score and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE and largest adjusted R^2, was identified as the best model to predict ACT reading from prior performance on the AzMERIT ELA test:

$\hat{Y}$ = – 290.65+0.12*X1+ 0.26*X2 – 2.35*X3 – 0.79*X4 +0.57*X5 -2.32*X6 – 1.79*X7– 2.40*X8 – 1.82*X9– 2.07*X10, where

$\hat{Y}$ =ACT Reading Scale Score
X1 = AZMERIT ELA Scale Score
X2 = Female-Male Contrast
X3=American Indian-White Contrast
X4 = Multi-ethnic Contrast
X5 = Asian Contrast
X6 = Hispanic-White Contrast
X7 = African American-White Contrast
X8= Native Hawaiian-White Contrast
X9 = Free and Reduced Lunch Contrast
X10= ELL Contrast

The overall model was statistically significant (F(10, 20388) = 1704.70, p < .0001; adjusted R^2 = 0.46). Table 1 shows the estimated model parameters. Application of this regression model indicates that an AzMERIT ELA scale score 2585 is associated with the ACT reading college ready cut score of 22.

Mathematics The records with missing ACT or AzMERIT scale scores were excluded from the analysis. Then the ACT mathematics scale scores for the remaining 13,777 students were regressed onto the applicable AzMERIT Algebra II test and demographic variables. Stepwise selection was used to identify the prediction model. The following regression equation, which has the smallest AIC, smallest RMSE and largest adjusted R^2, was identified as the best model to predict ACT mathematics scores from prior performance on the AzMERIT Algebra II test:

$\hat{Y}$ = – 305.7 + 0.08*X1 -0.55*X2 – 1.55*X3 – 0.48*X4 -0.44*X5 – 1.44*X6 – 1.41*X7 – 0.83*X8– 1.22*X9– 1.57*X10, where
$\hat{Y}$ =ACT Mathematics Scale Score
X1 = AZMERIT Reading Scale Score
X2 = Female-Male Contrast
X3=American Indian-White Contrast
X4 = Multi-ethnic Contrast
X5 = Asian Contrast
X6 = Hispanic-White Contrast
X7 = African American-White Contrast
X8= Native Hawaiian-White Contrast
X9 = Free and Reduced Lunch Contrast
X10= ELL Contrast

The overall model was statistically significant (F(10, 13768) =1764.13, p < .0001; adjusted R^2 = 0.5`). Table 2 shows the estimated model parameters. Application of this regression model indicates that an AzMERIT mathematics score of 3727 is associated with the ACT mathematics college ready cut score of 22.

The validation set approach is a type of resampling methods that estimates a model error rate by holding out a subset of the data from the fitting process (the testing dataset). The model is then built using the other set of observations (the training dataset). Then the model result is applied on the testing dataset in which we can then calculate the error. In summary, this general idea allows for the model to not overfit. In this study, the training dataset contains 50% randomly selected merged records and the testing dataset has the other 50% of students. The multiple regression built by the training set yield the same AzMERIT cut scores (ELA 2585, Math 3727) as the ones from the full data model. Then the predictive model was applied onto the testing set. The Root Mean Squared Error (RMSE) was calculated as the square root of the average squared errors found between the actual ACT score point and the model fitted values. Furthermore, we repeated this sampling and model fitting process 100 times to see how the RMSE varies. For ELA, the average RMSE is 5.03 and the standard deviation of the RMSE is 0.02 across 100 replications. For mathematics, the average RMSE is 2.79 and the standard deviation is 0.02. The standard deviation of the RMSE is very small indicates that the sample selected for the modeling has no significant impact on the model fitting.

In addition, the equipercentile equating method was used to verify the linking between ACT and AzMERIT test scores. The AzMERIT scale score associated with the ACT cut score 22 is 2585.72 for ELA and 3727.46 for mathematics. This is consistent with the cut scores identified using regression models

10. CONSTRUCTED-RESPONSE SCORING

The AzMERIT assessments in ELA and math utilize a variety of item types to assess students' mastery of the Arizona College and Career Ready Standards (ACCRS). ADE leverages AIR's item scoring technology to machine-score student responses to most items, including traditional selected-response (multiple choice) item types, and machine-scored constructed response (MSCR) items types. The MSCR item types are designed to capture and score a variety of response types, such as graphing, drawing or arranging graphic regions, selecting or rearranging sentences or phrases within passages, or entering equations or words, allowing AzMERIT items to assess a wide range of student knowledge and skills. In most cases, constructed response machine-scored items that are developed for online administration are adapted for paper and responses are captured in a format that allows machine-scoring.

In addition, some constructed response items are scored by human raters, also referred to as hand-scored. To support machine scoring of essay responses, a sample of essay responses was hand-scored through verification, and those responses and scores were used to develop the statistical scoring models used to score the remaining responses, and which will be used to score all essay responses in future test administrations. In addition, math assessments that were administered on paper included a small number of items that were scored by human raters. Generally, these were items that required students to produce an equation. The reading components of the ELA assessments, both online and paper, and the math assessments administered online are machine-scored in their entirety.

AIR partners with Measurement, Incorporated (MI) to fulfill all hand-scoring requirements. AIR provides the automated electronic scoring and MI provides all hand-scoring for the AzMERIT tests. This section describes the process for configuring and validating machine rubrics and the process for handscoring, including rules, descriptions of scorer training and systems used, and mechanisms for ensuring reliability and validity of item scores.

10.1 MACHINE SCORING

10.1.1 EXPLICIT RUBRICS

As part of the item development process for machine-scored item types which are scored with explicit rubrics, a rubric validation process was enacted to verify that rubrics are implemented as intended, and responses are scored correctly. This procedure is typically conducted following the initial administration of items, usually when the item is field-tested, and allows test developers to review the intent of the rubric versus the actual behavior. Actual student responses were reviewed by test development experts, along with resulting item scores, to ensure that the rubrics functioned as intended and awarded credit appropriately. Where necessary, test developers modified machine rubrics to address insufficiencies, automatically rescoring student responses for the item, and repeating the process to finalize and approve the machine-scored rubrics. Test developers reviewed a strategic sample of responses, including responses where high achieving students scored poorly on the item, lower achieving students scored well on the item, and randomly selected responses from the population.

10.1.2 ESSAY AUTOSCORING

As part of the spring 2016 administration of AzMERIT, students were administered one of six writing tasks (three informational/explanatory and three persuasive/argumentative) in the writing component of each of the ELA online assessments. Writing tasks were randomly administered to students. Random assignment of writing tasks helped ensure that writing tasks were calibrated on a representative sample of Arizona students, and served to reduce the effects of school clustering in the obtained samples, increasing the efficiency of the samples, and thereby yielding more precise item parameter estimates.

Two approaches were used to develop the statistical models that were used to score the essay responses. For AIRCore writing tasks that were administered online in the Florida field-test (grades 8-10), ADE adopted the scoring models generated from student responses in the Florida field test administration. Because the scoring models are based on semantic and syntactic features of the text that discriminate high versus low scoring essays as determined by human raters, the models are highly generalizable.

For the grades where scoring models did not already exist (3-7 and 11), an alternative approach was employed that allowed for autoscoring to be implemented as part of the spring 2016 essay scoring. Because the ELA window is split into separate writing and reading assessment windows, with the online writing window closing several weeks prior to close of the reading test administration, the dual window afforded an opportunity to build and implement the statistical scoring models in time to meet spring reporting timelines.

To facilitate development of the scoring models, MI conducted range finding, where possible, based on student responses from the Florida assessment. The range finding process is designed to calibrate a sample of responses for scorer training, qualification, and monitoring. Responses exemplifying each score point are identified and annotated for scorer training. Additional responses are identified for use in qualifying readers for scoring and for establishing validity sets that are used to monitor reader performance. Thus, for grades 4-7 which were included in the Florida field test, range finding activities to support AzMERIT rubric scoring were completed prior to the opening of the AzMERIT assessment window.

For the grade 3 and 11 assessments, which had not been previously administered, MI pulled a sample of essay responses following the first week of the testing window with which to conduct range finding activities. The development of training materials and training of raters followed immediately so that hand scoring could begin by the end of the fourth week of the testing window.

At the end of the second week of testing, AIR drew a random sample of 2,000 responses to each of the writing tasks administered at grades 3-7 and 11 for use in building the statistical scoring models. Those responses were routed to MI for hand-scoring. Each response was double scored, with any discrepancies routed for resolution scoring.

As hand-scoring activities were completed for each writing task, and scores were uploaded to AIR, work began to develop statistical scoring models for each rubric element, and to deploy those models to the test delivery system to score all remaining essay responses. [57]

To develop the scoring models, the random sample of 2,000 responses was divided into a model building sample of 1,500 responses and a cross-validation sample of 500 responses. Model performance was evaluated on the cross-validation sample to ensure that model fit indices were not based on the model building sample, which may inflate fit indicators.

The statistical scoring models also yield an indicator of score confidence based on 1) responses with unusual features, and 2) responses scoring near rubric thresholds. For each model, a confidence threshold defined as two standard deviations below the mean confidence value for the responses in the cross validation sample was identified. Any scored response with a confidence value below the threshold was automatically routed to MI for verification scoring.

The statistical rubrics used to develop the scoring models measure a broad set of features, some of which may be item specific and "learned" from a training set. During training, these features are related to human scores through a statistical model. The resulting estimates complete a prediction equation that predicts how a human would score a response with the measured features. Statistical rubrics are, effectively, proxy measures. Although they can directly measure some aspects of writing conventions (e.g., use of passive voice, misspellings, run-on sentences), they do not make direct measures of argument structure or content relevance. Hence, although statistical rubrics often prove useful for scoring essays and even for providing some diagnostic feedback in writing, they do not develop a sufficiently specific model of the correct semantic structure to score many propositional items. Further, they cannot provide the explanatory or diagnostic information available from an explicit rubric. For example, the frequency of incorrect spellings may *predict* whether a response to a factual item is correct—higher-performing students may also have better spelling skills. Spelling may prove useful in predicting the human score, but it is not the "reason" that the human scorer deducts points. Indeed, statistical rubrics are not about explanation or reason but rather about a prediction of how a human would score the response.

As noted, the engine employs a "training set," a set of essay responses scored with maximally valid scores, which we obtain by having all responses double-scored by expert scorers and a thorough adjudication process for adjacent or discrepant scores. The quality of the human assigned scores is critical to the identification of a valid model and final performance of the scoring engine. Approximately 1500 essay responses were selected at random from the set of scored essay responses to serve as the training set.

---

[57] Standard 4.19 – When automated algorithms are to be used to score complex examinee responses, characteristics of responses at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms.

For each dimension in the rubric, the system estimates an appropriate statistical model relating the measures to the score assigned by humans. This model, along with its final parameter estimates, is used to generate a predicted or "proxy" score.

In addition to the training set, we draw an independent random sample of responses for cross-validation of the identified scoring rubric. As with the training set, student responses in the cross-validation study are hand-scored, and agreement between human- and machine-assigned scores is examined. The cross-validation process ensures that the rubric generalizes across all responses and that the statistical model identified during training does not capitalize on peculiarities in the training set.

Exhibit 10.1.2.1 presents agreement indicators for the two initial human raters, and between the resolved human and statistical rubric score. [58] Indicators include percent exact agreement, Pearson's correlation, a quadratic weighted kappa statistic, and the standardized mean difference between the scores. Although absolute values for evaluating statistics have been advanced (Condon, 2013; Higgins, 2013), the focus of these comparisons is degradation of agreement when moving from human-human agreement to machine-human agreement. Agreement between human raters is an indicator of how reliably the responses can be scored by human raters. Since the statistical rubrics attempt to reproduce human assigned scores, evaluation of machine-human agreement is with respect to observed human-human agreement. Responses with poor human-human agreement will not be reliably scored by either humans or machines. Exhibit 10.1.2.2 presents the correlations among dimension scores for the summative and interim tests.

[58] Standard 6.8 – Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

Exhibit 10.1.2.1 Summary of Human and Machine Scores for Spring 2016 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ*	SMD*	% Exact	Pearson r	Weighted κ*	SMD*
3	13021	Conventions	2	2092	1.43	1.55	0.75	0.71	0.69	0.65	0.65	0.03	0.72	0.71	0.70	0.16
		Evidence	4		1.93	1.90	0.78	0.61	0.65	0.64	0.64	0.02	0.65	0.65	0.63	0.05
		Organization	4		1.93	2.00	0.76	0.66	0.66	0.67	0.67	0.00	0.67	0.66	0.65	0.10
3	13022	Conventions	2	2093	1.47	1.61	0.69	0.64	0.71	0.67	0.67	0.02	0.72	0.67	0.66	0.20
		Evidence	4		2.02	2.02	0.75	0.63	0.60	0.64	0.64	0.00	0.67	0.66	0.65	0.00
		Organization	4		2.12	2.10	0.74	0.66	0.64	0.68	0.68	0.00	0.68	0.65	0.64	0.03
3	13023	Conventions	2	2090	1.51	1.57	0.72	0.65	0.70	0.65	0.65	0.01	0.73	0.67	0.66	0.09
		Evidence	4		1.89	1.93	0.74	0.65	0.62	0.60	0.60	0.03	0.68	0.68	0.67	0.05
		Organization	4		1.95	1.92	0.77	0.61	0.64	0.66	0.66	0.02	0.68	0.66	0.64	0.04
3	13024	Conventions	2	2096	1.44	1.53	0.70	0.67	0.71	0.66	0.66	0.01	0.76	0.68	0.67	0.13
		Evidence	4		1.93	1.90	0.76	0.64	0.63	0.64	0.64	0.04	0.64	0.63	0.62	0.04
		Organization	4		1.96	1.96	0.80	0.65	0.63	0.66	0.66	0.05	0.64	0.64	0.63	0.00
3	13025	Conventions	2	2093	1.37	1.48	0.76	0.70	0.66	0.58	0.58	0.01	0.71	0.67	0.66	0.15
		Evidence	4		1.94	1.97	0.72	0.61	0.63	0.65	0.65	0.00	0.69	0.65	0.64	0.04
		Organization	4		1.92	1.86	0.82	0.71	0.61	0.64	0.64	0.01	0.64	0.68	0.68	0.08
3	13026	Conventions	2	2090	1.45	1.55	0.73	0.67	0.71	0.66	0.66	0.00	0.75	0.68	0.67	0.15
		Evidence	4		1.94	1.94	0.74	0.68	0.66	0.68	0.68	0.04	0.72	0.71	0.71	0.01
		Organization	4		2.04	2.02	0.80	0.71	0.64	0.68	0.68	0.05	0.63	0.65	0.65	0.03
4	13094	Conventions	2	2095	0.95	0.95	0.75	0.68	0.66	0.67	0.67	0.00	0.65	0.65	0.65	0.01
		Evidence	4		1.30	1.27	0.47	0.47	0.77	0.52	0.52	0.00	0.82	0.58	0.58	0.08
		Organization	4		1.40	1.34	0.51	0.49	0.74	0.56	0.56	0.01	0.83	0.66	0.66	0.11
4	13095	Conventions	2	2096	1.17	1.17	0.67	0.63	0.64	0.62	0.62	0.01	0.67	0.59	0.59	0.01
		Evidence	4		1.35	1.24	0.53	0.45	0.75	0.57	0.57	0.00	0.81	0.63	0.60	0.22
		Organization	4		1.54	1.51	0.59	0.54	0.71	0.59	0.59	0.03	0.73	0.56	0.56	0.06
4	13118	Conventions	2	2096	1.15	1.16	0.71	0.65	0.64	0.60	0.60	0.01	0.67	0.63	0.63	0.01
		Evidence	4		1.33	1.29	0.49	0.48	0.76	0.55	0.55	0.01	0.84	0.64	0.64	0.07
		Organization	4		1.56	1.53	0.61	0.56	0.71	0.59	0.59	0.03	0.77	0.67	0.67	0.04
4	13119	Conventions	2	2094	1.15	1.19	0.72	0.63	0.66	0.64	0.64	0.02	0.66	0.63	0.63	0.06
		Evidence	4*		1.38	1.30	0.54	0.49	0.73	0.53	0.53	0.05	0.77	0.57	0.56	0.14

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ*	SMD*	% Exact	Pearson r	Weighted κ*	SMD*
		Organization	4		1.51	1.46	0.60	0.52	0.72	0.60	0.60	0.01	0.75	0.60	0.60	0.10
4	13120	Conventions	2	2091	1.05	1.09	0.70	0.67	0.67	0.66	0.66	0.02	0.68	0.64	0.64	0.06
		Evidence	4		1.28	1.20	0.49	0.42	0.77	0.54	0.54	0.04	0.85	0.65	0.63	0.17
		Organization	4		1.49	1.43	0.58	0.53	0.74	0.63	0.63	0.03	0.79	0.65	0.64	0.11
4	13121	Conventions	2	2096	1.10	1.08	0.69	0.59	0.67	0.65	0.65	0.03	0.68	0.61	0.60	0.02
		Evidence	4*		1.34	1.27	0.54	0.49	0.77	0.60	0.60	0.03	0.81	0.65	0.64	0.14
		Organization	4*		1.53	1.45	0.58	0.54	0.72	0.61	0.61	0.03	0.74	0.59	0.59	0.13
5	13236	Conventions	2	2099	1.41	1.57	0.67	0.62	0.74	0.69	0.69	0.02	0.76	0.69	0.67	0.25
		Evidence	4		1.81	1.79	0.58	0.52	0.71	0.59	0.59	0.01	0.79	0.64	0.64	0.03
		Organization	4		1.92	1.88	0.68	0.58	0.70	0.65	0.65	0.03	0.73	0.67	0.66	0.05
5	13237	Conventions	2	2095	1.30	1.40	0.74	0.67	0.73	0.72	0.71	0.04	0.73	0.69	0.68	0.13
		Evidence	4		1.59	1.53	0.60	0.53	0.73	0.61	0.61	0.04	0.76	0.62	0.62	0.09
		Organization	4		1.75	1.75	0.66	0.57	0.72	0.66	0.66	0.01	0.72	0.64	0.64	0.01
5	13238	Conventions	2	2099	1.47	1.51	0.62	0.61	0.72	0.65	0.65	0.00	0.75	0.65	0.64	0.06
		Evidence	4		1.87	1.88	0.64	0.53	0.69	0.63	0.63	0.01	0.75	0.63	0.62	0.02
		Organization	4		1.95	1.99	0.68	0.56	0.70	0.65	0.65	0.01	0.74	0.62	0.61	0.06
5	13239	Conventions	2	2095	1.41	1.51	0.69	0.60	0.73	0.66	0.66	0.02	0.75	0.68	0.67	0.15
		Evidence	4		1.67	1.67	0.62	0.56	0.65	0.56	0.56	0.02	0.74	0.63	0.63	0.01
		Organization	4		1.92	1.93	0.64	0.52	0.71	0.65	0.65	0.02	0.76	0.63	0.61	0.03
5	13246	Conventions	2	2093	1.36	1.45	0.68	0.65	0.72	0.68	0.68	0.01	0.73	0.69	0.69	0.13
		Evidence	4		1.54	1.58	0.57	0.54	0.72	0.59	0.59	0.03	0.77	0.61	0.60	0.07
		Organization	4		1.81	1.82	0.66	0.57	0.71	0.65	0.65	0.01	0.73	0.64	0.64	0.02
5	13247	Conventions	2	2097	1.38	1.43	0.68	0.63	0.72	0.67	0.67	0.01	0.75	0.71	0.71	0.07
		Evidence	4		1.77	1.80	0.67	0.59	0.65	0.60	0.60	0.02	0.73	0.65	0.65	0.05
		Organization	4		2.00	1.97	0.69	0.57	0.69	0.66	0.66	0.02	0.72	0.63	0.62	0.04
6	13304	Conventions	2	2097	1.43	1.52	0.67	0.62	0.67	0.57	0.57	0.02	0.76	0.72	0.71	0.14
		Evidence	4		1.74	1.75	0.65	0.61	0.63	0.57	0.56	0.04	0.74	0.66	0.66	0.02
		Organization	4		1.89	1.86	0.74	0.65	0.62	0.61	0.61	0.01	0.69	0.68	0.67	0.04
6	13305	Conventions	2	2095	1.45	1.59	0.68	0.61	0.66	0.57	0.57	0.00	0.76	0.69	0.67	0.22
		Evidence	4		1.53	1.43	0.60	0.55	0.70	0.58	0.58	0.01	0.74	0.61	0.60	0.17
		Organization	4		1.62	1.60	0.68	0.62	0.65	0.59	0.59	0.01	0.70	0.62	0.62	0.02

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ*	SMD*	% Exact	Pearson r	Weighted κ*	SMD*
6	13306	Conventions	2	2097	1.47	1.54	0.69	0.64	0.71	0.64	0.64	0.02	0.75	0.68	0.67	0.11
		Evidence	4		1.67	1.63	0.64	0.57	0.65	0.55	0.55	0.00	0.71	0.61	0.60	0.08
		Organization	4		1.85	1.80	0.69	0.61	0.64	0.62	0.61	0.04	0.71	0.67	0.66	0.07
6	13307	Conventions	2	2095	1.36	1.42	0.69	0.65	0.66	0.64	0.64	0.04	0.72	0.68	0.68	0.09
		Evidence	4		1.54	1.52	0.68	0.65	0.67	0.62	0.62	0.04	0.72	0.64	0.64	0.03
		Organization	4		1.78	1.80	0.74	0.63	0.62	0.61	0.61	0.02	0.68	0.65	0.64	0.02
6	13308	Conventions	2	2097	1.41	1.50	0.67	0.62	0.62	0.54	0.54	0.06	0.74	0.68	0.67	0.13
		Evidence	4		1.46	1.37	0.62	0.57	0.69	0.57	0.57	0.03	0.71	0.59	0.58	0.15
		Organization	4		1.64	1.57	0.69	0.62	0.63	0.60	0.60	0.03	0.71	0.66	0.65	0.10
6	13309	Conventions	2	2093	1.39	1.48	0.65	0.56	0.68	0.58	0.58	0.06	0.76	0.68	0.67	0.15
		Evidence	4		1.69	1.60	0.73	0.67	0.65	0.59	0.59	0.02	0.72	0.71	0.70	0.13
		Organization	4		1.84	1.83	0.78	0.69	0.61	0.62	0.62	0.01	0.70	0.71	0.71	0.01
7	13400	Conventions	2	2082	1.35	1.45	0.66	0.63	0.70	0.67	0.67	0.02	0.74	0.70	0.69	0.14
		Evidence	4		1.84	1.83	0.61	0.53	0.66	0.60	0.60	0.07	0.77	0.65	0.65	0.03
		Organization	4		1.92	1.90	0.64	0.54	0.65	0.62	0.62	0.02	0.74	0.61	0.60	0.03
7	13401	Conventions	2	2084	1.65	1.72	0.56	0.49	0.79	0.62	0.62	0.04	0.80	0.64	0.63	0.14
		Evidence	4		1.86	1.87	0.58	0.50	0.72	0.63	0.63	0.01	0.79	0.64	0.63	0.03
		Organization	4		2.00	2.02	0.54	0.48	0.73	0.59	0.59	0.02	0.83	0.66	0.65	0.05
7	13402	Conventions	2	2088	1.49	1.55	0.63	0.62	0.69	0.60	0.60	0.03	0.75	0.67	0.67	0.10
		Evidence	4		1.83	1.87	0.51	0.43	0.73	0.59	0.59	0.04	0.88	0.74	0.72	0.08
		Organization	4		1.91	1.93	0.59	0.50	0.70	0.61	0.61	0.01	0.80	0.66	0.65	0.02
7	13403	Conventions	2	2085	1.56	1.62	0.57	0.55	0.77	0.61	0.61	0.03	0.81	0.70	0.70	0.11
		Evidence	4		1.65	1.58	0.60	0.57	0.73	0.66	0.66	0.00	0.80	0.72	0.72	0.12
		Organization	4		1.75	1.75	0.64	0.56	0.68	0.61	0.61	0.02	0.78	0.69	0.69	0.00
7	13405	Conventions	2	2093	1.46	1.48	0.61	0.62	0.75	0.63	0.63	0.03	0.77	0.68	0.68	0.02
		Evidence	4		1.63	1.66	0.59	0.62	0.74	0.70	0.70	0.03	0.79	0.72	0.72	0.04
		Organization	4		1.83	1.80	0.62	0.56	0.73	0.68	0.68	0.03	0.79	0.70	0.69	0.05
7	13406	Conventions	2	2090	1.44	1.47	0.62	0.58	0.72	0.63	0.63	0.01	0.76	0.67	0.67	0.05
		Evidence	4		1.80	1.81	0.54	0.46	0.73	0.58	0.58	0.03	0.79	0.58	0.57	0.01
		Organization	4		1.92	1.89	0.56	0.49	0.71	0.59	0.59	0.01	0.79	0.62	0.62	0.06
8	13437	Conventions	2	2391	1.47	1.53	0.68	0.63	0.74	0.69	0.69	0.03	0.77	0.72	0.72	0.09

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ*	SMD*	% Exact	Pearson r	Weighted κ*	SMD*
		Evidence	4		1.91	1.89	0.68	0.61	0.76	0.75	0.75	0.01	0.75	0.70	0.70	0.02
		Organization	4		2.05	2.01	0.77	0.69	0.73	0.75	0.75	0.01	0.75	0.75	0.74	0.06
8	13438	Conventions	2	2631	2.01	1.95	0.77	0.71	0.79	0.70	0.70	0.01	0.73	0.75	0.74	0.08
		Evidence	4		2.11	2.08	0.80	0.76	0.77	0.78	0.78	0.00	0.71	0.75	0.74	0.04
		Organization	4		1.55	1.57	0.63	0.59	0.73	0.76	0.76	0.01	0.79	0.72	0.72	0.03
8	13439	Conventions	2	2548	1.57	1.67	0.63	0.55	0.78	0.67	0.67	0.02	0.83	0.76	0.74	0.17
		Evidence	4		2.05	2.07	0.72	0.60	0.73	0.74	0.74	0.01	0.73	0.70	0.69	0.03
		Organization	4		2.15	2.16	0.81	0.73	0.71	0.74	0.74	0.01	0.69	0.72	0.71	0.01
8	13452	Conventions	2	2491	1.61	1.65	0.58	0.54	0.79	0.67	0.67	0.02	0.80	0.68	0.68	0.06
		Evidence	4		2.07	2.06	0.75	0.64	0.77	0.77	0.77	0.01	0.74	0.73	0.72	0.02
		Organization	4		2.20	2.18	0.76	0.67	0.74	0.75	0.75	0.01	0.75	0.76	0.76	0.03
8	13453	Conventions	2	2538	1.53	1.57	0.64	0.60	0.76	0.68	0.68	0.01	0.78	0.71	0.71	0.06
		Evidence	4		1.99	1.99	0.78	0.74	0.76	0.78	0.78	0.01	0.73	0.76	0.76	0.00
		Organization	4		2.14	2.12	0.79	0.73	0.75	0.79	0.79	0.02	0.74	0.77	0.77	0.03
8	13454	Conventions	2	2544	1.56	1.58	0.61	0.56	0.77	0.68	0.68	0.01	0.78	0.68	0.68	0.03
		Evidence	4		1.99	1.91	0.74	0.67	0.77	0.77	0.77	0.01	0.74	0.73	0.73	0.10
		Organization	4		2.04	2.06	0.76	0.74	0.75	0.77	0.77	0.01	0.74	0.76	0.76	0.02
9	13554	Conventions	2	2751	1.61	1.68	0.59	0.55	0.81	0.71	0.71	0.02	0.80	0.69	0.68	0.13
		Evidence	4		1.89	1.92	0.62	0.53	0.82	0.76	0.76	0.01	0.79	0.68	0.67	0.04
		Organization	4		2.02	2.03	0.65	0.60	0.79	0.76	0.76	0.02	0.80	0.74	0.73	0.01
9	13555	Conventions	2	2853	1.58	1.68	0.63	0.57	0.81	0.74	0.74	0.02	0.81	0.76	0.74	0.17
		Evidence	4		1.88	1.90	0.66	0.59	0.81	0.77	0.77	0.02	0.79	0.72	0.72	0.03
		Organization	4		2.02	1.99	0.70	0.66	0.80	0.79	0.79	0.01	0.80	0.78	0.78	0.05
9	13556	Conventions	2	1469	1.66	1.72	0.57	0.56	0.80	0.69	0.69	0.02	0.81	0.71	0.71	0.11
		Evidence	4		1.86	1.90	0.64	0.60	0.79	0.76	0.76	0.01	0.78	0.71	0.71	0.08
		Organization	4		2.00	2.01	0.70	0.62	0.77	0.77	0.77	0.02	0.77	0.74	0.73	0.01
9	13557	Conventions	2	2815	1.54	1.58	0.65	0.60	0.79	0.73	0.73	0.00	0.78	0.71	0.70	0.06
		Evidence	4		1.82	1.85	0.57	0.53	0.83	0.78	0.78	0.01	0.83	0.72	0.72	0.05
		Organization	4		1.99	1.97	0.69	0.65	0.79	0.78	0.78	0.01	0.80	0.77	0.77	0.02
9	13565	Conventions	2	2869	1.52	1.56	0.62	0.61	0.80	0.75	0.75	0.02	0.78	0.71	0.71	0.06
		Evidence	4		1.92	1.92	0.67	0.60	0.81	0.80	0.80	0.03	0.79	0.74	0.73	0.01

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ*	SMD*	% Exact	Pearson r	Weighted κ*	SMD*
		Organization	4		2.11	2.11	0.72	0.66	0.79	0.80	0.80	0.01	0.78	0.76	0.76	0.00
9	13566	Conventions	2	2852	1.54	1.59	0.63	0.60	0.81	0.76	0.76	0.03	0.81	0.76	0.76	0.09
		Evidence	4		1.93	1.93	0.62	0.54	0.84	0.80	0.80	0.00	0.82	0.74	0.73	0.01
		Organization	4		2.08	2.09	0.67	0.63	0.79	0.79	0.79	0.02	0.82	0.77	0.77	0.00
10	13635	Conventions	2	2436	1.61	1.65	0.55	0.53	0.71	0.60	0.60	0.02	0.77	0.61	0.61	0.07
		Evidence	4		2.04	2.08	0.77	0.71	0.69	0.73	0.73	0.01	0.75	0.76	0.76	0.05
		Organization	4		2.25	2.26	0.76	0.69	0.70	0.73	0.73	0.04	0.72	0.73	0.72	0.02
10	13636	Conventions	2	2344	1.69	1.78	0.49	0.45	0.72	0.58	0.57	0.01	0.83	0.63	0.62	0.19
		Evidence	4		1.99	1.96	0.74	0.66	0.74	0.73	0.73	0.00	0.76	0.76	0.76	0.04
		Organization	4		2.06	2.08	0.75	0.72	0.72	0.74	0.74	0.00	0.79	0.81	0.81	0.03
10	13637	Conventions	2	1314	1.58	1.65	0.60	0.53	0.70	0.59	0.59	0.02	0.76	0.62	0.61	0.11
		Evidence	4		1.89	1.88	0.70	0.66	0.76	0.72	0.72	0.01	0.77	0.75	0.75	0.02
		Organization	4		2.06	2.04	0.68	0.60	0.75	0.72	0.72	0.03	0.76	0.69	0.69	0.03
10	13638	Conventions	2	2475	1.62	1.69	0.54	0.48	0.70	0.57	0.56	0.03	0.79	0.60	0.59	0.14
		Evidence	4		1.99	2.00	0.74	0.64	0.71	0.71	0.71	0.02	0.74	0.74	0.73	0.02
		Organization	4		2.12	2.14	0.76	0.70	0.69	0.71	0.71	0.01	0.77	0.79	0.78	0.03
10	13639	Conventions	2	2306	1.66	1.73	0.53	0.47	0.70	0.56	0.56	0.03	0.80	0.60	0.59	0.14
		Evidence	4		1.97	1.96	0.72	0.62	0.72	0.69	0.69	0.04	0.74	0.72	0.71	0.01
		Organization	4		2.10	2.14	0.71	0.69	0.70	0.70	0.70	0.03	0.78	0.75	0.75	0.05
10	13640	Conventions	2	2399	1.68	1.71	0.52	0.51	0.74	0.61	0.61	0.03	0.81	0.65	0.65	0.05
		Evidence	4		2.14	2.11	0.81	0.71	0.69	0.71	0.71	0.02	0.71	0.75	0.74	0.04
		Organization	4		2.26	2.29	0.79	0.72	0.67	0.71	0.71	0.04	0.73	0.76	0.76	0.04
11	13720	Conventions	2	2091	1.56	1.56	0.61	0.58	0.74	0.65	0.65	0.03	0.72	0.58	0.58	0.00
		Evidence	4		1.91	1.88	0.77	0.72	0.62	0.67	0.67	0.00	0.70	0.72	0.72	0.05
		Organization	4		2.15	2.15	0.77	0.68	0.67	0.71	0.71	0.01	0.73	0.73	0.72	0.00
11	13721	Conventions	2	2090	1.56	1.62	0.61	0.57	0.74	0.62	0.62	0.01	0.74	0.62	0.61	0.10
		Evidence	4		2.18	2.16	0.77	0.66	0.60	0.66	0.66	0.03	0.68	0.69	0.68	0.03
		Organization	4		2.36	2.35	0.71	0.64	0.66	0.67	0.66	0.03	0.76	0.74	0.73	0.01
11	13722	Conventions	2	2090	1.58	1.65	0.60	0.57	0.78	0.68	0.68	0.01	0.81	0.74	0.73	0.12
		Evidence	4		2.18	2.20	0.80	0.70	0.62	0.67	0.67	0.02	0.70	0.74	0.73	0.03
		Organization	4		2.38	2.36	0.73	0.68	0.66	0.70	0.70	0.04	0.76	0.75	0.75	0.03

Grade	ITS ID	Dimensions	Score Point	N	Mean		SD		Human-Human Agreement				Human-Machine Agreement			
					Human	Engine	Human	Engine	% Exact	Pearson r	Weighted κ*	SMD*	% Exact	Pearson r	Weighted κ*	SMD*
11	13723	Conventions	2	2095	1.60	1.63	0.59	0.57	0.75	0.61	0.61	0.01	0.77	0.65	0.65	0.05
		Evidence	4		2.24	2.24	0.83	0.74	0.62	0.70	0.70	0.02	0.70	0.74	0.73	0.00
		Organization	4		2.47	2.47	0.74	0.68	0.64	0.69	0.69	0.00	0.74	0.74	0.73	0.01
11	13724	Conventions	2	2089	1.60	1.61	0.57	0.54	0.73	0.61	0.61	0.03	0.79	0.66	0.66	0.03
		Evidence	4		2.24	2.27	0.79	0.74	0.63	0.68	0.68	0.00	0.74	0.76	0.75	0.04
		Organization	4		2.28	2.32	0.75	0.68	0.64	0.66	0.66	0.01	0.75	0.76	0.75	0.06
11	13725	Conventions	2	2085	1.45	1.50	0.67	0.63	0.71	0.64	0.64	0.04	0.78	0.74	0.73	0.08
		Evidence	4		2.21	2.26	0.83	0.78	0.64	0.72	0.72	0.02	0.74	0.79	0.79	0.05
		Organization	4		2.36	2.34	0.81	0.71	0.66	0.73	0.73	0.00	0.74	0.78	0.77	0.02

Note. Weighted K = Quadratic weighted kappa; SMD = Standardized Mean Difference

*For asterisked items, no 4-point responses were identified in the training set, so at present statistical models for these items can only assign up to three points.

Exhibit 10.1.2.2. Summary of Dimension Intercorrelations for Spring 2016 Writing Prompts

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
3	13021	Conventions	2	2092		
		Evidence	4		0.61	
		Organization	4		0.60	0.79
3	13022	Conventions	2	2093		
		Evidence	4		0.65	
		Organization	4		0.63	0.87
3	13023	Conventions	2	2090		
		Evidence	4		0.67	
		Organization	4		0.72	0.86
3	13024	Conventions	2	2096		
		Evidence	4		0.72	
		Organization	4		0.61	0.89
3	13025	Conventions	2	2093		
		Evidence	4		0.65	
		Organization	4		0.64	0.83
3	13026	Conventions	2	2090		
		Evidence	4		0.66	
		Organization	4		0.61	0.88
4	13094	Conventions	2	2095		
		Evidence	4		0.63	
		Organization	4		0.62	0.72
4	13095	Conventions	2	2096		
		Evidence	4		0.55	
		Organization	4		0.67	0.64
4	13118	Conventions	2	2096		
		Evidence	4		0.51	
		Organization	4		0.67	0.61
4	13119	Conventions	2	2094		
		Evidence	4		0.52	
		Organization	4		0.64	0.63
4	13120	Conventions	2	2091		
		Evidence	4		0.55	
		Organization	4		0.72	0.70
4	13121	Conventions	2	2096		
		Evidence	4		0.49	
		Organization	4		0.57	0.66
5	13236	Conventions	2	2099		
		Evidence	4		0.52	
		Organization	4		0.57	0.81
5	13237	Conventions	2	2095		
		Evidence	4		0.59	
		Organization	4		0.63	0.71
5	13238	Conventions	2	2099		
		Evidence	4		0.57	
		Organization	4		0.53	0.78
5	13239	Conventions	2	2095		

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
		Evidence	4		0.63	
		Organization	4		0.66	0.77
		Conventions	2			
5	13246	Evidence	4	2093	0.64	
		Organization	4		0.59	0.68
		Conventions	2			
5	13247	Evidence	4	2097	0.58	
		Organization	4		0.46	0.86
		Conventions	2			
6	13304	Evidence	4	2097	0.72	
		Organization	4		0.68	0.90
		Conventions	2			
6	13305	Evidence	4	2095	0.53	
		Organization	4		0.62	0.77
		Conventions	2			
6	13306	Evidence	4	2097	0.66	
		Organization	4		0.69	0.76
		Conventions	2			
6	13307	Evidence	4	2095	0.69	
		Organization	4		0.68	0.72
		Conventions	2			
6	13308	Evidence	4	2097	0.42	
		Organization	4		0.62	0.76
		Conventions	2			
6	13309	Evidence	4	2093	0.74	
		Organization	4		0.71	0.77
		Conventions	2			
7	13400	Evidence	4	2082	0.63	
		Organization	4		0.65	0.73
		Conventions	2			
7	13401	Evidence	4	2084	0.66	
		Organization	4		0.49	0.80
		Conventions	2			
7	13402	Evidence	4	2088	0.64	
		Organization	4		0.64	0.87
		Conventions	2			
7	13403	Evidence	4	2085	0.57	
		Organization	4		0.71	0.63
		Conventions	2			
7	13405	Evidence	4	2093	0.58	
		Organization	4		0.62	0.76
		Conventions	2			
7	13406	Evidence	4	2090	0.61	
		Organization	4		0.62	0.69
		Conventions	2			
8	13437	Evidence	4	2391	0.55	
		Organization	4		0.42	0.85
		Conventions	2			
8	13438	Evidence	4	2631	0.89	
		Organization	4		0.61	0.49

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
8	13439	Conventions	2	2548		
		Evidence	4		0.47	
		Organization	4		0.51	0.83
8	13452	Conventions	2	2491		
		Evidence	4		0.56	
		Organization	4		0.61	0.86
8	13453	Conventions	2	2538		
		Evidence	4		0.60	
		Organization	4		0.61	0.85
8	13454	Conventions	2	2544		
		Evidence	4		0.51	
		Organization	4		0.53	0.80
9	13554	Conventions	2	2751		
		Evidence	4		0.47	
		Organization	4		0.50	0.79
9	13555	Conventions	2	2853		
		Evidence	4		0.72	
		Organization	4		0.67	0.81
9	13556	Conventions	2	1469		
		Evidence	4		0.51	
		Organization	4		0.57	0.82
9	13557	Conventions	2	2815		
		Evidence	4		0.39	
		Organization	4		0.49	0.76
9	13565	Conventions	2	2869		
		Evidence	4		0.63	
		Organization	4		0.59	0.81
9	13566	Conventions	2	2852		
		Evidence	4		0.58	
		Organization	4		0.59	0.78
10	13635	Conventions	2	2436		
		Evidence	4		0.52	
		Organization	4		0.51	0.87
10	13636	Conventions	2	2344		
		Evidence	4		0.46	
		Organization	4		0.51	0.79
10	13637	Conventions	2	1314		
		Evidence	4		0.51	
		Organization	4		0.50	0.81
10	13638	Conventions	2	2475		
		Evidence	4		0.51	
		Organization	4		0.57	0.83
10	13639	Conventions	2	2306		
		Evidence	4		0.49	
		Organization	4		0.58	0.80
10	13640	Conventions	2	2399		
		Evidence	4		0.43	
		Organization	4		0.64	0.82
11	13720	Conventions	2	2091		
		Evidence	4		0.61	

Grade	ITS ID	Dimensions	Score Point	N	Correlations Among Dimensions	
					Conventions	Evidence
		Organization	4		0.51	0.79
11	13721	Conventions	2	2090		
		Evidence	4		0.61	
		Organization	4		0.54	0.79
11	13722	Conventions	2	2090		
		Evidence	4		0.63	
		Organization	4		0.63	0.80
11	13723	Conventions	2	2095		
		Evidence	4		0.59	
		Organization	4		0.63	0.77
11	13724	Conventions	2	2089		
		Evidence	4		0.70	
		Organization	4		0.66	0.88
11	13725	Conventions	2	2085		
		Evidence	4		0.65	
		Organization	4		0.62	0.87

10.2 HAND-SCORING

Hand scoring of online essay responses for statistical model building, as well as hand scoring of all essay responses from paper test administrations were routed to MI for scoring. As noted in section 10.1, the sample of essay responses selected for statistical model building was independently scored by two readers. Any response assigned discrepant scores were routed for resolution scoring by a scoring trainer. In addition, all essay responses captured from paper test administrations were hand scored, with ten percent of all paper responses receiving a second reading (Reader 2) for the purpose of monitoring and maintaining sufficient inter-rater reliability, as discussed below. For ELA hand scoring, where scores from Reader 1 and Reader 2 were not in adjacent agreement, the response was sent for resolution scoring by a Team Leader or Scoring Director. The final item score was based on the resolution score, when present, or else on the initial read. For math hand scoring, where scores from Reader 1 and Reader 2 were not in exact agreement, the response was sent for resolution scoring by a Team Leader or Scoring Director. The final item score for math was based on the resolution score, when present, or else on the initial read.

10.2.1 HANDSCORING PROCESS

MI's hand-scoring efforts are managed via the Virtual Scoring Center VSC™ software, which is composed of two primary subsystems: VSC Capture™ and VSC Score™. Images of student responses to open ended items were sent to VSC Score™, which is a web-based environment for scoring constructed- response items by scorers working in an online environment. VSC Score is a secure, centrally administered environment used by site-based scorers. The interface enabled scorers to evaluate constructed response items and writing assessments from images. VSC Score has the following capabilities:

- Defining scorer roles and qualifications based on training, security requirements, or prior history
- Managing and randomly routing scorers' responses that require second readings in a double-blind manner
- Allowing project leaders to spot-check scorers, monitor reliability, and offer feedback
- Allowing scorers to flag responses for a variety of reasons (unusual approaches, nonscorable issues, etc.)
- Generating status reports at project milestones (such as percent of items scored)

- Generating individual scorer and item statistics (such as score distribution, interscorer reliability, and non-adjacent scores)
    - Accommodating paper-based scores when images are of insufficient quality
    - Outputting data easily into MI's score reporting applications

Paper-pencil tests were scanned into VSC. The images were displayed to trained and qualified scorers who score the images online. Scorers had access only to those items for which they had been qualified to score. Online assessment responses were also converted into images and displayed in an identical manner to paper-pencil student responses using the same VSC scoring application.

When logging onto VSC Score, scorers were presented with a scoring set, which is the images-scoring equivalent of a physical packet of student responses. The scoring set was generated by randomly selecting student responses from the pool of non-scored student responses. The resultant set of responses was checked out to the scorer. The images they receive had no demographic information on them. The scorer did not know the name, sex, school, or location of the student whose item was being scored. The scorer evaluated the first response, entered the score by clicking the appropriate values on the scoring toolbar, and clicked the Submit button. For multi-page responses, a scorer had to view each page of the response before a score was entered. Once the Submit button was clicked, the system recorded the score and the next response in the scoring set appeared for the scorer to score and submit. This process continued until all responses in the set had been scored.

When a scorer had a question about a response, he or she transferred the image (along with a virtual note including the question and/or comments) from the current scoring set to a review set assigned to a team leader or the scoring director. The team leader or scoring director submitted the appropriate score or returned the response to the scorer with comments. This procedure was used whenever a scorer had scoring concerns or found nonscorable responses (NSR) or responses requiring condition codes. Previously, condition codes were assigned to student responses by scoring leadership per Arizona specifications, such as a code noting that the response was left blank, the response was undecipherable or illegible, non-English, and so on. Condition codes other than blank were then recoded to the lowest score for each dimension for ability estimation. Because the statistical scoring engine cannot assign condition codes, all non-blank responses were assigned a rubric score directly, with responses that would otherwise have received a non-blank condition code being assigned the lowest score point for each dimension.

After scoring all of the responses in a set, the scorer reviewed any of the responses and modified the scores before committing them to the system. Once the scores had been committed, the set was checked in and responses are routed to other scorers as necessary. Regardless of the specific requirements, however, student responses were not marked as complete until the requisite number of independent scorers had scored the response.

VSC prioritized the available responses in the queue to make sure that the newer responses were placed toward the back of the queue.

10.2.2 HAND-SCORING QUALITY CONTROL

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students. The requirements for double scorings are defined to VSC at setup time. MI assumed a double-blind scoring rate of 10% for both the essays and math constructed responses.

10.2.3 HAND-SCORING RELIABILITY AND VALIDITY

MI uses a two-pronged approach to construct the scoring teams for AzMERIT. First, the scoring leadership recruits qualified, experienced scorers who have successfully scored large-scale assessments for MI, and therefore have experience understanding the approach to scoring. To ensure reliable and valid hand-scores, MI puts scoring directors, team leaders, and scorers through a rigorous screening and training process.[59]

Scoring directors, team leaders, and scorers are hired for AzMERIT based on experience and performance. Potential new scorers are given a comprehensive content screening for reading and math. This screening is used to identify potential scorers' aptitude for content area and grade level as well as their reading comprehension and deductive reasoning skills, which are directly related to what they may be scoring. In addition to writing an extemporaneous essay, new hires are required to read a passage and answer questions pertaining to that passage, proofread a sample essay for writing conventions, and solve a series of math problems. The results determine grade and content area placement if a scorer is to be offered a position on a project. New scorers are selected based on their scores on MI's content screening assessment given for language arts and math projects, the quality of their interview, their work history, and the references provided. The actual qualification for the scorers occurs at the end of training. In addition, the scorers are provided with ongoing validation that they are providing the state with consistency in their scoring through the use of validation sets that are incorporated into the ongoing live scoring.

All of the Arizona training materials provided for the initial operational ELA scoring were scoring guides composed of anchor responses as well as training, qualifying, and recalibration sets approved for use by the state as a result of approval of existing documentation from AIR's Item Tracking System (ITS), which is the repository for all item attributes, including scoring rubrics. In subsequent years, new items approved from the previous year's field test will be incorporated based on the materials used during the field test scoring. All materials and selected sets were submitted to Arizona for approval.[60]

MI's scoring directors ensured ELA scoring guides had detailed annotations to explain how the scoring criteria are applied to each response's specific features and why it should be assigned a particular score. The approach was to focus on the precise scoring rationale and which helped scorers define the lines between score points. All scoring guides and other training materials were presented to Arizona for review and approval prior to the start of scoring.

Training sets and qualifying sets consisted of items that are most representative of the type that will be scored. MI scoring leadership selected these responses and provided them to Arizona for approval prior to their use. The training and qualifying sets contained examples of responses from all score points arranged in random score-point order. MI created an appropriate number of training sets and qualifying sets based on the complexity of the item. Essay questions were more complex than single-point math items. The sets were designed to help the scorers learn to apply the criteria illustrated in the scoring guide, ensure that the scorers become familiar with the process of scoring student responses, and assess the scorers' understanding of the scoring criteria before they are allowed to begin live scoring. MI worked with Arizona to finalize the number of training and qualifying sets for each item and determine

---

[59] Standard 4.20 – The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.

[60] Standard 6.8 – Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

the appropriate qualifying percentage. All scoring decisions and supplemental responses were submitted more than one month before the start of scoring for review and approval by the state.

MI's scoring directors trained both new and experienced scorers within the scoring rooms, giving detailed explanations of all training materials.

MI's online training interface allowed observers from ADE to witness training in real-time. Through the use of TurboMeeting software, observers were able to visually see the responses being trained and discussed as each training set progressed. Observers were also allowed to hear the training through the software's audio function. In addition to observing the training of leadership virtually, representatives from Arizona also traveled to individual scoring sites to observe training in-person. This allowed Arizona to observe MI's training techniques and interact with project leadership. The State was able to provide additional guidance on scoring rationale during the training process. These observations allowed MI to further ensure reliability in the hand-scoring efforts.

Recruited staff followed established training methodologies to ensure the reliability and validity of scores. Scorers were trained as a group, not individually, and all scorers (whether experienced or not) were required to train on all the scoring sets and, at the end of training, pass the qualifying sets with acceptable scores to prove that they were able to understand and apply the criteria. Unless a scorer was trained and qualified for a project successfully, he or she was not permitted to score any student responses.

Each member of MI's scoring staff was required to qualify for the scoring of student responses based on standards established by Arizona following our vigorous training process. Each staff member was also expected to maintain a consistent level of scoring quality throughout the scoring effort or he or she was released from the project. MI continually monitored performance in order to guarantee scoring accuracy.

For math, MI trained scorers to hand-score a limited number of math items from the paper assessment that could not be machine-scored. Scoring leadership reviewed all hand-scored math items prior to training. Using the scoring rubrics provided from ITS, leadership provided feedback and questions to both AIR and Arizona to ensure consistency in training methodology. Math items were trained and scored individually with the use of the provided scoring rubrics. Qualified math scorers received training that included all possible answers to each individual item.

Math hand-scoring was monitored in the same was as essay scoring, with consistent read behind and validation sets incorporated into the daily scoring schedule to ensure that scorers were providing accurate scoring on a consistent basis.

10.2.4 MACHINE-SCORING VERIFICATION

In addition to the regular ELA hand scoring activities, MI also provided a percentage of second readings on items that were machine-scored. These read behind scores were used to help ensure consistency and reliability with the ELA machine-scoring. Responses requiring read behind were generated and sent to MI where the most experienced scorers, team leaders and scoring directors provided a second read verification. This process utilized blind scoring, with the scorer unaware of the first score provided by machine. Where scores from Reader 1 (machine) and Reader 2 (human) were in exact agreement or adjacent, the final item score was based on the initial machine read. Where scores from Reader 1 (machine) and Reader 2 (human) were not in exact agreement or adjacent, the final item score was based on the second human read.

11. QUALITY ASSURANCE PROCEDURES

Quality assurance procedures are enforced through all stages of AzMERIT test development, administration, and scoring and reporting of results. This section describes quality assurance procedures associated with

- Test construction
- Test production
- Answer document processing
- Data preparation
- Equating and scaling
- Scoring and reporting

Because quality assurance procedures pervade all aspects of test development, we note that discussion of quality assurance procedures is not limited to this section, but is also included in sections describing all phases of test development and implementation.

11.1 QUALITY ASSURANCE IN TEST CONSTRUCTION

Section 5.5 details the form construction process. Each form is built to exactly match the detailed test blueprint, and match the target distribution of item difficulty and test information. Together, these constitute the definition of the instrument. The blueprint describes the content to be covered, the depth of knowledge with which it will be covered, the type of items that will measure the constructs, and every other content-relevant aspect of the test. The statistical targets ensure that students will receive scores of similar precision, regardless of which form of the test they receive.

The form construction process is managed through AIR's FormBuilder software which automates important form construction activities to ensure development of equated test forms. FormBuilder interfaces with AIR's Item Tracking System (ITS) to extract test information and interactively creates test characteristics curves (TCCs), test information curves (TICs), and Standard Error of Measurement Curves (SEMCs) as test developers build a test map. This helps our content specialists ensure that the test forms are statistically parallel, in addition to ensuring content parallelism.

Immediately upon generation of a test form, the FormBuilder generates a blueprint match report to ensure that all elements of the test blueprint have been satisfied. In addition, the FormBuilder produces a statistical summary of form characteristics to ensure consistency of test characteristics across test forms. The summary report also flags items with low biserial correlations, as well as very easy and very difficult items. Although items in the operational pool have passed through data review, construction of fixed form assessments allows another opportunity to ensure that poorly performing items are not included in operational test forms.

When submitting test forms for review by ADE, AIR produces a form evaluation workbook that includes an evaluation summary checklist, as well as summary statistics and test characteristic graphs.

All bookmaps (test maps), key files, and conversion tables were produced directly from FormBuilder to eliminate the possibility of human error in the construction of these important files. Bookmaps, key files, conversion tables, and other critical documents are generated directly from information maintained in ITS. The information stored in ITS is rigorously reviewed by multiple skilled reviewers, to protect against errors. Automated production of these critical files (such as key files) virtually eliminates opportunities for errors.

11.2 QUALITY ASSURANCE IN TEST PRODUCTION

The production of computer-delivered assessments involves two distinct types of products, each of which follows an appropriate quality assurance process:

1. Content for online delivery shares some processes with paper versions, but also requires additional, unique steps.
2. Online test delivery software must deliver the content reliably (and, with the right tools, the accommodations, layouts, etc.).

The AzMERIT test delivery system also has a real-time quality-monitoring component built in. As students are administered assessments, data flow through the test delivery system's Quality Monitor (QM) software. QM conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. QM scores the test, recalculates performance level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

QM also aggregates data to detect problems that become apparent only in the aggregate. For example, QM monitors item fit and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem.

11.2.1 PRODUCTION OF CONTENT

While the online workflow requires some additional steps, it actually removes a substantial amount of work from the time critical path, reducing the likelihood of errors. Like a test book, an online system can deliver a sequence of items; however, the online system makes the layout of that sequence algorithmic. A paper form must await final forms construction before blackline proofs can show how the item will look in the booklet. Online, the appearance of the item screen can be known with certainty before the final test form is ever constructed. This characteristic of online forms enables us to lock down the final presentation of each item well before forms are constructed. In turn, this moves the final blueline review of items much earlier in the process, removing it from the critical path.

The production of computer-based tests includes five key steps:

1. Final content is previewed and approved in a process called web approval. Web approval packages the item exactly as it will be displayed to the student.
2. Forms are finalized using the process described in Section 4.6, and final forms are approved in our FormBuilder software.
3. Complete test packages are created with our test packager, which gathers the content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.
4. Forms are initially deployed to a test site where they undergo platform review, a process during which we ensure that each item displays properly on a large number of platforms representative of those used in the field.
5. The final system is deployed to a staging environment accessible to ADE for user acceptance testing and final review.

11.2.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT

The Item Tracking System (ITS) integrates directly with the test delivery system (TDS) display module, and displays each item exactly as it will appear to the student. This process is called web preview, and web preview is tied to specific item review levels. Upon approval at those levels, the system locks content as it will be displayed to the student, transforming the item representation to the exact representation that will be rendered to the student. No change to the display content can occur without a subsequent web preview. This process freezes the display code that will present the item to the student.

Web approval functions as an item-by-item blueline review. It is the final rendering of the item as the student will see it. Layout changes can be made after this process in two ways:

1. Content can be revised and re-approved for web display.
2. Online style sheets can change to revise the layout of all items on the test.

Both of these processes are subject to strict change control protocols to ensure that accidental changes are not introduced. Below, we discuss automated quality control processes during content publication that raise warnings if item content has changed after the most recent web-approved content was generated. The web approval process offers the benefit of allowing final layout review much earlier in the process, reducing the work that must be done during the very busy period just before tests go live.

11.2.3 APPROVAL OF FINAL FORMS

Section 5.6. describes our process for constructing operational test forms, including the approval of test forms by ADE. The forms are built in FormBuilder (a component of our ITS), and upon approval, they are ready for preliminary publication.

11.2.4 PACKAGING

The test packaging system performs two simultaneous roles in the preparation of computer-based products: It compiles the form definitions and other information about how the test is to be administered (e.g., where any embedded field-test items might be inserted) and pulls together the content packaged during web approval.

The test packager assigns form identifiers to each form, evaluates the form against the blueprint, and performs a quality check against the content. The content quality check includes checks to see that every asset (e.g., graphics) referenced in the item is included in the package, confirms that the item has not changed since it was web approved, and ensures that the items have received all the approvals necessary for publication.

11.2.5 PLATFORM REVIEW

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on approximately 15 platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to see that it renders as expected.

11.2.6 USER ACCEPTANCE TESTING AND FINAL REVIEW

Prior to deployment, the testing system and content are deployed to a staging server where they are subject to user acceptance testing (UAT). UAT of the test delivery system serves both a software evaluation and content approval role. The UAT period provides the Department with an opportunity to interact with the exact test with which the students will interact.

11.2.7 FUNCTIONALITY AND CONFIGURATION

The items, both in themselves and as configured onto the tests, form one type of online product. The delivery of that test can be thought of as an independent service. Here, we document quality assurance procedures for delivering the online assessments.

One area of quality unique to online delivery is the quality of the delivery system. Three activities provide for the predictable, reliable, quality performance of our system:

1. Testing on the system itself to ensure function, performance, and capacity
2. Capacity planning
3. Continuous monitoring

AIR statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and AIR contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. Applications log not only errors and exceptions, but latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem.

In addition, latency data is captured for each assessed student—data about how long it takes to load, view, or respond to an item. All of this information is logged as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

11.3 QUALITY ASSURANCE IN DOCUMENT PROCESSING

11.3.1 SCANNING ACCURACY

When test documents are scanned, a quality control sample of documents consisted of ten test cases per document type (normally between five and six hundred documents) were created so that all possible responses and all demographic grids were verified including various typical errors that required editing via MI's Data Inspection, Correction, and Entry (DICE) application program. This structured method of testing provided exact test parameters and a methodical way of determining that the output received from the scanner(s) was correct. MI staff carefully compared the documents and the data file created from them to further ensure that results from the scanner, editing process (validation and data correction), transfer to the project database, and scoring were accurate according to the reporting rules provided by ADE.

11.3.2 QUALITY ASSURANCE IN EDITING AND DATA INPUT

At a minimum, MI implemented, maintained, and constantly updated the following quality assurance controls:

- Score key verification Post analysis of item keys
- Response analyses to determine score frequency distribution by item verification of bank values of item statistics
- Live data checks to verify that data/results conform to approved specifications comprehensive software test plan
- Double data entry correction process to verify student response and demographic information report data verification
- Reviewed and proofread all electronic and printed report deliverables

MI utilized a double data correction process to achieve the highest level of quality and accuracy in Arizona CBT and PBT assessment student data. Data correction operators used their sophisticated Data Inspection, Correction and Entry (DICE) application, which retrieved flagged data records and highlighted the problem field on a computer screen for resolution. The operator compared the highlighted data on the answer document template, retrieved the original document for resolution, and made any necessary correction.

After an operator corrected a flagged record, the same flagged record was routed to a second data correction operator who repeated the data correction process. After a flagged record was edited by two independent operators, the data correction application checked to verify that both operators made identical corrections. If the two corrections differed, the record was routed to a supervisor for a third and final resolution. Agreement rate statistics were generated for the individual data correction operators, allowing the supervisor to monitor their job performance. This process continued until all flagged records are examined and resolved.

Thorough training significantly improves the accuracy of data correction. To ensure that goal, MI trained their data correction staff on the use of the data correction application and on the specific validation errors and procedures associated with the specific project. Practice sets generated by the programming staff allowed data correction staff to learn on samples of answer documents that simulated the kinds of errors they were expected to correct for the actual assessment prior to actually processing live data. Additionally, each user had an electronic copy of the data correction user’s guide for reference.

MI developed verification routines as part of their standard data validation to detect duplicate student tests in the assessment, whether in a single LEA (local educational agency) or across LEAs, and student moves between schools. MI staff then worked closely with ADE to resolve these discrepancies through processes called Barcode Processing and Tested Roster. These processes and the business rules governing them were described in a set of requirements developed in conjunction with ADE. They involved direct data transfer in several steps between the MI and ADE databases, with the goal of ensuring that each student final report was sent to the school where the test was taken, that it had accurate demographic data, and that the test reported was the correct test per the business rules.

11.4 QUALITY ASSURANCE IN DATA PREPARATION

AIR’s test delivery system has a real-time quality-monitoring component built in. As students test, data flow through our Quality Monitor (QM) software. QM conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. QM scores the test, recalculates performance level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

QM also aggregates data to detect problems that become apparent only in the aggregate. For example, QM monitors item fit and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem. This automated process is similar to the sorts of checks that are done for data review, but (a) they are done on operational data and (b) they are conducted in real time so that our psychometricians can catch and correct any problems before they have an opportunity to do any harm.

Data pass directly from the QM to the database of record (DoR), which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator (DEG) is the tool that is used to pull data from the DoR for delivery to ADE and their quality assurance contractor. AIR psychometricians ensure that data in the extract files matches the DoR prior to delivery to ADE.

11.5 QUALITY ASSURANCE IN TEST FORM EQUATING

Item information necessary for statistical and psychometric analyses is provided to ADE and HumRRO, ADE's independent quality assurance contractor, prior to test administration. Item information is published as part of the configuration of the online assessment system that AIR employs for administering, scoring, and reporting test scores. Information contained in these workbooks includes, but is not limited to, unique item ID used for item tracking, test form ID, location on the test form, correct answer, item difficulty, and information about the strand, standard, and benchmark each item measures. These item files are used in quality control checks of the assessment data scoring and analysis.

To ensure security, all data is shared using ADE's SFTP site.

Prior to operational work, AIR produces simulated datasets for the purpose of testing software and analysis procedures, and shares with ADE and the QA contactor. All parties complete a dry run of calibration and post-equating activities and compare results. The practice runs serve two functions:

1. To verify accuracy of program code and procedures.
2. To evaluate the communication and work flow among participants. If necessary, the team will reconcile differences and correct production or verification programs.

Following the completion of these activities and resolution of questions that arise, analysis specifications are finalized.

11.6 QUALITY ASSURANCE IN SCORING AND REPORTING

11.6.1 QUALITY ASSURANCE IN HAND SCORING

DOUBLE SCORING RATES, AGREEMENT RATES, VALIDITY SETS, AND ONGOING READ-BEHINDS.

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students.

VSC provides the infrastructure for extensive quality control procedures. Through the VSC platform, project leadership can perform spot checks (read behinds) of each scorer to evaluate scoring performance; provide feedback and respond to questions; deliver retraining and/or recalibration items on demand and at regularly scheduled intervals; and prevent scorers from scoring live responses in the event that they require additional monitoring.

Once scoring is underway, quality results are achieved by consistent monitoring of each scorer. The scoring director and team leaders read behind each scorer's performance every day to ensure that he or she is on target and conduct one-on-one retraining sessions when necessary. MI's quality assurance procedures allow scoring staff to identify struggling scorers very early and begin retraining immediately.

We monitor their scoring intensively to ensure all responses are scored accurately. If through read-behinds (or data monitoring) it becomes apparent that a scorer is experiencing difficulties, he or she is given interactive feedback and mentoring on the responses that have been scored incorrectly and is expected to change the scores. Retraining is an ongoing process throughout the scoring effort to ensure more accurate scoring. Daily analyses of the scorer status reports alert management personnel to individual or group retraining needs.

If a scorer's interrater agreement rate falls below the expected standard, the scorer will be re-trained. Should the scorer still be unable to score reliably, the scorer is assigned to another, non-Arizona-related project or dismissed.

In addition to using validity responses as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. Validity responses can be culled from approved existing anchor or validity responses, but they also may be generated from live scoring and included in the pool following Arizona's review and approval. MI periodically administers validity sets to each of MI's scorers working on the scoring effort. VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whatever number of items is preferred by the state.

With the VSC program, the way in which the student responses are presented prevents scorers from having any knowledge about which responses are being single or double read, or which responses are validity set responses. A performance threshold of 75% is set to specify validity agreement standards as well as the frequency and total number of validity responses evaluated by each scorer based on client specifications.

HANDSCORING QA MONITORING REPORTS

MI generates detailed scorer status reports for each scoring project utilizing a comprehensive system for collecting and analyzing score data. The scores are validated and processed according to the specifications set out by Arizona. This allows MI to manage the quality of the scorers and take any corrective actions immediately. Updated real-time reports are available that show both daily and cumulative (project-to-date) data. These reports are available to Arizona 24 hours a day via a secure website. Project leadership reviews these reports regularly. This mechanism allows project leadership to spot-check scores at any time and offer feedback to ensure that each scorer is on target.

Scorers are released when they are unable to demonstrate the ability to score responses according to the criteria and standards established by MI and Arizona and perform to the level of client expectation. Should Arizona request that certain responses be rescored, we are prepared to do so if necessary. The reporting system can produce a list of all the responses a selected scorer has scored. In these situations, all responses scored by a scorer during the time frame in question can be identified, reset, and released back into the scoring pool. The aberrant scorer's scores are deleted, and the responses are redistributed to other qualified scorers for rescoring.

MONITORING BY ARIZONA DEPARTMENT OF EDUCATION

ADE also directly observes MI activities, both onside and virtually. MI provides virtual access to the training activities through the online training interface, as well as on-site training and on-site scoring. Arizona monitors the scoring process through the Client Command Center (CCC) with access to view and run specific reports during the scoring process. This ability to attend the training, qualification, and initial scoring virtually provides Arizona the most efficient use of oversight by reducing the travel requirements for on-site attendance for ADE staff.

IDENTIFYING, EVALUATING, AND INFORMING THE STATE ON ALERT PAPERS

MI implements a formal process for informing clients when student responses reflect a possibly dangerous situation for the examinee. We also flag potential security breaches identified during scoring. For possible dangerous situations, scoring project management and staff employ a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties.

This process is also used to notify Arizona of possible instances of teacher or proctor interference or student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow up.

11.6.2 TEST SCORING

AIR verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the state. The ability of each of these simulated students is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they provide a check of the full range of item responses and test scores in fixed-form tests as well. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the online reporting system, we merge item response data with the demographic information taken either from previous year assessment data, or if current year enrollment data is available by the time simulated data files are created, we can verify online reporting using current year testing information. By populating the simulated data files with real school information, it is possible to verify that special school types and special districts are being handled properly in the reporting system.

Specifications for generating simulated data files are included in the Analysis Specifications document submitted to the Department each year. Although ADE does not currently provide immediate reporting, review of all simulated data is scheduled to be completed prior to the opening of the test administration, so that the integrity of item administration, data capture, item and test scoring and reporting can be verified before the system goes live.

To monitor the performance of the assessment system during the test administration window, a series of Quality Assurance Reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational test window. In the context of adaptive test administrations, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to specifications.

An additional set of forensic analysis reports flags unlikely patterns of behavior in testing administrations aggregated at the test administration, test administrator, and school level that may indicate cheating. The quality assurance reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the test window to ensure that test administrations conform to blueprint and items are performing as anticipated.

Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Exhibit 11.6.2.1 presents an overview of the quality assurance (QA) reports.

Exhibit 11.6.2.1 Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology items)
Blueprint Match Rates	To monitor unexpected low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)	Early detection of any oversight in the blueprint specification
Forensic Analysis	To monitor testing irregularities	Early detection of testing irregularities

ITEM ANALYSIS REPORT

The item analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as IRT based item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

Item p-Value. For multiple-choice items, the proportion of students selecting each of response option is computed; for constructed-response, performance, and technology items, the proportion of student responses classified at each score point is computed. For multiple-choice items, if the keyed response is not the modal response, the item is also flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

Item Discrimination. Biserial correlations for the keyed response for selected-response items and polyserial correlations for polytomous constructed response, performance, and technology items are computed. AIR psychometric staff evaluates all items with biserial correlations below a target level, even if the obtained values are consistent with past item performance.

Item Fit. In addition to the item difficulty and item discrimination indices, an item fit index is produced for each item. For each student, a residual between observed and expected score given the student's ability is computed for each item. The residuals for each are averaged across all students, and the average residual is used to flag an item.

Let X_{ij} be the variable for the response of student j to item i, and $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ be the probability that student j gets a score of x_{ij} to item i given his or her ability estimate $\hat{\theta}_j$. $P\left(X_{ij} = x_{ij}\middle|\hat{\theta}_j\right)$ is calculated using three parameter logistic IRT model

$$P(X_{ij} = x_{ij}|\hat{\theta}_j) = c_i + (-c_i)\frac{\exp\left(a_i(\hat{\theta}_j - b_i)\right)}{+\exp\left(a_i(\hat{\theta}_j - b_i)\right)},$$

where a_i, b_i, and c_i are parameters of item i. If item i is a polytomously scored item, $P(X_{ij} = x_{ij}|\hat{\theta}_j)$ is calculated using the Generalized Partial Credit model,

$$P(X_{ij} = x_{ij}|\hat{\theta}_j) = \frac{\exp\sum_{k=0}^{x_{ij}} a_i(\hat{\theta}_j - b_{ki})}{\sum_{l=0}^{m_i}\exp\sum_{k=0}^{l} a_i(\hat{\theta}_j - b_{ki})}$$

The expected score for student j with estimated ability $\hat{\theta}_j$ on an item i with a maximum possible score of m_i is calculated as

$$E(X_{ij}|\hat{\theta}_j) = \sum_{x_{ij}=0}^{m_i} x_{ij}P(X_{ij} = x_{ij}|\hat{\theta}_j).$$

For item i, the residual between observed and expected score for student j is defined as

$$\delta_{ij} = x_{ij} - E(X_{ij}|\hat{\theta}_j).$$

The statistic δ_{ij} is aggregated across all n students for item i,

$$\bar{\delta}_i = \frac{}{n}\sum_{i=}^{n}(\delta_{ij}).$$

The report can be configured to report all items or flag and report only those items where the fit index is above a given threshold (e.g., items could be flagged when

$$\frac{\bar{\delta}_j}{se(\bar{\delta}_j)} > .96$$

where $se(\bar{\delta}_j) = \frac{SD(\delta_{ij})}{\sqrt{n}}$.

FORENSIC ANALYSIS

Another component in the suite of QA reports is geared toward detection testing irregularities that may indicate possible cheating. The forensic analysis component of the QA reports are described in detail in Section 6.6. Evidence evaluated includes changes in test scores across administrations, item response time, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be changed by the user. Analyses are performed at student-level and summarized for each aggregate unit, including testing session, test administrator, and school.

11.6.3 REPORTING

Scores for online assessments are assigned by automated systems in real time. For machine scored portions of assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field-testing. The review process "locks down" the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The hand-scoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Hand-scored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the quality monitor (QM) system where the integrated record is passed for scoring. Once the integrated scores are sent to the QM, the records are rescored in the test-scoring system, a mature, well-tested real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculating performance- level indicators, subscale scores and other features, which then pass automatically to the reporting system and Database of Record (DoR). The scoring system is tested extensively prior to deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

After passing through the series of validation checks in the QM system, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring there is only one place where the “official” record is stored. Only after scores have passed the QM checks and are uploaded to the DoR are they passed to the Online Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the Online Reporting System until it passes all of the QM system’s validation checks.

12. REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Council of Chief State School Officers. 2014. Criteria for Procuring and Evaluating High-Quality Assessments. Retrieved from www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(), 67–86.

Estrada, S., Burnham, C., Feld, J. K., Bergan, J. R., & Bergan, J. R. (2015). *Research Commentary: Can Local Assessment Data be Successfully Used as Part of an Arizona A-F Accountability System?* Assessment Technology Incorporated.

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*, *21*(3), 187-206.

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics*, *4*, 231-246.

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003). Separate versus concurrent calibration methods in vertical scaling. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Linacre, J. M. (2004). A user's guide to WINSTEPS: Rasch-Model Computer Program. Chicago, IL: MESA Press.

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measuremen*t, *16*(4), 247-260.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179-197.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*(2), 443-451.

McLaughlin, D., Scarloss, B. A., Stancavage, F. B., & Blankenship, C. D. (2005). Using State Assessments to Impute Achievement of Students Absent from NAEP: An Empirical Study in Four States. Washington, DC: American Institutes for Research. Retrieved from www.air.org/files/McLaughlin_AbsentStudents.pdf.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, , 115-135.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187.

Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Phillipine Statistician*, *52*(–4), 81–92.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265-276.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(), -10.

Scott, L. (2015). *Analysis of Mode Comparability of AzMERIT's Online and Paper Administrations for Spring 2015.* In Arizona Department of Education, *Recommending AzMERIT Performance Standards* (pp. I-28 – I-40), Retrieved from http://www.azed.gov/assessment/files/2014/11/spring-2015-azmerit-standard-setting_091415-full-report.pdf.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. In annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Webb, N. L. (2002). *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments for Four States.* Washington, DC: Council of Chief State School Officers.

Webb, N. (2005, April). Issues related to judging the alignment of curriculum standards and assessments. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, *27*(7), 909-921.

AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

English Language Arts Assessment Blueprint

Grade 3		
Strands	**Min**	**Max**
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 4		
Strands	**Min**	**Max**
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 5		
Strands	**Min**	**Max**
Reading Standards for Literature	26%	35%
Reading Standards for Informational Text	26%	35%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 6		
Strands	**Min**	**Max**
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 7		
Strands	**Min**	**Max**
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 8		
Strands	**Min**	**Max**
Reading Standards for Literature	24%	31%
Reading Standards for Informational Text	30%	38%
Listening Comprehension (Informational)	0%	13%
Language	13%	19%
Writing	17%	19%

Grade 9		
Strands	**Min**	**Max**
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 10		
Strands	**Min**	**Max**
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Grade 11		
Strands	**Min**	**Max**
Reading Standards for Literature	23%	30%
Reading Standards for Informational Text	31%	40%
Listening Comprehension (Informational)	0%	13%
Language	13%	18%
Writing	16%	18%

Listening Standards will only be assessed on the computer-based assessment.

In Grades 3-5 some items in the Reading and Language Strands will also be aligned to the standards for Reading: Foundational Skills.

Percentage of Points by Depth of Knowledge Level				
Grade	DOK Level 1	DOK Level 2	DOK Level 3	DOK Level 4
3-11	10%-20%	50%-60%	15%-25%	16%-19% (Writing)

For more information go to www.azed.gov/AzMERIT

AzMERIT

Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Mathematics Assessment Blueprint

Grade 3		
Domain	**Min.**	**Max.**
Operations, Algebraic Thinking, and Numbers in Base Ten	49%	53%
Number and Operations-Fractions	18%	22%
Measurement, Data, and Geometry	26%	30%

Grade 4		
Domain	**Min.**	**Max.**
Operations, Algebraic Thinking, and Numbers in Base Ten	46%	54%
Number and Operations-Fractions	29%	33%
Measurement, Data, and Geometry	15%	19%

Grade 5		
Domain	**Min.**	**Max.**
Operations, Algebraic Thinking, and Numbers in Base Ten	38%	42%
Number and Operations-Fractions	31%	35%
Measurement, Data, and Geometry	24%	28%

Grade 6		
Domain	**Min.**	**Max.**
Ratio and Proportional Relationships	19%	23%
The Number System	25%	29%
Expressions and Equations	29%	33%
Geometry, Statistics and Probability	17%	21%

Grade 7		
Domain	**Min.**	**Max.**
Ratio and Proportional Relationships	19%	23%
The Number System	19%	23%
Expressions and Equations	23%	27%
Geometry, Statistics and Probability	27%	35%

Grade 8		
Domain	**Min.**	**Max.**
Expressions and Equations	32%	36%
Functions	21%	25%
Geometry	23%	27%
Statistics and Probability and The Number System	15%	19%

Algebra I		
Conceptual Categories	**Min.**	**Max.**
Algebra	40%	44%
Functions	36%	40%
Statistics	17%	21%

Geometry		
Domain	**Min.**	**Max.**
Congruence	23%	27%
Similarity, Right Triangles and Trigonometry	27%	31%
Circles , Geometric Measurement and Geometric Properties with Equations	23%	27%
Modeling with Geometry	17%	21%

Algebra II		
Conceptual Categories	**Min.**	**Max.**
Algebra	34%	38%
Functions	32%	36%
Statistics	27%	31%

Percentage of Points by Depth of Knowledge Level			
Grade	DOK Level 1	DOK Level 2	DOK Level 3
3-11	10%-20%	60%-70%	12%-30%

Within a test, approximately 70% of the assessment will be on major content within that grade or course.

Revised by ADE on 8/19/15

For more information go to www.azed.gov/AzMERIT

Appendix B.1a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	13733.772	1804				
Metric	13956.001	1847	Configural	222.230 (43)	< .01	.000
Scalar	14638.886	1890	Metric	682.885 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	7117.742	1804				
Metric	7271.625	1847	Configural	153.883 (43)	< .01	.000
Scalar	7347.689	1890	Metric	76.064 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	11819.287	1804				
Metric	12378.423	1847	Configural	559.136 (43)	< .01	.000
Scalar	12648.918	1890	Metric	270.495 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6852.658	1804				
Metric	6900.130	1847	Configural	47.473 (43)	0.30	.000
Scalar	7101.063	1890	Metric	200.932 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	7034.814	1804				
Metric	7278.311	1847	Configural	243.498 (43)	< .01	.000
Scalar	7436.232	1890	Metric	157.921 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	7047.919	1804				
Metric	7113.779	1847	Configural	65.860 (43)	0.01	.001
Scalar	7172.388	1890	Metric	58.609 (43)	0.06	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	13509.610	1804				
Metric	13867.551	1847	Configural	357.941 (43)	< .01	.000
Scalar	14304.595	1890	Metric	437.045 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	13726.756	1804				
Metric	14038.892	1847	Configural	312.136 (43)	< .01	.000
Scalar	14103.428	1890	Metric	64.536 (43)	0.02	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	13549.468	1804				
Metric	14031.524	1847	Configural	482.057 (43)	< .01	.000
Scalar	14284.069	1890	Metric	252.545 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	13220.548	1804				
Metric	14268.173	1847	Configural	1047.625 (43)	< .01	.001
Scalar	14879.019	1890	Metric	610.846 (43)	< .01	.001

Appendix B.1b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	14638.886	1890	< .01	0.945	0.039
Model B-1	7347.689	1890	< .01	0.954	0.035
Model B-2	12648.918	1890	< .01	0.941	0.039
Model B-3	7101.063	1890	< .01	0.956	0.033
Model B-4	7436.232	1890	< .01	0.971	0.023
Model B-5	7172.388	1890	< .01	0.954	0.034
Model C	14304.595	1890	< .01	0.944	0.036
Model D	14103.428	1890	< .01	0.946	0.039
Model E	14284.069	1890	< .01	0.966	0.025
Model F	14879.019	1890	< .01	0.956	0.027

Appendix B.2a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	10163.750	1804				
Metric	10328.986	1847	Configural	165.236 (43)	< .01	.000
Scalar	10978.449	1890	Metric	649.463 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5563.647	1804				
Metric	5758.299	1847	Configural	194.652 (43)	< .01	.000
Scalar	5871.122	1890	Metric	112.823 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8716.399	1804				
Metric	9243.675	1847	Configural	527.276 (43)	< .01	.001
Scalar	9913.604	1890	Metric	669.929 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5405.843	1804				
Metric	5488.063	1847	Configural	82.220 (43)	< .01	.001
Scalar	5701.817	1890	Metric	213.754 (43)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5509.142	1804				
Metric	5806.752	1847	Configural	297.610 (43)	< .01	.001
Scalar	6041.940	1890	Metric	235.189 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5468.639	1804				
Metric	5521.446	1847	Configural	52.807	0.15	.001
Scalar	5556.291	1890	Metric	34.845	0.81	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	10234.397	1804				
Metric	10695.816	1847	Configural	461.419 (43)	< .01	.000
Scalar	11194.600	1890	Metric	498.784 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	10133.666	1804				
Metric	10554.761	1847	Configural	421.096 (43)	< .01	.001
Scalar	10709.722	1890	Metric	154.961 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	10034.335	1804				
Metric	10648.370	1847	Configural	614.035 (43)	< .01	.001
Scalar	10960.887	1890	Metric	312.516 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9881.793	1804				
Metric	11189.067	1847	Configural	1307.274 (43)	< .01	.002
Scalar	11780.373	1890	Metric	591.305 (43)	< .01	.000

Appendix B.2b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	10978.449	1890	< .01	0.947	0.036
Model B-1	5871.122	1890	< .01	0.986	0.016
Model B-2	9913.604	1890	< .01	0.968	0.024
Model B-3	5701.817	1890	< .01	0.947	0.031
Model B-4	6041.940	1890	< .01	0.984	0.016
Model B-5	5556.291	1890	< .01	0.974	0.021
Model C	11194.600	1890	< .01	0.943	0.033
Model D	10709.722	1890	< .01	0.946	0.036
Model E	10960.887	1890	< .01	0.986	0.016
Model F	11780.373	1890	< .01	0.938	0.034

Appendix B.3a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	11011.824	1804				
Metric	11305.868	1847	Configural	294.044 (43)	< .01	.000
Scalar	11780.179	1890	Metric	474.311 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	6128.336	1804				
Metric	6312.812	1847	Configural	184.476 (43)	< .01	.000
Scalar	6406.820	1890	Metric	94.008 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	9646.302	1804				
Metric	10044.148	1847	Configural	397.846 (43)	< .01	.001
Scalar	10327.027	1890	Metric	282.879 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6076.608	1804				
Metric	6143.701	1847	Configural	67.094 (43)	0.01	.000
Scalar	6278.654	1890	Metric	134.953 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	6160.021	1804				
Metric	6479.098	1847	Configural	319.077 (43)	< .01	.001
Scalar	6674.439	1890	Metric	195.341 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	6003.154	1804				
Metric	6075.136	1847	Configural	71.982 (43)	< .01	.000
Scalar	6129.819	1890	Metric	54.682 (43)	0.11	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	10971.222	1804				
Metric	11588.948	1847	Configural	617.727 (43)	< .01	.000
Scalar	12189.102	1890	Metric	600.153 (43)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	11119.735	1804				
Metric	11347.250	1847	Configural	227.515 (43)	< .01	.000
Scalar	11426.483	1890	Metric	79.233 (43)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	11049.637	1804				
Metric	11494.150	1847	Configural	444.512 (43)	< .01	.000
Scalar	11680.610	1890	Metric	186.460 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	10695.809	1804				
Metric	12027.699	1847	Configural	1331.889 (43)	< .01	.002
Scalar	12617.063	1890	Metric	589.364 (43)	< .01	.001

Appendix B.3b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	11780.179	1890	< .01	0.970	0.031
Model B-1	6406.820	1890	< .01	0.986	0.016
Model B-2	10327.027	1890	< .01	0.964	0.032
Model B-3	6278.654	1890	< .01	0.969	0.027
Model B-4	6674.439	1890	< .01	0.984	0.016
Model B-5	6129.819	1890	< .01	0.970	0.027
Model C	12189.102	1890	< .01	0.966	0.029
Model D	11426.483	1890	< .01	0.970	0.031
Model E	11680.610	1890	< .01	0.987	0.015
Model F	12617.063	1890	< .01	0.981	0.017

Appendix B.4a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	9698.728	1804				
Metric	10093.239	1847	Configural	394.510 (43)	< .01	.001
Scalar	10841.059	1890	Metric	747.820 (43)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5611.090	1804				
Metric	5743.655	1847	Configural	132.565 (43)	< .01	.000
Scalar	5860.740	1890	Metric	117.085 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8409.068	1804				
Metric	8938.388	1847	Configural	529.319 (43)	< .01	.001
Scalar	9380.459	1890	Metric	442.072 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5466.053	1804				
Metric	5516.666	1847	Configural	50.614	0.20	.000
Scalar	5655.420	1890	Metric	138.753	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5652.413	1804				
Metric	5960.932	1847	Configural	308.519 (43)	< .01	.000
Scalar	6181.560	1890	Metric	220.628 (43)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5444.331	1804				
Metric	5480.300	1847	Configural	35.969 (43)	0.77	.001
Scalar	5526.887	1890	Metric	46.587 (43)	0.33	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	9547.579	1804				
Metric	10051.098	1847	Configural	503.520 (43)	< .01	.001
Scalar	10564.126	1890	Metric	513.028 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	9673.726	1804				
Metric	10094.442	1847	Configural	420.715 (43)	< .01	.001
Scalar	10208.083	1890	Metric	113.641 (43)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	9627.312	1804				
Metric	9925.714	1847	Configural	298.403 (43)	< .01	.000
Scalar	10293.927	1890	Metric	368.213 (43)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9313.958	1804				
Metric	9980.107	1847	Configural	666.149 (43)	< .01	.000
Scalar	10761.774	1890	Metric	781.667 (43)	< .01	.001

Appendix B.4b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	10841.059	1890	< .01	0.880	0.032
Model B-1	5860.740	1890	< .01	0.966	0.030
Model B-2	9380.459	1890	< .01	0.961	0.033
Model B-3	5655.420	1890	< .01	0.989	0.013
Model B-4	6181.560	1890	< .01	0.986	0.016
Model B-5	5526.887	1890	< .01	0.969	0.027
Model C	10564.126	1890	< .01	0.870	0.031
Model D	10208.083	1890	< .01	0.890	0.030
Model E	10293.927	1890	< .01	0.880	0.031
Model F	10761.774	1890	< .01	0.986	0.015

Appendix B.5a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	7386.057	1804				
Metric	7573.876	1847	Configural	187.819 (43)	< .01	.000
Scalar	8502.762	1890	Metric	928.886 (43)	< .01	.002
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	4703.545	1804				
Metric	4856.004	1847	Configural	152.459 (43)	< .01	.000
Scalar	4973.729	1890	Metric	117.725 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	6684.172	1804				
Metric	7103.969	1847	Configural	419.798 (43)	< .01	.000
Scalar	7451.795	1890	Metric	347.826 (43)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	4718.096	1804				
Metric	4764.669	1847	Configural	46.573 (43)	0.33	.000
Scalar	4939.561	1890	Metric	174.891 (43)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	4623.233	1804				
Metric	4784.118	1847	Configural	160.885 (43)	< .01	.000
Scalar	4959.031	1890	Metric	174.913 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	4665.775	1804				
Metric	4707.603	1847	Configural	41.828 (43)	0.52	.001
Scalar	4769.442	1890	Metric	61.839 (43)	0.03	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	7391.785	1804				
Metric	7699.688	1847	Configural	307.903 (43)	< .01	.001
Scalar	8152.133	1890	Metric	452.445 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	7522.564	1804				
Metric	7856.478	1847	Configural	333.913 (43)	< .01	.000
Scalar	7923.163	1890	Metric	66.685 (43)	0.01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	7541.658	1804				
Metric	7791.811	1847	Configural	250.154 (43)	< .01	.000
Scalar	8127.541	1890	Metric	335.730 (43)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	7431.922	1804				
Metric	7840.706	1847	Configural	408.784 (43)	< .01	.001
Scalar	8519.549	1890	Metric	678.843 (43)	< .01	.001

Appendix B.5b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	8502.762	1890	< .01	0.967	0.030
Model B-1	4973.729	1890	< .01	0.988	0.013
Model B-2	7451.795	1890	< .01	0.961	0.029
Model B-3	4939.561	1890	< .01	0.966	0.024
Model B-4	4959.031	1890	< .01	0.988	0.013
Model B-5	4769.442	1890	< .01	0.969	0.023
Model C	8152.133	1890	< .01	0.989	0.013
Model D	7923.163	1890	< .01	0.966	0.028
Model E	8127.541	1890	< .01	0.990	0.012
Model F	8519.549	1890	< .01	0.988	0.013

Appendix B.6a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	14027.579	1804				
Metric	14325.678	1847	Configural	298.099 (43)	< .01	.000
Scalar	15084.733	1890	Metric	759.055 (43)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	7738.802	1804				
Metric	7889.631	1847	Configural	150.829 (43)	< .01	.000
Scalar	8035.171	1890	Metric	145.540 (43)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	12202.711	1804				
Metric	12716.273	1847	Configural	513.562 (43)	< .01	.000
Scalar	12922.089	1890	Metric	205.816 (43)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	7613.130	1804				
Metric	7667.109	1847	Configural	53.979 (43)	0.12	.000
Scalar	7793.346	1890	Metric	126.237 (43)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	7776.336	1804				
Metric	7970.933	1847	Configural	194.597 (43)	< .01	.000
Scalar	8130.334	1890	Metric	159.402 (43)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	7529.398	1804				
Metric	7562.979	1847	Configural	33.581 (43)	0.85	.001
Scalar	7616.296	1890	Metric	53.317 (43)	0.13	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	13905.703	1804				
Metric	14240.091	1847	Configural	334.388 (43)	< .01	.000
Scalar	14864.115	1890	Metric	624.024 (43)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	14078.981	1804				
Metric	14327.866	1847	Configural	248.885 (43)	< .01	.000
Scalar	14421.238	1890	Metric	93.372 (43)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	14095.956	1804				
Metric	14211.952	1847	Configural	115.996 (43)	< .01	.000
Scalar	14414.906	1890	Metric	202.954 (43)	< .01	.001
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	13861.740	1804				
Metric	14406.544	1847	Configural	544.805 (43)	< .01	.000
Scalar	15104.898	1890	Metric	698.353 (43)	< .01	.000

Appendix B.6b. Global Model Fit Indices of Scalar Invariance Model for Grade 8 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	15084.733	1890	< .01	0.948	0.044
Model B-1	8035.171	1890	< .01	0.957	0.041
Model B-2	12922.089	1890	< .01	0.951	0.044
Model B-3	7793.346	1890	< .01	0.957	0.037
Model B-4	8130.334	1890	< .01	0.958	0.040
Model B-5	7616.296	1890	< .01	0.961	0.036
Model C	14864.115	1890	< .01	0.947	0.038
Model D	14421.238	1890	< .01	0.947	0.042
Model E	14414.906	1890	< .01	0.950	0.035
Model F	15104.898	1890	< .01	0.971	0.038

Appendix B.7a. Global Model Fit Indices of Measurement Invariance Tests for Grade 9 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	9785.018	1978				
Metric	10049.839	2023	Configural	264.821 (45)	< .01	.000
Scalar	10769.979	2068	Metric	720.140 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5881.394	1978				
Metric	5997.772	2023	Configural	116.378 (45)	< .01	.000
Scalar	6098.463	2068	Metric	100.691 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	8602.375	1978				
Metric	9013.777	2023	Configural	411.402 (45)	< .01	.000
Scalar	9358.174	2068	Metric	344.397 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	5811.836	1978				
Metric	5858.039	2023	Configural	46.203 (45)	0.42	.001
Scalar	5961.451	2068	Metric	103.412 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5876.617	1978				
Metric	6018.198	2023	Configural	141.581 (45)	< .01	.000
Scalar	6266.719	2068	Metric	248.521 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5702.498	1978				
Metric	5741.660	2023	Configural	39.162 (45)	0.72	.001
Scalar	5776.327	2068	Metric	34.667 (45)	0.87	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	9819.145	1978				
Metric	10020.292	2023	Configural	201.147 (45)	< .01	.000
Scalar	10344.222	2068	Metric	323.930 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	9860.151	1978				
Metric	10147.583	2023	Configural	287.432 (45)	< .01	.000
Scalar	10258.257	2068	Metric	110.674 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	9970.516	1978				
Metric	10173.098	2023	Configural	202.582 (45)	< .01	.000
Scalar	10314.595	2068	Metric	141.497 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	9912.833	1978				
Metric	10125.393	2023	Configural	212.560 (45)	< .01	.000
Scalar	10352.306	2068	Metric	226.913 (45)	< .01	.000

Appendix B.7b. Global Model Fit Indices of Scalar Invariance Model for Grade 9 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	10769.979	2068	< .01	0.965	0.023
Model B-1	6098.463	2068	< .01	0.972	0.019
Model B-2	9358.174	2068	< .01	0.963	0.023
Model B-3	5961.451	2068	< .01	0.974	0.017
Model B-4	6266.719	2068	< .01	0.973	0.019
Model B-5	5776.327	2068	< .01	0.978	0.015
Model C	10344.222	2068	< .01	0.970	0.019
Model D	10258.257	2068	< .01	0.967	0.022
Model E	10314.595	2068	< .01	0.973	0.018
Model F	10352.306	2068	< .01	0.986	0.012

Appendix B.8a. Global Model Fit Indices of Measurement Invariance Tests for Grade 10 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	11398.967	1978				
Metric	11643.610	2023	Configural	244.644 (45)	< .01	.000
Scalar	12454.369	2068	Metric	810.759 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	6787.701	1978				
Metric	6887.953	2023	Configural	100.253 (45)	< .01	.000
Scalar	7000.945	2068	Metric	112.992 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	9962.149	1978				
Metric	10396.724	2023	Configural	434.575 (45)	< .01	.001
Scalar	10754.618	2068	Metric	357.894 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	6718.523	1978				
Metric	6794.553	2023	Configural	76.030 (45)	< .01	.000
Scalar	6955.160	2068	Metric	160.607 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	6677.657	1978				
Metric	6909.453	2023	Configural	231.797 (45)	< .01	.001
Scalar	7122.427	2068	Metric	212.974 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	6731.650	1978				
Metric	6784.798	2023	Configural	53.148 (45)	0.19	.000
Scalar	6837.360	2068	Metric	52.562 (45)	0.20	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	11516.948	1978				
Metric	11743.401	2023	Configural	226.452 (45)	< .01	.000
Scalar	12092.078	2068	Metric	348.678 (45)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	11482.165	1978				
Metric	11785.845	2023	Configural	303.680 (45)	< .01	.000
Scalar	11878.384	2068	Metric	92.539 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	11538.971	1978				
Metric	11651.848	2023	Configural	112.877 (45)	< .01	.000
Scalar	11768.221	2068	Metric	116.373 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	11479.163	1978				
Metric	11665.006	2023	Configural	185.844 (45)	< .01	.000
Scalar	11828.482	2068	Metric	163.476 (45)	< .01	.000

Appendix B.8b. Global Model Fit Indices of Scalar Invariance Model for Grade 10 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	12454.369	2068	< .01	0.962	0.036
Model B-1	7000.945	2068	< .01	0.960	0.021
Model B-2	10754.618	2068	< .01	0.947	0.025
Model B-3	6955.160	2068	< .01	0.958	0.021
Model B-4	7122.427	2068	< .01	0.960	0.021
Model B-5	6837.360	2068	< .01	0.963	0.018
Model C	12092.078	2068	< .01	0.965	0.031
Model D	11878.384	2068	< .01	0.965	0.035
Model E	11768.221	2068	< .01	0.978	0.028
Model F	11828.482	2068	< .01	0.978	0.027

Appendix B.9a. Global Model Fit Indices of Measurement Invariance Tests for Grade 11 ELA

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	7351.566	1978				
Metric	7591.812	2023	Configural	240.246 (45)	< .01	.000
Scalar	8303.823	2068	Metric	712.011 (45)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	5120.192	1978				
Metric	5209.060	2023	Configural	88.868 (45)	< .01	.000
Scalar	5319.957	2068	Metric	110.897 (45)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	6619.298	1978				
Metric	6981.561	2023	Configural	362.263 (45)	< .01	.001
Scalar	7298.593	2068	Metric	317.032 (45)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	4973.414	1978				
Metric	5009.619	2023	Configural	36.205 (45)	0.82	.000
Scalar	5129.101	2068	Metric	119.482 (45)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	5244.742	1978				
Metric	5422.163	2023	Configural	177.422 (45)	< .01	.001
Scalar	5688.003	2068	Metric	265.839 (45)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	5065.796	1978				
Metric	5119.452	2023	Configural	53.657 (45)	0.18	.000
Scalar	5145.823	2068	Metric	26.371 (45)	0.99	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	7513.769	1978				
Metric	7678.313	2023	Configural	164.544 (45)	< .01	.000
Scalar	8058.493	2068	Metric	380.180 (45)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	7488.459	1978				
Metric	7754.618	2023	Configural	266.159 (45)	< .01	.001
Scalar	7914.265	2068	Metric	159.647 (45)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	7717.892	1978				
Metric	7812.260	2023	Configural	94.368 (45)	< .01	.000
Scalar	7923.082	2068	Metric	110.822 (45)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	7704.274	1978				
Metric	7847.220	2023	Configural	142.946 (45)	< .01	.000
Scalar	7993.301	2068	Metric	146.081 (45)	< .01	.000

Appendix B.9b. Global Model Fit Indices of Scalar Invariance Model for Grade 11 ELA

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	8303.823	2068	< .01	0.959	0.034
Model B-1	5319.957	2068	< .01	0.973	0.020
Model B-2	7298.593	2068	< .01	0.964	0.023
Model B-3	5129.101	2068	< .01	0.975	0.018
Model B-4	5688.003	2068	< .01	0.972	0.021
Model B-5	5145.823	2068	< .01	0.977	0.016
Model C	8058.493	2068	< .01	0.969	0.027
Model D	7914.265	2068	< .01	0.960	0.032
Model E	7923.082	2068	< .01	0.985	0.014
Model F	7993.301	2068	< .01	0.983	0.015

Appendix B.10a. Global Model Fit Indices of Measurement Invariance Tests for Grade 3 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	178020.127	1890				
Metric	178975.724	1934	Configural	955.597 (44)	< .01	.001
Scalar	183551.597	1978	Metric	4575.874 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	74981.430	1890				
Metric	76809.764	1934	Configural	1828.334 (44)	< .01	.000
Scalar	77843.804	1978	Metric	1034.040 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	143550.580	1890				
Metric	149168.011	1934	Configural	5617.431 (44)	< .01	.000
Scalar	154151.396	1978	Metric	4983.385 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	71303.946	1890				
Metric	71673.151	1934	Configural	369.205 (44)	< .01	.001
Scalar	72140.060	1978	Metric	466.909 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	74961.905	1890				
Metric	77173.889	1934	Configural	2211.984 (44)	< .01	.000
Scalar	78790.231	1978	Metric	1616.342 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	71635.066	1890				
Metric	71792.403	1934	Configural	157.337 (44)	< .01	.000
Scalar	71909.228	1978	Metric	116.825 (44)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	170307.577	1890				
Metric	177157.840	1934	Configural	6850.263 (44)	< .01	.001
Scalar	180389.601	1978	Metric	3231.762 (44)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	175711.661	1890				
Metric	179583.439	1934	Configural	3871.778 (44)	< .01	.000
Scalar	181201.055	1978	Metric	1617.616 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	174317.255	1890				
Metric	180824.968	1934	Configural	6507.712 (44)	< .01	.000
Scalar	182097.212	1978	Metric	1272.244 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	165211.122	1890				
Metric	181890.732	1934	Configural	16679.610 (44)	< .01	.002
Scalar	184706.069	1978	Metric	2815.336 (44)	< .01	.001

Appendix B.10b. Global Model Fit Indices of Scalar Invariance Model for Grade 3 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	183551.597	1978	< .01	0.917	0.058
Model B-1	77843.804	1978	< .01	0.913	0.056
Model B-2	154151.396	1978	< .01	0.909	0.056
Model B-3	72140.060	1978	< .01	0.922	0.050
Model B-4	78790.231	1978	< .01	0.917	0.055
Model B-5	71909.228	1978	< .01	0.915	0.054
Model C	180389.601	1978	< .01	0.905	0.058
Model D	181201.055	1978	< .01	0.914	0.057
Model E	182097.212	1978	< .01	0.916	0.057
Model F	184706.069	1978	< .01	0.906	0.057

Appendix B.11a. Global Model Fit Indices of Measurement Invariance Tests for Grade 4 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	78513.118	1890				
Metric	79737.405	1934	Configural	1224.287 (44)	< .01	.000
Scalar	84155.308	1978	Metric	4417.903 (44)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	31709.452	1890				
Metric	33546.915	1934	Configural	1837.464 (44)	< .01	.000
Scalar	34341.647	1978	Metric	794.731 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	62665.155	1890				
Metric	66540.482	1934	Configural	3875.328 (44)	< .01	.001
Scalar	69648.954	1978	Metric	3108.472 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	29285.910	1890				
Metric	29504.006	1934	Configural	218.096 (44)	< .01	.000
Scalar	30032.709	1978	Metric	528.703 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	31111.359	1890				
Metric	32984.692	1934	Configural	1873.333 (44)	< .01	.001
Scalar	34001.235	1978	Metric	1016.544 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	29303.472	1890				
Metric	29364.693	1934	Configural	61.220 (44)	.044	.000
Scalar	29494.527	1978	Metric	129.835 (44)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	74429.878	1890				
Metric	80370.099	1934	Configural	5940.221 (44)	< .01	.001
Scalar	84016.074	1978	Metric	3645.976 (44)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	77747.392	1890				
Metric	80662.164	1934	Configural	2914.772 (44)	< .01	.000
Scalar	81460.601	1978	Metric	798.437 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	76537.210	1890				
Metric	82372.857	1934	Configural	5835.648 (44)	< .01	.001
Scalar	83624.630	1978	Metric	1251.772 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	71418.760	1890				
Metric	84391.883	1934	Configural	12973.123 (44)	< .01	.002
Scalar	87919.282	1978	Metric	3527.398 (44)	< .01	.001

Appendix B.11b. Global Model Fit Indices of Scalar Invariance Model for Grade 4 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	84155.308	1978	< .01	0.962	0.035
Model B-1	34341.647	1978	< .01	0.961	0.032
Model B-2	69648.954	1978	< .01	0.957	0.033
Model B-3	30032.709	1978	< .01	0.960	0.029
Model B-4	34001.235	1978	< .01	0.961	0.031
Model B-5	29494.527	1978	< .01	0.964	0.029
Model C	84016.074	1978	< .01	0.959	0.033
Model D	81460.601	1978	< .01	0.962	0.034
Model E	83624.630	1978	< .01	0.964	0.033
Model F	87919.282	1978	< .01	0.957	0.033

Appendix B.12a. Global Model Fit Indices of Measurement Invariance Tests for Grade 5 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	72135.653	1890				
Metric	73067.921	1934	Configural	932.268 (44)	< .01	.000
Scalar	76282.010	1978	Metric	3214.089 (44)	< .01	.000
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	29168.642	1890				
Metric	31046.373	1934	Configural	1877.730 (44)	< .01	.001
Scalar	31621.282	1978	Metric	574.909 (44)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	57946.466	1890				
Metric	62395.316	1934	Configural	4448.850 (44)	< .01	.001
Scalar	64315.473	1978	Metric	1920.157 (44)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	26768.137	1890				
Metric	27112.072	1934	Configural	343.935 (44)	< .01	.000
Scalar	27481.486	1978	Metric	369.413 (44)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	29519.527	1890				
Metric	32104.916	1934	Configural	2585.389 (44)	< .01	.001
Scalar	32872.831	1978	Metric	767.915 (44)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	26763.882	1890				
Metric	26846.021	1934	Configural	82.140 (44)	< .01	.001
Scalar	26943.191	1978	Metric	97.169 (44)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	68366.428	1890				
Metric	74117.853	1934	Configural	5751.424 (44)	< .01	.001
Scalar	79295.211	1978	Metric	5177.358 (44)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	71225.495	1890				
Metric	74139.995	1934	Configural	2914.500 (44)	< .01	.000
Scalar	74628.714	1978	Metric	488.719 (44)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	70880.121	1890				
Metric	74815.092	1934	Configural	3934.970 (44)	< .01	.000
Scalar	77058.496	1978	Metric	2243.404 (44)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	66416.026	1890				
Metric	77274.731	1934	Configural	10858.705 (44)	< .01	.002
Scalar	83493.323	1978	Metric	6218.592 (44)	< .01	.001

Appendix B.12b. Global Model Fit Indices of Scalar Invariance Model for Grade 5 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	76282.010	1978	< .01	0.968	0.032
Model B-1	31621.282	1978	< .01	0.968	0.029
Model B-2	64315.473	1978	< .01	0.964	0.031
Model B-3	27481.486	1978	< .01	0.968	0.027
Model B-4	32872.831	1978	< .01	0.968	0.029
Model B-5	26943.191	1978	< .01	0.970	0.027
Model C	79295.211	1978	< .01	0.963	0.031
Model D	74628.714	1978	< .01	0.967	0.031
Model E	77058.496	1978	< .01	0.968	0.030
Model F	83493.323	1978	< .01	0.962	0.031

Appendix B.13a. Global Model Fit Indices of Measurement Invariance Tests for Grade 6 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	90030.261	2068				
Metric	91811.680	2114	Configural	1781.419 (46)	< .01	.000
Scalar	98752.423	2160	Metric	6940.743 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	37453.505	2068				
Metric	40103.361	2114	Configural	2649.856 (46)	< .01	.001
Scalar	40679.803	2160	Metric	576.442 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	71421.466	2068				
Metric	78474.546	2114	Configural	7053.080 (46)	< .01	.002
Scalar	80280.423	2160	Metric	1805.877 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	35345.538	2068				
Metric	35560.721	2114	Configural	215.182 (46)	< .01	.001
Scalar	35866.572	2160	Metric	305.851 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	37208.036	2068				
Metric	39896.741	2114	Configural	2688.705 (46)	< .01	.001
Scalar	40737.236	2160	Metric	840.495 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	34775.418	2068				
Metric	34927.089	2114	Configural	151.671 (46)	< .01	.000
Scalar	34986.399	2160	Metric	59.310 (46)	.090	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	80371.915	2068				
Metric	87601.812	2114	Configural	7229.897 (46)	< .01	.001
Scalar	96412.066	2160	Metric	8810.254 (46)	< .01	.002
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	88516.533	2068				
Metric	93104.084	2114	Configural	4587.551 (46)	< .01	.001
Scalar	93599.712	2160	Metric	495.628 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	86690.679	2068				
Metric	91830.076	2114	Configural	5139.397 (46)	< .01	.000
Scalar	94418.244	2160	Metric	2588.169 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	77015.305	2068				
Metric	89657.183	2114	Configural	12641.879 (46)	< .01	.003
Scalar	98108.681	2160	Metric	8451.497 (46)	< .01	.001

Appendix B.13b. Global Model Fit Indices of Scalar Invariance Model for Grade 6 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	98752.423	2160	< .01	0.958	0.037
Model B-1	40679.803	2160	< .01	0.960	0.033
Model B-2	80280.423	2160	< .01	0.953	0.036
Model B-3	35866.572	2160	< .01	0.957	0.031
Model B-4	40737.236	2160	< .01	0.960	0.033
Model B-5	34986.399	2160	< .01	0.962	0.031
Model C	96412.066	2160	< .01	0.952	0.034
Model D	93599.712	2160	< .01	0.957	0.036
Model E	94418.244	2160	< .01	0.964	0.033
Model F	98108.681	2160	< .01	0.956	0.033

Appendix B.14a. Global Model Fit Indices of Measurement Invariance Tests for Grade 7 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	37314.410	2068				
Metric	38506.401	2114	Configural	1191.991 (46)	< .01	.000
Scalar	44483.010	2160	Metric	5976.609 (46)	< .01	.002
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	18571.581	2068				
Metric	20221.959	2114	Configural	1650.378 (46)	< .01	.001
Scalar	20875.847	2160	Metric	653.888 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	30441.823	2068				
Metric	35078.471	2114	Configural	4636.648 (46)	< .01	.001
Scalar	36819.342	2160	Metric	1740.870 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	17534.361	2068				
Metric	17791.556	2114	Configural	257.195 (46)	< .01	.000
Scalar	18399.719	2160	Metric	608.162 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	18001.839	2068				
Metric	19957.793	2114	Configural	1955.954 (46)	< .01	.001
Scalar	21012.517	2160	Metric	1054.724 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	17639.958	2068				
Metric	17735.461	2114	Configural	95.503 (46)	< .01	.000
Scalar	17822.776	2160	Metric	87.315 (46)	< .01	.001
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	35767.378	2068				
Metric	39249.591	2114	Configural	3482.213 (46)	< .01	.001
Scalar	44955.521	2160	Metric	5705.930 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	36685.569	2068				
Metric	39848.742	2114	Configural	3163.173 (46)	< .01	.001
Scalar	40105.854	2160	Metric	257.113 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	36944.840	2068				
Metric	39359.374	2114	Configural	2414.535 (46)	< .01	.001
Scalar	41609.312	2160	Metric	2249.938 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non-Accommodation)						
Configural	34993.541	2068				
Metric	39720.647	2114	Configural	4727.106 (46)	< .01	.001
Scalar	45818.461	2160	Metric	6097.814 (46)	< .01	.002

Appendix B.14b. Global Model Fit Indices of Scalar Invariance Model for Grade 7 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	44483.010	2160	< .01	0.980	0.022
Model B-1	20875.847	2160	< .01	0.981	0.021
Model B-2	36819.342	2160	< .01	0.977	0.022
Model B-3	18399.719	2160	< .01	0.982	0.020
Model B-4	21012.517	2160	< .01	0.982	0.020
Model B-5	17822.776	2160	< .01	0.983	0.019
Model C	44955.521	2160	< .01	0.980	0.020
Model D	40105.854	2160	< .01	0.981	0.021
Model E	41609.312	2160	< .01	0.985	0.019
Model F	45818.461	2160	< .01	0.981	0.020

Appendix B.15a. Global Model Fit Indices of Measurement Invariance Tests for Grade 8 Math

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	51923.973	2068				
Metric	53374.054	2114	Configural	1450.081 (46)	< .01	.000
Scalar	57215.968	2160	Metric	3841.913 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	24553.604	2068				
Metric	25354.154	2114	Configural	800.550 (46)	< .01	.000
Scalar	26162.288	2160	Metric	808.134 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	43941.130	2068				
Metric	45642.765	2114	Configural	1701.635 (46)	< .01	.000
Scalar	46780.427	2160	Metric	1137.662 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	23020.720	2068				
Metric	23325.360	2114	Configural	304.640 (46)	< .01	.000
Scalar	23646.081	2160	Metric	320.721 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	24766.121	2068				
Metric	25703.123	2114	Configural	937.002 (46)	< .01	.000
Scalar	26643.128	2160	Metric	940.004 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	23056.447	2068				
Metric	23120.850	2114	Configural	64.402 (46)	.038	.000
Scalar	23224.440	2160	Metric	103.590 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	48778.924	2068				
Metric	52531.621	2114	Configural	3752.697 (46)	< .01	.001
Scalar	57853.129	2160	Metric	5321.508 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	52081.658	2068				
Metric	53241.940	2114	Configural	1160.282 (46)	< .01	.000
Scalar	53724.454	2160	Metric	482.513 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	51790.117	2068				
Metric	52710.250	2114	Configural	920.133 (46)	< .01	.000
Scalar	54441.917	2160	Metric	1731.667 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	48522.307	2068				
Metric	51976.900	2114	Configural	3454.593 (46)	< .01	.001
Scalar	58394.447	2160	Metric	6417.548 (46)	< .01	.001

Appendix B.15b. Global Model Fit Indices of Scalar Invariance Model for Grade 8 Math

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	57215.968	2160	< .01	0.965	0.028
Model B-1	26162.288	2160	< .01	0.964	0.026
Model B-2	46780.427	2160	< .01	0.964	0.027
Model B-3	23646.081	2160	< .01	0.964	0.026
Model B-4	26643.128	2160	< .01	0.962	0.027
Model B-5	23224.440	2160	< .01	0.966	0.025
Model C	57853.129	2160	< .01	0.965	0.024
Model D	53724.454	2160	< .01	0.966	0.026
Model E	54441.917	2160	< .01	0.971	0.023
Model F	58394.447	2160	< .01	0.966	0.024

Appendix B.17a. Global Model Fit Indices of Measurement Invariance Tests for Algebra I

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	31880.312	2068				
Metric	32905.318	2114	Configural	1025.006 (46)	< .01	.000
Scalar	36406.798	2160	Metric	3501.480 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	16142.158	2068				
Metric	17121.893	2114	Configural	979.735 (46)	< .01	.001
Scalar	17604.115	2160	Metric	482.221 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	26017.305	2068				
Metric	29072.903	2114	Configural	3055.599 (46)	< .01	.001
Scalar	30186.879	2160	Metric	1113.975 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	15857.962	2068				
Metric	16107.596	2114	Configural	249.634 (46)	< .01	.000
Scalar	16536.207	2160	Metric	428.611 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	16758.206	2068				
Metric	18161.297	2114	Configural	1403.091 (46)	< .01	.001
Scalar	18954.656	2160	Metric	793.359 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	15465.159	2068				
Metric	15518.118	2114	Configural	52.959 (46)	.224	.001
Scalar	15611.302	2160	Metric	93.185 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	30806.225	2068				
Metric	32582.106	2114	Configural	1775.881 (46)	< .01	.000
Scalar	34778.963	2160	Metric	2196.858 (46)	< .01	.001
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	31432.467	2068				
Metric	33610.533	2114	Configural	2178.066 (46)	< .01	.001
Scalar	34009.880	2160	Metric	399.346 (46)	< .01	.001
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	31591.441	2068				
Metric	32369.885	2114	Configural	778.443 (46)	< .01	.000
Scalar	33674.958	2160	Metric	1305.073 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	31292.754	2068				
Metric	32158.284	2114	Configural	865.530 (46)	< .01	.000
Scalar	33945.363	2160	Metric	1787.079 (46)	< .01	.000

Appendix B.17b. Global Model Fit Indices of Scalar Invariance Model for Algebra I

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	36406.798	2160	< .01	0.929	0.024
Model B-1	17604.115	2160	< .01	0.981	0.021
Model B-2	30186.879	2160	< .01	0.974	0.023
Model B-3	16536.207	2160	< .01	0.981	0.021
Model B-4	18954.656	2160	< .01	0.979	0.022
Model B-5	15611.302	2160	< .01	0.983	0.020
Model C	34778.963	2160	< .01	0.927	0.024
Model D	34009.880	2160	< .01	0.930	0.023
Model E	33674.958	2160	< .01	0.936	0.023
Model F	33945.363	2160	< .01	0.982	0.020

Appendix B.16a. Global Model Fit Indices of Measurement Invariance Tests for Geometry

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	53224.910	2068				
Metric	54454.502	2114	Configural	1229.591 (46)	< .01	.000
Scalar	57940.176	2160	Metric	3485.674 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	27938.686	2068				
Metric	28828.529	2114	Configural	889.843 (46)	< .01	.000
Scalar	29338.108	2160	Metric	509.579 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	43389.350	2068				
Metric	47107.528	2114	Configural	3718.178 (46)	< .01	.001
Scalar	48016.657	2160	Metric	909.129 (46)	< .01	.000
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	27092.053	2068				
Metric	27384.954	2114	Configural	292.902 (46)	< .01	.000
Scalar	27629.871	2160	Metric	244.916 (46)	< .01	.001
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	27989.160	2068				
Metric	29872.199	2114	Configural	1883.040 (46)	< .01	.001
Scalar	30830.067	2160	Metric	957.867 (46)	< .01	.000
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	26391.366	2068				
Metric	26454.861	2114	Configural	63.495 (46)	.045	.001
Scalar	26525.935	2160	Metric	71.074 (46)	.010	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	52492.263	2068				
Metric	53783.151	2114	Configural	1290.888 (46)	< .01	.000
Scalar	55970.200	2160	Metric	2187.049 (46)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	51894.256	2068				
Metric	54836.635	2114	Configural	2942.380 (46)	< .01	.001
Scalar	55244.890	2160	Metric	408.255 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	53475.540	2068				
Metric	54059.554	2114	Configural	584.014 (46)	< .01	.000
Scalar	54692.029	2160	Metric	632.476 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	52971.934	2068				
Metric	53842.768	2114	Configural	870.834 (46)	< .01	.000
Scalar	55431.388	2160	Metric	1588.620 (46)	< .01	.000

Appendix B.16b. Global Model Fit Indices of Scalar Invariance Model for Geometry

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	57940.176	2160	< .01	0.930	0.038
Model B-1	29338.108	2160	< .01	0.950	0.033
Model B-2	48016.657	2160	< .01	0.929	0.035
Model B-3	27629.871	2160	< .01	0.947	0.034
Model B-4	30830.067	2160	< .01	0.951	0.032
Model B-5	26525.935	2160	< .01	0.954	0.031
Model C	55970.200	2160	< .01	0.942	0.031
Model D	55244.890	2160	< .01	0.934	0.036
Model E	54692.029	2160	< .01	0.944	0.032
Model F	55431.388	2160	< .01	0.955	0.027

Appendix B.18a. Global Model Fit Indices of Measurement Invariance Tests for Algebra II

Invariance Model	χ^2	*df*	χ^2 Difference Test			Change in RMSEA
			Comparison	$\chi^2(df)$	*p* value	
Model A: Students' Gender (Female vs. Male)						
Configural	17718.329	2068				
Metric	18874.760	2114	Configural	1156.431 (46)	< .01	.000
Scalar	20911.565	2160	Metric	2036.806 (46)	< .01	.001
Model B-1: Students' Ethnicity (African American vs. White)						
Configural	10607.551	2068				
Metric	11002.537	2114	Configural	394.986 (46)	< .01	.000
Scalar	11497.152	2160	Metric	494.615 (46)	< .01	.000
Model B-2: Students' Ethnicity (Hispanics vs. White)						
Configural	15096.010	2068				
Metric	16321.688	2114	Configural	1225.678 (46)	< .01	.001
Scalar	17382.791	2160	Metric	1061.103 (46)	< .01	.001
Model B-3: Students' Ethnicity (Asian vs. White)						
Configural	10235.537	2068				
Metric	10522.133	2114	Configural	286.595 (46)	< .01	.001
Scalar	10828.678	2160	Metric	306.545 (46)	< .01	.000
Model B-4: Students' Ethnicity (American Indian vs. White)						
Configural	10485.655	2068				
Metric	11186.459	2114	Configural	700.803 (46)	< .01	.001
Scalar	12270.158	2160	Metric	1083.699 (46)	< .01	.001
Model B-5: Students' Ethnicity (Multi-Ethnics vs. White)						
Configural	10048.537	2068				
Metric	10133.402	2114	Configural	84.865 (46)	< .01	.001
Scalar	10216.730	2160	Metric	83.328 (46)	< .01	.000
Model C: Students' SPED Status (Special Education vs. Non-SPED)						
Configural	17849.740	2068				
Metric	18266.012	2114	Configural	416.272 (46)	< .01	.000
Scalar	19649.191	2160	Metric	1383.179 (46)	< .01	.000
Model D: Students' Low Income Status (Low Income vs. Non-Low Income)						
Configural	17496.271	2068				
Metric	18725.869	2114	Configural	1229.598 (46)	< .01	.001
Scalar	19370.546	2160	Metric	644.677 (46)	< .01	.000
Model E: Students' LEP Status (Limited English Proficiency vs. Non-LEP)						
Configural	17883.903	2068				
Metric	18249.789	2114	Configural	365.886 (46)	< .01	.000
Scalar	18641.474	2160	Metric	391.685 (46)	< .01	.000
Model F: Students' Accommodation Status (Accommodation vs. Non- Accommodation)						
Configural	18128.269	2068				
Metric	18473.548	2114	Configural	345.279 (46)	< .01	.000
Scalar	19474.733	2160	Metric	1001.185 (46)	< .01	.000

Appendix B.18b. Global Model Fit Indices of Scalar Invariance Model for Algebra II

Model	Chi-Square Test			CFI	RMSEA
	Value	*df*	P-Value		
Model A	20911.565	2160	< .01	0.979	0.017
Model B-1	11497.152	2160	< .01	0.986	0.014
Model B-2	17382.791	2160	< .01	0.978	0.017
Model B-3	10828.678	2160	< .01	0.985	0.015
Model B-4	12270.158	2160	< .01	0.986	0.015
Model B-5	10216.730	2160	< .01	0.988	0.013
Model C	19649.191	2160	< .01	0.986	0.013
Model D	19370.546	2160	< .01	0.982	0.016
Model E	18641.474	2160	< .01	0.986	0.013
Model F	19474.733	2160	< .01	0.989	0.011

Appendix C: Regression Model Parameter Estimates of Differential Growth across Subgroups – ELA

Parameter	2015_G3E to 2016_G4E			2015_G4E to 2016_G5E		
	Estimate	Standard Error	*p* value	Estimate	Standard Error	*p* value
Intercept (*β₀₀*)	2521.59	0.14	<.0001	2539.23	0.14	<.0001
Female *vs.* Male (*β₀₁*)	1.42	0.13	<.0001	3.63	0.14	<.0001
Special Education Status *vs*. Non-SPED (*β₀₂*)	-6.70	0.28	<.0001	-8.89	0.29	<.0001
Limited English Proficiency vs. Non-LEP (*β₀₃*)	-11.12	0.41	<.0001	-9.94	0.49	<.0001
Low income *vs.* Non-Low Income(*β₀₄*)	-2.26	0.15	<.0001	-1.69	0.16	<.0001
Asian *vs.* White (*β₀₅*)	3.37	0.49	<.0001	3.49	0.49	<.0001
Hispanic *vs.* White (*β₀₆*)	-3.41	0.16	<.0001	-3.55	0.16	<.0001
African American *vs.* White (*β₀₇*)	-3.44	0.34	<.0001	-5.99	0.34	<.0001
Hawaiian/Pacific Islander *vs.* White (*β₀₈*)	-1.40	1.06	0.1846	-0.69	1.23	0.5744
American Indian *vs.* White (*β₀₉*)	-6.71	0.37	<.0001	-7.67	0.37	<.0001
Multiple *vs.* White (*β₀₁₀*)	-0.44	0.42	0.2939	-1.59	0.45	0.0004
Slope (*β₁₀*)	0.85	0.00	<.0001	0.92	0.00	<.0001
Female *vs.* Male (*β₁₁*)	0.00	0.00	0.8830	0.03	0.00	<.0001
Special Education Status *vs.* Non-SPED (*β₁₂*)	0.04	0.01	<.0001	-0.03	0.01	<.0001
Limited English Proficiency *vs.* Non-LEP (*β₁₃*)	-0.19	0.01	<.0001	-0.20	0.01	<.0001
Low income vs. Non-Low Income (*β₁₄*)	-0.04	0.01	<.0001	-0.03	0.01	<.0001
Asian *vs.* White (*β₁₅*)	-0.01	0.01	0.3709	-0.04	0.01	0.0036
Hispanic *vs.* White (*β₁₆*)	0.02	0.01	<.0001	-0.02	0.01	0.0003
African American *vs.* White (*β₁₇*)	0.06	0.01	<.0001	-0.04	0.01	0.0013
Hawaiian/Pacific Islander *vs.* White (*β₁₈*)	0.02	0.04	0.5411	0.00	0.05	0.9591
American Indian *vs.* White (*β₁₉*)	-0.05	0.01	<.0001	-0.07	0.01	<.0001
Multiple *vs.* White (*β₁₁₀*)	0.00	0.01	0.8149	0.00	0.02	0.9071

Parameter	2015 G5E to 2016 G6E			2015 G6E to 2016 G7E		
	Estimate	Standard Error	*p* value	Estimate	Standard Error	*p* value
Intercept (*β₀₀*)	2544.30	0.14	<.0001	2555.45	0.12	<.0001
Female *vs.* Male (*β₀₁*)	0.82	0.13	<.0001	0.62	0.12	<.0001
Special Education Status *vs*. Non-SPED (*β₀₂*)	-7.46	0.32	<.0001	-8.01	0.31	<.0001
Limited English Proficiency vs. Non-LEP (*β₀₃*)	-13.23	0.56	<.0001	-9.31	0.62	<.0001
Low income *vs.* Non-Low Income(*β₀₄*)	-2.00	0.15	<.0001	-1.90	0.14	<.0001
Asian *vs.* White (*β₀₅*)	4.30	0.46	<.0001	2.51	0.42	<.0001
Hispanic *vs.* White (*β₀₆*)	-4.83	0.16	<.0001	-2.70	0.14	<.0001
African American *vs.* White (*β₀₇*)	-5.42	0.33	<.0001	-2.81	0.29	<.0001
Hawaiian/Pacific Islander *vs.* White (*β₀₈*)	1.19	1.23	0.3313	-3.26	1.14	0.0042
American Indian *vs.* White (*β₀₉*)	-7.48	0.38	<.0001	-5.70	0.37	<.0001
Multiple *vs.* White (*β₀₁₀*)	-1.96	0.45	<.0001	-1.41	0.43	0.0010
Slope (*β₁₀*)	0.98	0.00	<.0001	0.80	0.00	<.0001
Female *vs.* Male (*β₁₁*)	0.02	0.00	<.0001	-0.01	0.00	0.0018
Special Education Status *vs.* Non-SPED (*β₁₂*)	-0.12	0.01	<.0001	0.00	0.01	0.8637
Limited English Proficiency *vs.* Non-LEP (*β₁₃*)	-0.31	0.01	<.0001	-0.11	0.02	<.0001
Low income vs. Non-Low Income (*β₁₄*)	-0.06	0.01	<.0001	0.00	0.01	0.7904
Asian *vs.* White (*β₁₅*)	0.01	0.01	0.3688	0.01	0.01	0.3462
Hispanic *vs.* White (*β₁₆*)	-0.04	0.01	<.0001	0.01	0.00	0.0962

African American *vs.*White ($β_{17}$)	-0.04	0.01	0.0013	0.01	0.01	0.2224
Hawaiian/Pacific Islander *vs.*White ($β_{18}$)	0.06	0.05	0.1736	0.01	0.04	0.8099
American Indian *vs.*White ($β_{19}$)	-0.11	0.01	<.0001	0.01	0.01	0.2810
Multiple *vs.*White ($β_{110}$)	0.02	0.02	0.1585	0.01	0.01	0.3085

Parameter	2015 G7E to 2016 G8E			2015 G8E to 2016 G9E		
	Estimate	**Standard Error**	***p* value**	**Estimate**	**Standard Error**	***p* value**
Intercept ($β_{00}$)	2557.47	0.13	<.0001	2568.20	0.13	<.0001
Female *vs.*Male ($β_{01}$)	1.53	0.13	<.0001	2.65	0.13	<.0001
Special Education Status *vs.* Non-SPED ($β_{02}$)	-8.45	0.33	<.0001	-7.16	0.39	<.0001
Limited English Proficiency vs. Non-LEP ($β_{03}$)	-4.25	0.54	<.0001	-2.46	0.49	<.0001
Low income *vs.* Non-Low Income($β_{04}$)	-0.66	0.15	<.0001	-4.07	0.17	<.0001
Asian *vs.*White ($β_{05}$)	4.53	0.45	<.0001	4.34	0.46	<.0001
Hispanic *vs.*White ($β_{06}$)	-2.38	0.15	<.0001	-4.10	0.15	<.0001
African American *vs.*White ($β_{07}$)	-3.00	0.30	<.0001	-4.07	0.33	<.0001
Hawaiian/Pacific Islander *vs.*White ($β_{08}$)	-1.73	1.23	0.1590	-1.43	1.11	0.1982
American Indian *vs.*White ($β_{09}$)	-5.18	0.38	<.0001	-4.06	0.41	<.0001
Multiple *vs.*White ($β_{010}$)	-0.83	0.45	0.0679	-0.34	0.54	0.5256
Slope ($β_{10}$)	0.91	0.00	<.0001	0.84	0.00	<.0001
Female *vs.*Male ($β_{11}$)	-0.01	0.00	0.0195	-0.01	0.00	0.0351
Special Education Status *vs.* Non-SPED ($β_{12}$)	-0.09	0.01	<.0001	-0.05	0.01	<.0001
Limited English Proficiency *vs.* Non-LEP ($β_{13}$)	-0.09	0.01	<.0001	0.01	0.01	0.5765
Low income vs. Non-Low Income ($β_{14}$)	-0.01	0.01	0.0779	-0.05	0.01	<.0001
Asian *vs.*White ($β_{15}$)	-0.01	0.01	0.6154	0.04	0.01	0.0027
Hispanic *vs.*White ($β_{16}$)	0.00	0.01	0.6914	-0.02	0.01	0.0012
African American *vs.*White ($β_{17}$)	0.01	0.01	0.5551	0.01	0.01	0.4781
Hawaiian/Pacific Islander *vs.*White ($β_{18}$)	-0.01	0.04	0.7772	0.04	0.04	0.3859
American Indian *vs.*White ($β_{19}$)	-0.06	0.01	<.0001	-0.05	0.01	0.0010
Multiple *vs.*White ($β_{110}$)	0.00	0.02	0.7872	0.00	0.02	0.9819

Parameter	2015 G9E to 2016 G10E			2015 G10E to 2016 G11E		
	Estimate	**Standard Error**	***p* value**	**Estimate**	**Standard Error**	***p* value**
Intercept ($β_{00}$)	2569.76	0.13	<.0001	2571.39	0.15	<.0001
Female *vs.*Male ($β_{01}$)	-0.43	0.14	0.0021	1.57	0.16	<.0001
Special Education Status *vs.* Non-SPED ($β_{02}$)	-5.72	0.40	<.0001	-8.01	0.48	<.0001
Limited English Proficiency vs. Non-LEP ($β_{03}$)	-4.23	0.48	<.0001	-2.70	0.62	<.0001
Low income *vs.* Non-Low Income($β_{04}$)	-3.51	0.18	<.0001	-4.68	0.22	<.0001
Asian *vs.*White ($β_{05}$)	2.66	0.47	<.0001	2.58	0.52	<.0001
Hispanic *vs.*White ($β_{06}$)	-3.66	0.16	<.0001	-3.86	0.18	<.0001
African American *vs.*White ($β_{07}$)	-4.45	0.33	<.0001	-4.48	0.40	<.0001
Hawaiian/Pacific Islander *vs.*White ($β_{08}$)	-3.40	1.38	0.0139	-2.88	1.45	0.0476
American Indian *vs.*White ($β_{09}$)	-5.15	0.41	<.0001	-4.25	0.47	<.0001
Multiple *vs.*White ($β_{010}$)	-0.30	0.56	0.5959	-0.43	0.64	0.5073
Intercept ($β_{00}$)	0.86	0.01	<.0001	0.91	0.01	<.0001
Female *vs.*Male ($β_{01}$)	-0.01	0.01	0.0638	0.00	0.01	0.9377
Special Education Status *vs.* Non-SPED ($β_{02}$)	-0.06	0.01	<.0001	-0.09	0.01	<.0001

Limited English Proficiency vs. Non-LEP (β_{03})	-0.02	0.02	0.2390	-0.05	0.02	0.0261
Low income *vs.* Non-Low Income(β_{04})	-0.07	0.01	<.0001	-0.03	0.01	0.0021
Asian *vs.* White (β_{05})	0.06	0.01	<.0001	0.08	0.02	<.0001
Hispanic *vs.* White (β_{06})	-0.03	0.01	<.0001	-0.03	0.01	0.0001
African American *vs.* White (β_{07})	-0.03	0.01	0.0323	-0.01	0.02	0.3888
Hawaiian/Pacific Islander *vs.* White (β_{08})	0.07	0.05	0.1502	0.01	0.06	0.8333
American Indian *vs.* White (β_{09})	-0.10	0.02	<.0001	-0.11	0.02	<.0001
Multiple *vs.* White (β_{010})	0.00	0.02	0.8904	0.00	0.02	0.9491

Note: G3E through G11E refers grade 3 ELA to grade 11 ELA. The significance of the effect is less than *p* value of .05.

Appendix C cont. Regression Model Parameter Estimates of Differential Growth across Subgroups – Math

	2015 G3M to 2016 G4M			2015 G4M to 2016 G5M		
Parameter	**Estimate**	**Standard Error**	***p* value**	**Estimate**	**Standard Error**	***p* value**
Intercept (β_{00})	3558.95	0.17	<.0001	3590.54	0.18	<.0001
Female *vs.* Male (β_{01})	-0.01	0.16	0.9286	3.55	0.17	<.0001
Special Education Status *vs.* Non-SPED (β_{02})	-8.44	0.33	<.0001	-10.00	0.35	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-9.16	0.43	<.0001	-6.60	0.56	<.0001
Low income *vs.* Non-Low Income(β_{04})	-0.36	0.18	0.0498	-1.40	0.19	<.0001
Asian *vs.* White (β_{05})	3.79	0.65	<.0001	7.97	0.64	<.0001
Hispanic *vs.* White (β_{06})	-6.32	0.19	<.0001	-2.11	0.20	<.0001
African American *vs.* White (β_{07})	-9.13	0.42	<.0001	-4.04	0.43	<.0001
Hawaiian/Pacific Islander *vs.* White (β_{08})	-3.93	1.30	0.0025	-1.11	1.51	0.4595
American Indian *vs.* White (β_{09})	-9.11	0.44	<.0001	-6.60	0.46	<.0001
Multiple *vs.* White (β_{010})	-3.15	0.50	<.0001	-0.67	0.55	0.2224
Slope (β_{10})	0.76	0.00	<.0001	0.82	0.00	<.0001
Female *vs.* Male (β_{11})	-0.01	0.00	0.0099	-0.02	0.00	<.0001
Special Education Status *vs.* Non-SPED (β_{12})	0.10	0.00	<.0001	0.08	0.01	<.0001
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.05	0.00	<.0001	-0.01	0.01	0.5093
Low income vs. Non-Low Income (β_{14})	-0.03	0.00	<.0001	-0.01	0.01	0.1949
Asian *vs.* White (β_{15})	-0.01	0.00	0.2898	0.01	0.01	0.4984
Hispanic *vs.* White (β_{16})	0.00	0.00	0.6600	0.02	0.01	<.0001
African American *vs.* White (β_{17})	0.03	0.00	0.0012	0.05	0.01	0.0001
Hawaiian/Pacific Islander *vs.* White (β_{18})	-0.01	0.00	0.7891	-0.04	0.04	0.4039
American Indian *vs.* White (β_{19})	0.00	0.00	0.9644	0.00	0.01	0.9941
Multiple *vs.* White (β_{110})	0.01	0.00	0.3142	0.01	0.01	0.3647

	2015 G5M to 2016 G6M			2015 G6M to 2016 G7M		
Parameter	**Estimate**	**Standard Error**	***p* value**	**Estimate**	**Standard Error**	***p* value**
Intercept (β_{00})	3620.97	0.17	<.0001	3637.50	0.14	<.0001
Female *vs.* Male (β_{01})	2.07	0.16	<.0001	0.28	0.14	0.0448
Special Education Status *vs.* Non-SPED (β_{02})	-11.02	0.37	<.0001	-10.27	0.34	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-12.58	0.59	<.0001	-11.15	0.59	<.0001
Low income *vs.* Non-Low Income(β_{04})	-1.22	0.19	<.0001	-2.74	0.16	<.0001
Asian *vs.* White (β_{05})	2.25	0.61	0.0002	2.51	0.55	<.0001
Hispanic *vs.* White (β_{06})	-5.52	0.19	<.0001	-3.82	0.16	<.0001
African American *vs.* White (β_{07})	-6.41	0.42	<.0001	-5.22	0.35	<.0001
Hawaiian/Pacific Islander *vs.* White (β_{08})	-0.86	1.51	0.5666	-2.60	1.30	0.0455

American Indian *vs.* White (β_{09})	-8.41	0.45	<.0001	-8.51	0.38	<.0001
Multiple *vs.* White (β_{010})	-2.11	0.55	0.0001	-0.89	0.49	0.0676
Slope (β_{10})	0.83	0.00	<.0001	0.81	0.00	<.0001
Female *vs.* Male (β_{11})	0.00	0.00	0.4006	-0.01	0.00	0.0122
Special Education Status *vs.* Non-SPED (β_{12})	0.11	0.01	<.0001	-0.01	0.01	0.5219
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.07	0.01	<.0001	-0.14	0.02	<.0001
Low income vs. Non-Low Income (β_{14})	0.01	0.01	0.2319	0.01	0.01	0.0479
Asian *vs.* White (β_{15})	0.02	0.01	0.1103	0.01	0.01	0.6488
Hispanic *vs.* White (β_{16})	0.01	0.01	0.0058	-0.01	0.00	0.0385
African American *vs.* White (β_{17})	0.07	0.01	<.0001	0.02	0.01	0.0349
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.09	0.04	0.0219	0.01	0.04	0.7645
American Indian *vs.* White (β_{19})	0.02	0.01	0.0702	-0.04	0.01	0.0002
Multiple *vs.* White (β_{110})	0.03	0.01	0.0515	0.01	0.01	0.5148

Parameter	2015 G7M to 2016 G8M			2015 G8M to 2016 AlgI		
	Estimate	Standard Error	*p* value	Estimate	Standard Error	*p* value
Intercept (β_{00})	3653.40	0.17	<.0001	3666.15	0.18	<.0001
Female *vs.* Male (β_{01})	2.48	0.16	<.0001	2.44	0.17	<.0001
Special Education Status *vs.* Non-SPED (β_{02})	-7.96	0.36	<.0001	-7.29	0.39	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-7.65	0.61	<.0001	-3.90	0.55	<.0001
Low income *vs.* Non-Low Income(β_{04})	-1.55	0.18	<.0001	-3.87	0.21	<.0001
Asian *vs.* White (β_{05})	5.31	0.67	<.0001	3.79	0.71	<.0001
Hispanic *vs.* White (β_{06})	-2.07	0.19	<.0001	-2.49	0.19	<.0001
African American *vs.* White (β_{07})	-2.06	0.38	<.0001	-2.36	0.40	<.0001
Hawaiian/Pacific Islander *vs.* White (β_{08})	0.75	1.53	0.6267	-1.74	1.47	0.2389
American Indian *vs.* White (β_{09})	-4.85	0.42	<.0001	-2.65	0.46	<.0001
Multiple *vs.* White (β_{010})	-1.00	0.58	0.0843	-1.54	0.73	0.0333
Slope (β_{10})	0.86	0.00	<.0001	0.76	0.01	<.0001
Female *vs.* Male (β_{11})	0.00	0.00	0.6631	-0.01	0.01	0.0114
Special Education Status *vs.* Non-SPED (β_{12})	-0.04	0.01	<.0001	-0.13	0.01	<.0001
Limited English Proficiency *vs.* Non-LEP (β_{13})	-0.12	0.02	<.0001	-0.07	0.02	<.0001
Low income vs. Non-Low Income (β_{14})	-0.03	0.01	<.0001	-0.03	0.01	<.0001
Asian *vs.* White (β_{15})	0.03	0.01	0.0546	0.04	0.02	0.0666
Hispanic *vs.* White (β_{16})	0.01	0.01	0.0981	-0.05	0.01	<.0001
African American *vs.* White (β_{17})	-0.01	0.01	0.6058	-0.06	0.01	<.0001
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.00	0.04	0.9952	-0.07	0.06	0.1891
American Indian *vs.* White (β_{19})	-0.04	0.01	0.0065	-0.08	0.02	<.0001
Multiple *vs.* White (β_{110})	0.03	0.02	0.1548	-0.03	0.03	0.2957

Parameter	2015 AlgI to 2016 Geo			2015 Geo to 2016 AlgII		
	Estimate	Standard Error	*p* value	Estimate	Standard Error	*p* value
Intercept (β_{00})	3688.61	0.18	<.0001	3699.15	0.19	<.0001
Female *vs.* Male (β_{01})	0.25	0.18	0.1716	1.37	0.20	<.0001
Special Education Status *vs.* Non-SPED (β_{02})	-8.23	0.50	<.0001	-7.22	0.61	<.0001
Limited English Proficiency vs. Non-LEP (β_{03})	-3.27	0.56	<.0001	1.28	0.70	0.0676
Low income *vs.* Non-Low Income(β_{04})	-3.49	0.25	<.0001	-4.97	0.28	<.0001

Asian *vs.* White (β_{05})	0.90	0.63	0.1512	4.86	0.64	<.0001
Hispanic *vs.* White (β_{06})	-5.06	0.21	<.0001	-4.00	0.23	<.0001
African American *vs.* White (β_{07})	-5.83	0.47	<.0001	-3.35	0.52	<.0001
Hawaiian/Pacific Islander *vs.* White (β_{08})	-5.64	1.76	0.0014	-4.30	1.90	0.0233
American Indian *vs.* White (β_{09})	-4.27	0.53	<.0001	-7.36	0.56	<.0001
Multiple *vs.* White (β_{010})	-2.65	0.74	0.0003	-0.30	0.79	0.7065
Slope (β_{10})	0.84	0.00	<.0001	0.79	0.01	<.0001
Female *vs.* Male (β_{11})	-0.01	0.01	0.0887	-0.01	0.01	0.3208
Special Education Status *vs.* Non-SPED (β_{12})	-0.08	0.01	<.0001	-0.10	0.02	<.0001
Limited English Proficiency *vs.* Non-LEP (β_{13})	0.02	0.02	0.2392	0.08	0.02	0.0004
Low income vs. Non-Low Income (β_{14})	-0.05	0.01	<.0001	-0.07	0.01	<.0001
Asian *vs.* White (β_{15})	0.03	0.01	0.0369	0.09	0.02	<.0001
Hispanic *vs.* White (β_{16})	-0.05	0.01	<.0001	-0.10	0.01	<.0001
African American *vs.* White (β_{17})	-0.04	0.02	0.0106	-0.08	0.02	<.0001
Hawaiian/Pacific Islander *vs.* White (β_{18})	0.04	0.06	0.4755	0.01	0.06	0.9300
American Indian *vs.* White (β_{19})	-0.07	0.02	<.0001	-0.15	0.02	<.0001
Multiple *vs.* White (β_{110})	0.00	0.02	0.9952	0.00	0.02	0.8723

Note: G3M through G8M refers grade 3 Math to grade 8 Math. AlgI and AlgII refers Algebra I and II, respectively. Geo refers Geometry. The significance of the effect is less than *p* value of .05.

Appendix D.1 – Student Participation by Demographic Subgroup – Fall 2015 Online Administration_online

Group	G9E	G10E	G11E	AlgI	Geo	AlgII
All students	2934	3304	5097	6084	4909	4756
Female	1354	1502	2484	2923	2258	2398
Male	1580	1802	2613	3161	2651	2358
African American	187	189	351	438	405	292
Asian	59	79	119	192	90	162
Native Hawaiian/Pacific	11	14	17	29	14	24
Hispanic/Latino	1285	1348	2340	2500	2356	1826
American Indian or Alaskan	268	228	289	234	194	233
White	1059	1331	1867	2563	1705	2078
Multiple	65	115	114	128	145	141
Limited English Proficiency	72	39	62	97	75	62
Special Education	273	275	383	270	379	206

Appendix D.2 – Student Participation by Demographic Subgroup – Fall 2015 Online Administration_paper

Group	G9E	G10E	G11E	AlgI	Geo	AlgII
All students	864	739	996	1275	978	656
Female	411	351	481	591	472	340
Male	453	388	515	684	506	316
African American	50	46	53	94	45	46
Asian	9	6	7	20	20	13
Native Hawaiian/Pacific	3	3	4	6	5	2
Hispanic/Latino	426	365	546	566	455	334
American Indian or Alaskan	38	27	41	51	28	12
White	317	267	331	506	393	237
Multiple	18	18	11	27	31	12
Limited English Proficiency	13	10	17	21	16	13
Special Education	77	91	140	179	138	53

Appendix E. Equations and Formula for Estimating Reliability

E.1 Standard Error Formula

For the AzMERIT assessments scored using MLE, according to Masters (1982), the asymptotic estimate of the standard error for ability θ is given by

$$SE(\theta) = \left[\sum_{i=1}^{N}\sum_{x_i=0}^{m_i} x_i^2\, P(X_i = x_i|\theta) - \sum_{i=1}^{N}\left[\sum_{x_i=0}^{m_i} x_i\, P(X_i = x_i|\theta)\right]^2\right]^{-\frac{1}{2}},$$

which is further placed onto the reporting scale by the following transformation:

$$SE_{vs} = a \times SE(\theta),$$

where a is the slope of the scaling constants that take θ to the reporting scale. For both ELA and Mathematics tests, $a = 30$.

E.2 Student Classification Consistency Formula

For a student with estimated ability $\hat{\theta}$ and associated standard error $\mathrm{se}(\hat{\theta})$, we can assume that $\hat{\theta}$ follows a normal distribution with mean of true ability θ and standard deviation of $\mathrm{se}(\hat{\theta})$, that is, $\hat{\theta} \sim N\left(\theta, \mathrm{se}(\hat{\theta})^2\right)$. The probability of the true score *at or above* the cut score θ_c is estimated as

$$P(\theta \geq \theta_c) = P\left(\frac{\theta - \hat{\theta}}{\mathrm{se}(\hat{\theta})} \geq \frac{\theta_c - \hat{\theta}}{\mathrm{se}(\hat{\theta})}\right) = P\left(\frac{\hat{\theta} - \theta}{\mathrm{se}(\hat{\theta})} < \frac{\hat{\theta} - \theta_c}{\mathrm{se}(\hat{\theta})}\right) = \Phi\left(\frac{\hat{\theta} - \theta_c}{\mathrm{se}(\hat{\theta})}\right),$$

where $\Phi(\cdot)$ is the cumulative function of standard normal distribution. Similarly, the probability of the true score being *below* the cut score is estimated as

$$P(\theta < \theta_c) = 1 - \Phi\left(\frac{\hat{\theta} - \theta_c}{\mathrm{se}(\hat{\theta})}\right).$$

E.2.1 Classification Accuracy Formula

The probability of a student with true ability θ being classified *at or above* the cut score θ_c, given the student's item scores $\boldsymbol{x} = (x_1, \cdots, x_N)$, can be estimated as

$$P(\theta \geq \theta_c|\boldsymbol{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta},$$

where the likelihood function is

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{N} P(x_i|\theta),$$

and $P(x_i|\theta)$ is calculated from the Rasch model or partial credit model based on the estimated item parameters.

Similarly, we can estimate the probability of *below* the cut score as:

$$P(\theta < \theta_c|\boldsymbol{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}$$

Mathematically, we have

$$N_{11} = \sum\nolimits_{i \in N_1} P(\theta_i \geq \theta_c|\boldsymbol{x}),$$

$$N_{01} = \sum\nolimits_{i \in N_1} P(\theta_i < \theta_c|\boldsymbol{x}),$$

$$N_{10} = \sum\nolimits_{i \in N_0} P(\theta_i \geq \theta_c|\boldsymbol{x})\text{, and}$$

$$N_{00} = \sum\nolimits_{i \in N_0} P(\theta_i < \theta_c|\boldsymbol{x}),$$

where N_1 consists of the students with estimated $\hat{\theta}_i$ being *at and above* the cut score, and N_0 contains the students with estimated $\hat{\theta}_i$ being *below* the cut score. The accuracy index is then computed as:

$$\frac{N_{11} + N_{00}}{N_1 + N_0}\ .$$

E.2.2 Classification Consistency Formula

To estimate the consistency, we assume the students are tested twice independently; hence, the probability of the student being classified as *at or above* the cut score θ_c in both tests can be estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c) = P(\theta_1 \geq \theta_c)P(\theta_2 \geq \theta_c) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

Similarly, the probability of consistency for *at or above* the cut score is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c|\boldsymbol{x}) = \left(\frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

The probability of consistency for *below* the cut score is estimated as

$$P(\theta_1 < \theta_c, \theta_2 < \theta_c|\boldsymbol{x}) = \left(\frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta}\right)^2.$$

The probability of inconsistency is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 < \theta_c|\boldsymbol{x}) = \frac{\int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta \int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta\right]^2}, \text{and}$$

$$P(\theta_1 < \theta_c, \theta_2 \geq \theta_c|\boldsymbol{x}) = \frac{\int_{-\infty}^{\theta_c} L(\theta|\boldsymbol{x})d\theta \int_{\theta_c}^{+\infty} L(\theta|\boldsymbol{x})d\theta}{\left[\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{x})d\theta\right]^2}.$$

The consistent index is computed as $\frac{N_{11} + N_{00}}{N}$, where

$$N_{11} = \sum_{i \in N} P(\theta_{i,1} \geq \theta_c, \theta_{i,2} \geq \theta_c|\boldsymbol{x}),$$

$$N_{01} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} \geq \theta_c|\boldsymbol{x}),$$

$$N_{10} = \sum_{i \in N} P(\theta_i \geq \theta_c, \theta_{i,2} < \theta_c|\boldsymbol{x}),$$

$$N_{00} = \sum_{i \in N} P(\theta_i < \theta_c, \theta_{i,2} < \theta_c|\boldsymbol{x}), \text{and}$$

$$N = N_{11} + N_{10} + N_{01} + N_{00}.$$

Appendix F: Standard Errors of Measurement – Fall 2015 Administration

Test	% Minimally Proficient	% Partially Proficient	% Proficient	% Highly Proficient	Overall
Grade 9 ELA	9.62	8.81	9.52	11.32	9.45
Grade 10 ELA	9.11	8.20	8.71	10.83	8.93
Grade 11 ELA	9.46	8.73	9.10	10.54	9.32
Algebra I	10.96	9.48	9.75	13.60	10.73
Geometry	12.80	10.53	10.16	11.31	11.77
Algebra II	13.79	10.96	10.43	11.23	12.24

Appendix G.1 – Number of Participating Students by Demographic Subgroups – ELA Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	62244	61069	60212	57639	58050	57350	51007	46818	41488
African American	3192	3194	3157	3120	3083	3077	2558	2373	1919
Asian	1458	1386	1384	1347	1329	1369	1231	1200	1109
Native Hawaiian/Pacific	235	244	208	170	155	144	205	149	143
Hispanic/Latino	29158	28182	27559	25946	25415	25172	21357	19155	16529
American Indian or Alaskan	3385	3412	3466	3166	3105	3048	3046	2818	2556
White	22945	22938	22897	22530	23711	23374	21841	20447	18653
Multiple	1871	1713	1541	1360	1252	1166	769	676	579
Female	30424	29986	29628	28073	28414	28138	24953	23085	20502
Male	31820	31083	30584	29566	29636	29212	26054	23733	20986
Limited English Proficiency	6505	6371	4921	3598	2998	2504	2394	2096	1465
Special Education	6797	7218	7236	6580	6085	5734	4395	3809	3127
Free/Reduced Lunch	26525	25856	25061	23567	22826	21773	18404	16146	13341
Accommodation	11724	11273	9492	7824	6533	5565	2495	1854	1422

Appendix G.2 – Number of Participating Students by Demographic Subgroups – ELA Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11
All students	25549	25256	25213	27012	26088	25429	29123	26585	23346
African American	1245	1238	1273	1398	1404	1478	1747	1622	1412
Asian	964	936	1082	1195	1090	1040	1192	1102	1011
Native Hawaiian/Pacific	78	105	69	102	90	84	83	58	74
Hispanic/Latino	11169	11023	10613	11623	11708	11450	14024	12334	10673
American Indian or Alaskan	969	921	960	1127	896	1015	854	756	597
White	10278	10212	10521	10904	10235	9771	10644	10136	9091
Multiple	832	809	687	658	649	579	528	522	452
Female	12710	12466	12474	13117	13004	12619	14535	13422	11706
Male	12825	12778	12731	13890	13068	12799	14540	13118	11610
Unknown Gender	14	12	8	5	16	11	48	45	30
Limited English Proficiency	1986	1940	1549	1318	1107	843	1093	520	290
Special Education	2784	2965	2968	2969	2818	2819	2551	2215	1869
Free/Reduced Lunch	11222	10999	10999	11899	11973	11659	11599	10341	8850
Accommodation	1715	1744	1439	1403	1079	905	616	337	227

Appendix G.3 – Number of Participating Students by Demographic Subgroups – Mathematics Online

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geometry	Algebra II
All students	62594	61395	60452	57874	57471	49787	53725	45784	39250
African American	3229	3241	3177	3141	3096	2898	2727	2161	1862
Asian	1465	1392	1386	1342	1246	947	1175	1203	1048
Native Hawaiian/Pacific	236	244	210	170	155	130	180	165	126
Hispanic/Latino	29321	28336	27651	26072	25421	23423	22997	18622	15576
American Indian or Alaskan	3416	3427	3483	3191	3178	2876	3567	2868	2346
White	23041	23034	22993	22589	23151	18504	22249	20079	17754
Multiple	1886	1721	1552	1369	1224	1009	830	686	538
Female	30576	30125	29723	28168	28201	24257	26023	22664	19805
Male	32018	31270	30729	29706	29270	25530	27702	23120	19445
Limited English Proficiency	6615	6450	4972	3639	3033	2310	2595	2135	1336
Special Education	6906	7287	7310	6621	6165	5729	4834	3286	2255
Free/Reduced Lunch	26077	25874	24986	23452	22820	20281	19568	16090	12676
Accommodation	11924	11518	9770	8083	6692	5652	2725	1665	1099

Appendix G.4 – Number of Participating Students by Demographic Subgroups – Mathematics Paper

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Geometry	Algebra II
All students	25709	25316	25267	26801	24358	20071	28898	25870	21650
African American	1259	1250	1277	1402	1341	1306	1779	1501	1334
Asian	970	949	1074	1011	638	468	1227	1153	903
Native Hawaiian/Pacific	80	104	69	101	88	70	81	57	66
Hispanic/Latino	11229	11019	10647	11666	11519	10167	14031	12020	10036
American Indian or Alaskan	988	926	968	1142	896	927	886	727	574
White	10330	10234	10535	10809	9244	6655	10320	9883	8255
Multiple	840	822	689	664	616	458	503	473	439
Female	12773	12482	12498	13005	12106	9846	14402	12988	10963
Male	12923	12822	12761	13790	12236	10206	14428	12829	10645
Unknown Gender	13	12	8	6	16	19	68	53	42
Limited English Proficiency	2007	1952	1560	1328	1110	843	1152	722	326
Special Education	2825	2978	2989	3007	2828	2727	2582	2033	1321
Free/Reduced Lunch	11080	10916	10946	11863	11867	10249	12292	10018	8632
Accommodation	1594	1616	1236	1226	919	750	343	221	148

Appendix H.1—Spring 16 Operational Item Parameter Estimates — Grade 3 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13021_e	ER	-1.13009	2.10828	4.15506	1.71108
2	13021_o	ER	-1.10203	1.71104	4.1758	1.59494
3	13021_c	ER	-1.31327	-0.98811		-1.15069
4	13025_e	ER	-1.28491	2.22063	3.90277	1.61283
5	13025_o	ER	-0.6838	1.77292	4.29041	1.79318
6	13025_c	ER	-1.58499	-0.47921		-1.0321
7	13026_e	ER	-1.11168	1.96713	4.20208	1.68584
8	13026_o	ER	-1.02194	1.83366	4.15705	1.65626
9	13026_c	ER	-1.53632	-0.61273		-1.07453
10	13022_e	ER	-1.50579	1.76305	4.40933	1.55553
11	13022_o	ER	-1.67709	1.56754	4.24577	1.37874
12	13022_c	ER	-1.75863	-0.83246		-1.29555
13	13024_e	ER	-1.10833	2.26727	4.05905	1.73933
14	13024_o	ER	-1.24303	2.04797	4.13136	1.64543
15	13024_c	ER	-1.6353	-0.94878		-1.29204
16	13023_e	ER	-1.15476	2.25355	4.12678	1.74186
17	13023_o	ER	-1.34689	2.14582	3.94719	1.58204
18	13023_c	ER	-1.69753	-0.82307		-1.2603
19	9698	MC4	-1.11745			-1.11745
20	9691	MC4	-0.37757			-0.37757
21	9687	HT	1.38239			1.38239
22	9700	MC4	1.23623			1.23623
23	9690	MC4	-1.209			-1.209
24	9697	HT	-0.26674			-0.26674
25	9694	EBSR4	1.70532			1.70532
26	9692	MC4	-0.60073			-0.60073
27	9699	MC4	-0.30132			-0.30132
28	8708	ETC	-1.85419			-1.85419
29	8710	ETC	-1.27986			-1.27986
30	8711	ETC	-0.1057			-0.1057
31	9359	MC4	0.28114			0.28114
32	9356	HT	1.30318			1.30318
33	9357	MC4	0.82249			0.82249
34	9353	MC4	0.10251			0.10251
35	9355	MC4	-0.3629			-0.3629
36	10268	MC4	0.2005			0.2005
37	9330	MC4	-0.2646			-0.2646
38	9336	MC4	-0.0956			-0.0956

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9331	MC4	0.62573			0.62573
40	9335	MC4	-0.64679			-0.64679
41	9333	MC4	0.30718			0.30718
42	835	MC4	-1.54772			-1.54772
43	837	MC4	-0.18142			-0.18142
44	9402	MS5	1.66101			1.66101
45	9398	MC4	-0.0567			-0.0567
46	9401	EBSR4	1.0949			1.0949
47	9347	MC4	0.27951			0.27951
48	9349	MC4	0.81415			0.81415
49	9404	MC4	-0.26242			-0.26242
50	9400	MC4	0.49505			0.49505
51	9414	MC4	-0.03304			-0.03304
52	9422	MC4	-0.87555			-0.87555
53	9418	MC4	0.70984			0.70984
54	10632	MC4	0.48179			0.48179
55	10634	MC4	1.17725			1.17725
56	9377	ETC	0.18504			0.18504
57	9378	ETC	-1.0865	-0.3551		-0.7208
58	9379	ETC	-1.03119			-1.03119
59	9380	ETC	-0.87865			-0.87865

Appendix H.2—Spring 16 Operational Item Parameter Estimates — Grade 4 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	9395	MC4	-0.2812			-0.2812
2	9392	MC4	-0.68057			-0.68057
3	9426	HT	0.92555			0.92555
4	10266	MC4	-0.74849			-0.74849
5	9391	MC4	0.26422			0.26422
6	1122	MC4	1.3626			1.3626
7	1124	MS4	-0.66063			-0.66063
8	9598	MC4	-0.0095			-0.0095
9	9595	MC4	-0.42531			-0.42531
10	9597	MC4	-0.09234			-0.09234
11	9596	MC4	0.65108			0.65108
12	9603	MS6	2.72225			2.72225
13	10644	ETC	-0.05707			-0.05707
14	10645	ETC	-2.09722	0.11686		-0.99018
15	10647	ETC	-0.85463			-0.85463
16	10263	MC4	-0.72314			-0.72314
17	9425	MC4	-0.64738			-0.64738
18	9382	MC4	-1.1843			-1.1843
19	9387	MC4	1.48043			1.48043
20	9386	MS6	0.46331			0.46331
21	9388	HT	-0.33297			-0.33297
22	9397	EBSR4	0.81106			0.81106
23	9389	MC4	0.04148			0.04148
24	9616	MC4	-0.19377			-0.19377
25	9390	MC4	-0.88945			-0.88945
26	9903	HT	0.55857			0.55857
27	10278	EBSR4	2.12205			2.12205
28	10277	MC4	1.11434			1.11434
29	9899	MC4	0.52233			0.52233
30	9902	MC4	-1.2141			-1.2141
31	9906	MC4	-0.53878			-0.53878
32	9900	HT	2.72673			2.72673
33	9446	MC4	-0.77651			-0.77651
34	9439	MS6	2.08491			2.08491
35	9437	MC4	-0.03208			-0.03208
36	9451	HT	1.17836			1.17836
37	9450	MC4	-0.16915			-0.16915
38	9438	MS6	1.75326			1.75326

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9428	ETC	0.79794			0.79794
40	9429	ETC	-2.21688	0.57637		-0.82026
41	9431	ETC	0.47241			0.47241
42	13094_e	ER	1.87077	4.87517	6.53587	4.42727
43	13094_o	ER	1.25092	4.69117	6.1041	4.0154
44	13094_c	ER	-0.79626	1.71384		0.45879
45	13095_e	ER	1.87872	4.8299	4.92767	3.87876
46	13095_o	ER	0.47386	4.35392	4.92003	3.24927
47	13095_c	ER	-1.45942	1.23827		-0.11058
48	13119_e	ER	1.54979	4.4619	6.46226	4.15798
49	13119_o	ER	0.64594	4.34742	5.35763	3.45033
50	13119_c	ER	-1.65829	1.05358		-0.30236
51	13121_e	ER	1.87054	4.48532	7.08457	4.48014
52	13121_o	ER	0.80218	4.25587	8.11115	4.38973
53	13121_c	ER	-1.70071	1.64402		-0.02835
54	13118_e	ER	1.69157	4.56961	4.98225	3.74781
55	13118_o	ER	0.73557	4.29715	4.84718	3.2933
56	13118_c	ER	-1.60239	1.2628		-0.1698
57	13120_e	ER	1.98955	5.05842	4.4995	3.84916
58	13120_o	ER	0.90148	4.27097	6.19001	3.78749
59	13120_c	ER	-1.25847	1.5145		0.12802

Appendix H.3—Spring 16 Operational Item Parameter Estimates — Grade 5 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13246_e	ER	0.33484	4.37175	6.34064	3.68241
2	13246_o	ER	-0.55282	3.33662	5.17957	2.65446
3	13246_c	ER	-2.03231	0.35683		-0.83774
4	13236_e	ER	-0.76876	3.70295	4.83054	2.58824
5	13236_o	ER	-0.83522	2.91473	5.00997	2.36316
6	13236_c	ER	-2.13985	-0.09297		-1.11641
7	13237_e	ER	0.76049	4.48983	5.72602	3.65878
8	13237_o	ER	-0.3275	3.62724	5.71883	3.00619
9	13237_c	ER	-1.61986	0.18987		-0.715
10	13238_e	ER	-0.90788	3.36594	4.06761	2.17522
11	13238_o	ER	-1.30216	2.84958	4.12669	1.89137
12	13238_c	ER	-2.13488	0.01614		-1.05937
13	13239_e	ER	0.02633	3.71815	4.71382	2.81943
14	13239_o	ER	-1.10817	3.18175	4.64235	2.23864
15	13239_c	ER	-2.15823	0.01608		-1.07108
16	13247_e	ER	-0.45605	3.2947	4.52191	2.45352
17	13247_o	ER	-1.00143	2.7942	4.59779	2.13019
18	13247_c	ER	-2.16137	0.43573		-0.86282
19	9305	MC4	-1.06814			-1.06814
20	9303	MC4	0.14403			0.14403
21	9304	MS6	0.5267			0.5267
22	9302	MC4	0.397			0.397
23	10264	HT	2.62702			2.62702
24	9290	ETC	-0.70663			-0.70663
25	9291	ETC	-1.14142	0.69588		-0.22277
26	9292	ETC	-1.08057	0.56956		-0.25551
27	9754	MS5	3.23528			3.23528
28	9755	MS6	1.2478			1.2478
29	9751	MC4	-0.23706			-0.23706
30	9757	MC4	-0.08355			-0.08355
31	9752	HT	1.19823			1.19823
32	9767	EBSR4	0.0675			0.0675
33	9753	MS6	0.81962			0.81962
34	9312	HT	2.72402			2.72402
35	9308	MC4	-0.32336			-0.32336
36	9307	MC4	-0.67813			-0.67813
37	9299	MC4	-0.1376			-0.1376
38	9298	MC4	-0.82066			-0.82066

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9311	HT	2.10824			2.10824
40	9306	MC4	-0.33688			-0.33688
41	9320	EBSR4	1.17902			1.17902
42	9783	MC4	-0.86303			-0.86303
43	9784	MC4	0.73684			0.73684
44	9842	HT	0.63012			0.63012
45	9833	MC4	0.8069			0.8069
46	9808	MC4	0.19637			0.19637
47	9809	MC4	0.24291			0.24291
48	9758	HT	-0.48322			-0.48322
49	9762	MC4	0.05262			0.05262
50	9765	MC4	1.41073			1.41073
51	9763	MC4	0.47559			0.47559
52	9600	HT	0.29219			0.29219
53	9599	HT	0.11816			0.11816
54	9294	MC4	-1.04402			-1.04402
55	9309	HT	3.39654			3.39654
56	9601	MS5	-0.19919			-0.19919
57	10659	ETC	-0.82784			-0.82784
58	10661	ETC	-2.20002	-0.66798		-1.434
59	10662	ETC	-1.01469			-1.01469

Appendix H.4—Spring 16 Operational Item Parameter Estimates — Grade 6 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13307_e	ER	0.4746	3.03111	4.52238	2.67603
2	13307_o	ER	-0.54732	2.76978	4.2732	2.16522
3	13307_c	ER	-2.10835	-0.10872		-1.10854
4	13308_e	ER	1.2363	3.33177	5.17142	3.2465
5	13308_o	ER	0.1599	3.083	4.37256	2.53849
6	13308_c	ER	-2.10091	-0.21027		-1.15559
7	13309_e	ER	0.33264	2.78999	4.58425	2.56896
8	13309_o	ER	-0.4541	2.19863	4.51841	2.08765
9	13309_c	ER	-2.8389	0.01884		-1.41003
10	13304_e	ER	-0.40648	2.83619	5.19328	2.541
11	13304_o	ER	-0.7602	2.34515	4.23866	1.9412
12	13304_c	ER	-2.38004	-0.22638		-1.30321
13	13305_e	ER	0.79496	3.84657	4.56174	3.06776
14	13305_o	ER	0.0947	3.01629	4.70018	2.60372
15	13305_c	ER	-1.96936	-0.65918		-1.31427
16	13306_e	ER	0.07056	3.38151	6.19445	3.21551
17	13306_o	ER	-0.59192	2.51057	5.24377	2.38747
18	13306_c	ER	-2.06099	-0.58055		-1.32077
19	9083	ETC	-2.37343	0.69063		-0.8414
20	9084	ETC	-0.529	1.94599		0.7085
21	9145	MC4	-0.95899			-0.95899
22	9143	MC4	0.38775			0.38775
23	9139	MC4	-0.90857			-0.90857
24	9140	MC4	0.29413			0.29413
25	10265	MC4	-1.04948			-1.04948
26	9142	HT	1.591			1.591
27	9807	EBSR4	1.168			1.168
28	9797	HT	0.97683			0.97683
29	9799	MC4	-0.16807			-0.16807
30	9802	MC4	0.42132			0.42132
31	9804	MC4	-0.49607			-0.49607
32	9803	MC4	-1.76786			-1.76786
33	9798	MC4	0.11274			0.11274
34	9800	EBSR4	1.21402			1.21402
35	9266	MC4	-1.04177			-1.04177
36	9272	MC4	-0.01006			-0.01006
37	9264	MC4	1.01395			1.01395
38	9262	MC4	1.18676			1.18676

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9267	HT	-0.59143			-0.59143
40	9263	MC4	-0.51352			-0.51352
41	9268	HT	2.08178			2.08178
42	10280	MC4	0.14517			0.14517
43	9872	HT	1.87785			1.87785
44	9870	HT	-0.18734			-0.18734
45	9869	MC4	-1.28088			-1.28088
46	9889	MC4	0.04903			0.04903
47	9867	MC4	0.88477			0.88477
48	9865	MC4	0.65339			0.65339
49	9866	MC4	0.07074			0.07074
50	9138	MC4	-1.48057			-1.48057
51	9168	EBSR4	0.35054			0.35054
52	9153	MC4	-0.41536			-0.41536
53	9130	MC4	-0.45134			-0.45134
54	9135	MS6	0.52718			0.52718
55	9134	MC4	-0.24354			-0.24354
56	9169	EBSR4	1.4272			1.4272
57	9131	MC4	0.87684			0.87684
58	9108	ETC	-1.32061	1.24437		-0.03812
59	9109	ETC	-1.49476	0.62001		-0.43738

Appendix H.5—Spring 16 Operational Item Parameter Estimates — Grade 7 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13400_e	ER	-0.89746	3.17022	4.70894	2.32723
2	13400_o	ER	-1.10875	2.81094	4.67186	2.12468
3	13400_c	ER	-1.93478	0.24135		-0.84672
4	13401_e	ER	-1.08712	3.58298	4.84348	2.44645
5	13401_o	ER	-2.01399	2.99796	5.04182	2.0086
6	13401_c	ER	-2.93019	-0.87734		-1.90377
7	13402_e	ER	-1.37884	3.89633	4.60481	2.3741
8	13402_o	ER	-1.49334	3.23935	5.1219	2.2893
9	13402_c	ER	-2.56472	-0.19658		-1.38065
10	13403_e	ER	0.48951	3.69526	5.16521	3.11666
11	13403_o	ER	-0.54572	3.33949	5.13751	2.64376
12	13403_c	ER	-2.82848	-0.61627		-1.72238
13	13405_e	ER	0.02255	3.38944	4.52755	2.64651
14	13405_o	ER	-0.88948	3.04194	4.92252	2.35833
15	13405_c	ER	-2.87726	-0.07362		-1.47544
16	13406_e	ER	-0.83081	4.35574	5.68822	3.07105
17	13406_o	ER	-1.24126	3.45562	5.9479	2.72075
18	13406_c	ER	-2.53963	0.10313		-1.21825
19	9103	ETC	0.33393			0.33393
20	9104	ETC	-1.21947	0.61781		-0.30083
21	9105	ETC	-0.14173			-0.14173
22	9147	MC4	-0.54315			-0.54315
23	9146	MC4	-1.19732			-1.19732
24	9152	MC4	-0.24903			-0.24903
25	9128	MS6	1.28632			1.28632
26	9610	MC4	0.90991			0.90991
27	9711	MC4	0.29644			0.29644
28	9611	HT	0.44176			0.44176
29	9713	MC4	0.2301			0.2301
30	9614	MC4	-1.02121			-1.02121
31	10613	MS5	1.17647			1.17647
32	10695	MC4	-0.51732			-0.51732
33	9750	MS6	-0.17847			-0.17847
34	9709	MC4	-0.24287			-0.24287
35	9177	MC4	0.5476			0.5476
36	9220	MC4	-0.2262			-0.2262
37	9176	MS5	1.43059			1.43059
38	10619	MC4	-0.10558			-0.10558

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9982	EBSR5	0.37426			0.37426
40	9962	MC4	0.59182			0.59182
41	10623	MC4	0.30353			0.30353
42	10621	EBSR4	1.03655			1.03655
43	9189	MC4	1.13439			1.13439
44	9778	MC4	0.42645			0.42645
45	9777	MC4	0.77082			0.77082
46	9779	HT	0.32439			0.32439
47	9781	MC4	0.22608			0.22608
48	9817	MS5	-0.10239			-0.10239
49	9811	MC4	-0.37913			-0.37913
50	9785	MS6	1.35871			1.35871
51	9786	MS5	1.03456			1.03456
52	9787	EBSR4	0.58622			0.58622
53	10274	MC4	1.70658			1.70658
54	9793	HT	1.62756			1.62756
55	9791	MC4	0.26709			0.26709
56	9789	MC4	-0.24857			-0.24857
57	10656	ETC	0.00688			0.00688
58	10657	ETC	-1.30701	-0.11629		-0.71165
59	10658	ETC	-0.75798	1.82429		0.53316

Appendix H.6—Spring 16 Operational Item Parameter Estimates — Grade 8 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13437_e	ER	-1.09959	1.99003	3.43357	1.44134
2	13437_o	ER	-1.21497	1.49481	3.52606	1.26863
3	13437_c	ER	-2.194	-0.55737		-1.37569
4	13438_e	ER	-1.30982	1.60181	3.36366	1.21855
5	13438_o	ER	-1.25185	1.38879	3.19409	1.11034
6	13438_c	ER	-2.48946	-0.96722		-1.72834
7	13439_e	ER	-1.77948	1.72887	2.90301	0.9508
8	13439_o	ER	-1.71597	1.09631	3.03584	0.80539
9	13439_c	ER	-2.40518	-1.1545		-1.77984
10	13452_e	ER	-1.67899	1.57768	3.33549	1.07806
11	13452_o	ER	-1.89107	1.08197	3.42027	0.87039
12	13452_c	ER	-2.54121	-1.11311		-1.82716
13	13453_e	ER	-1.06493	1.5231	3.24717	1.23511
14	13453_o	ER	-1.23977	0.91325	3.22926	0.96758
15	13453_c	ER	-1.99191	-0.7686		-1.38026
16	13454_e	ER	-1.16469	1.93019	3.31328	1.35959
17	13454_o	ER	-1.30159	1.37223	3.31286	1.12783
18	13454_C	ER	-2.12549	-0.79592		-1.46071
19	9116	MC4	-0.28257			-0.28257
20	9170	EBSR6	0.50345			0.50345
21	9115	HT	0.74599			0.74599
22	9240	MC4	-0.35422			-0.35422
23	9172	EBSR4	0.50692			0.50692
24	9113	MC4	-0.18363			-0.18363
25	9171	EBSR4	1.84183			1.84183
26	9255	HT	0.69969			0.69969
27	9259	MC4	0.64706			0.64706
28	9256	MC4	-0.24916			-0.24916
29	9253	MC4	-0.62194			-0.62194
30	9254	MC4	-0.32173			-0.32173
31	9017	MC4	-0.09267			-0.09267
32	9019	MC4	0.04892			0.04892
33	9014	MC4	-0.9944			-0.9944
34	9230	MC4	-1.15492			-1.15492
35	9018	MC4	0.80965			0.80965
36	9046	MC4	-0.82499			-0.82499
37	9015	MS5	1.20929			1.20929
38	9076	ETC	-0.64308			-0.64308

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9077	ETC	-1.7963	0.32962		-0.73334
40	9078	ETC	-0.93582	0.80924		-0.06329
41	10639	HT	2.91528			2.91528
42	9881	MC4	0.56406			0.56406
43	9880	MC4	-0.04633			-0.04633
44	9883	MC4	-1.4832			-1.4832
45	9885	MC4	-0.67563			-0.67563
46	10626	HT	1.26297			1.26297
47	9026	MC4	0.5053			0.5053
48	9022	MC4	0.99092			0.99092
49	9024	MC4	-0.53744			-0.53744
50	9023	MC4	0.95894			0.95894
51	9028	MC4	0.17704			0.17704
52	9020	MS5	1.09383			1.09383
53	9728	ETC	-1.65694	-0.20084		-0.92889
54	9729	ETC	-1.73003	0.26659		-0.73172
55	8965	MC4	-1.58511			-1.58511
56	8973	HT	-0.21957			-0.21957
57	9044	MC4	0.89154			0.89154
58	8967	MC4	-0.76092			-0.76092
59	8971	HT	1.15521			1.15521

Appendix H.7—Spring 16 Operational Item Parameter Estimates — Grade 9 ELA

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
1		13554_e	ER	-1.58337	2.72154	4.17335	1.77051
2		13554_o	ER	-1.49156	1.80507	4.60922	1.64091
3		13554_c	ER	-2.17867	-1.14465		-1.66166
4		13555_e	ER	-1.46735	2.51036	4.0225	1.6885
5		13555_o	ER	-1.7157	1.64362	3.78329	1.23707
6		13555_c	ER	-2.4999	-1.03399		-1.76695
7		13557_e	ER	-1.48371	2.29755	4.26558	1.69314
8		13557_o	ER	-1.33444	1.51754	4.19585	1.45965
9		13557_c	ER	-2.4354	-0.8662		-1.6508
10		13565_e	ER	-1.41497	1.96395	3.80007	1.44968
11		13565_o	ER	-1.69345	1.40948	3.82485	1.18029
12		13565_c	ER	-2.20831	-0.65189		-1.4301
13		13566_e	ER	-1.80017	2.23734	4.14528	1.52748
14		13566_o	ER	-1.66128	1.36788	4.09913	1.26858
15		13566_c	ER	-2.39075	-0.81852		-1.60464
16		13556_e	ER	-1.38759	2.43865	3.83392	1.62833
17		13556_o	ER	-1.76241	1.77841	3.89649	1.30416
18		13556_c	ER	-2.38602	-1.37075		-1.87839
19		10596	MC4	-1.63193			-1.63193
20		10595	MC4	-1.00777			-1.00777
21		10597	MC4	-1.30579			-1.30579
22		10603	MS6	0.37171			0.37171
23		9064	HT	1.21086			1.21086
24		9060	HT	1.2773			1.2773
25		9066	MC4	0.37042			0.37042
26		9065	MC4	0.29345			0.29345
27		9058	MC4	0.35867			0.35867
28		9063	MC4	-0.41625			-0.41625
29		9069	HT	1.02115			1.02115
30		1444	MC4	-0.4155			-0.4155
31		1446	MS4	-0.02528			-0.02528
32		9050	HT	1.37944			1.37944
33		9047	MS5	1.04482			1.04482
34		9048	MC4	-0.38622			-0.38622
35		9051	EBSR4	0.20427			0.20427
36		9053	MC4	-0.83869			-0.83869
37		11097	MC4	-0.45154			-0.45154
38		11098	MC4	0.4152			0.4152

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
39		9032	MC4	0.81518			0.81518
40		9040	MC4	1.00112			1.00112
41		9038	MC4	0.55427			0.55427
42		9033	MC4	-0.15469			-0.15469
43		9034	HT	1.18483			1.18483
44		8987	MC4	-0.57348			-0.57348
45		8989	MC4	-1.00653			-1.00653
46		9003	MC4	0.30221			0.30221
47		9005	HT	0.43076			0.43076
48		9043	EBSR4	0.37237			0.37237
49		8990	MC4	0.50835			0.50835
50		9004	MC4	0.0071			0.0071
51		9042	GI	0.59316			0.59316
52		9006	MC4	0.07488			0.07488
53		9007	MC4	-0.14351			-0.14351
54		9041	EBSR4	1.83768			1.83768
55		9012	MC4	0.3965			0.3965
56		8948	ETC	0.18506			0.18506
57		8951	ETC	-0.89183	1.4474		0.27779
58		9734	ETC	-0.38567			-0.38567
59		9735	ETC	-1.01417	1.57881		0.28232
60		9736	ETC	-0.28792			-0.28792
61		9737	ETC	-1.41093			-1.41093

Appendix H.8—Spring 16 Operational Item Parameter Estimates — Grade 10 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13638_e	ER	-1.78924	1.61688	3.39067	1.07277
2	13638_o	ER	-1.97581	0.93168	3.19227	0.71605
3	13638_c	ER	-3.08122	-1.11156		-2.09639
4	13635_e	ER	-1.64858	1.09969	3.31762	0.92291
5	13635_o	ER	-2.00564	0.76273	3.12769	0.62826
6	13635_c	ER	-3.09387	-1.09958		-2.09673
7	13639_e	ER	-1.16138	1.59043	3.81864	1.4159
8	13639_o	ER	-1.47773	1.06526	3.81875	1.13543
9	13639_c	ER	-2.96755	-1.16419		-2.06587
10	13636_e	ER	-1.14411	1.51623	4.30412	1.55875
11	13636_o	ER	-1.25137	0.71571	3.49201	0.98545
12	13636_c	ER	-2.77526	-1.56572		-2.17049
13	13637_e	ER	-1.2983	2.05466	3.36557	1.37398
14	13637_o	ER	-1.89522	1.40099	3.39837	0.96805
15	13637_c	ER	-2.97674	-0.96661		-1.97168
16	13640_e	ER	-1.90091	1.21808	2.96704	0.7614
17	13640_o	ER	-2.36052	0.51171	3.10403	0.41841
18	13640_c	ER	-3.03408	-1.28609		-2.16009
19	8785	MC4	-1.24117			-1.24117
20	8925	MC4	-0.35929			-0.35929
21	8788	MC4	-0.79543			-0.79543
22	8795	HT	1.52988			1.52988
23	8844	EBSR4	3.08365			3.08365
24	9623	MC4	-0.33285			-0.33285
25	9624	HT	0.96931			0.96931
26	9627	MC4	0.01553			0.01553
27	9626	MC4	0.39998			0.39998
28	9630	MC4	0.53771			0.53771
29	8815	MC4	0.55677			0.55677
30	8972	MC4	0.46067			0.46067
31	8816	MC4	0.64997			0.64997
32	8926	MC4	0.42232			0.42232
33	8821	MC4	0.18705			0.18705
34	8818	MC4	0.40891			0.40891
35	8819	MC4	0.62246			0.62246
36	9822	MC4	-0.54876			-0.54876
37	9824	MS5	0.82801			0.82801
38	9827	MC4	0.33355			0.33355

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	9826	MC4	-0.43866			-0.43866
40	9821	MC4	0.03241			0.03241
41	9825	MC4	-0.52709			-0.52709
42	9887	MS5	1.98708			1.98708
43	9888	MC4	-0.02807			-0.02807
44	8760	ETC	0.86409			0.86409
45	8761	ETC	-0.1046			-0.1046
46	8762	ETC	0.57993			0.57993
47	8764	ETC	-0.66888	1.12054		0.22583
48	9831	MC4	-1.13658			-1.13658
49	9830	HT	1.81934			1.81934
50	9829	MC4	-0.02291			-0.02291
51	9837	MC4	-0.3871			-0.3871
52	9839	MC4	-0.068			-0.068
53	9840	MC4	0.4925			0.4925
54	9769	MC4	0.02052			0.02052
55	9773	MC4	0.8572			0.8572
56	9774	MC4	-0.07955			-0.07955
57	9772	MC4	-1.27366			-1.27366
58	9771	MC4	-0.44234			-0.44234
59	8757	ETC	-1.36906			-1.36906
60	8758	ETC	0.32886			0.32886
61	8759	ETC	-1.17573	-0.21852		-0.69713

Appendix H.9—Spring 16 Operational Item Parameter Estimates — Grade 11 ELA

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	13723_e	ER	-1.9275	0.81378	2.56933	0.4852
2	13723_o	ER	-3.05795	0.1766	2.29503	-0.19544
3	13723_c	ER	-2.93586	-1.14637		-2.04112
4	13725_e	ER	-1.76303	0.57642	2.89886	0.57075
5	13725_o	ER	-2.23547	0.75545	2.49751	0.33916
6	13725_c	ER	-2.62307	-0.55033		-1.5867
7	13720_e	ER	-1.06366	1.77264	3.26664	1.32521
8	13720_o	ER	-2.24515	1.13742	3.0911	0.66112
9	13720_c	ER	-3.06998	-0.79431		-1.93215
10	13724_e	ER	-1.89212	0.54532	3.32819	0.66046
11	13724_o	ER	-2.63027	0.506	3.14192	0.33922
12	13724_c	ER	-2.90567	-0.96473		-1.9352
13	13721_e	ER	-2.24616	0.86028	3.32082	0.64498
14	13721_o	ER	-2.99104	0.50579	3.23517	0.24997
15	13721_c	ER	-2.98695	-0.98931		-1.98813
16	13722_e	ER	-1.92023	0.91135	3.08764	0.69292
17	13722_o	ER	-2.77119	0.40624	2.72972	0.12159
18	13722_C	ER	-2.88156	-1.06637		-1.97397
19	9853	HT	-0.7513			-0.7513
20	9856	MC4	-0.13181			-0.13181
21	9860	MC4	-0.3386			-0.3386
22	9858	MC4	0.99024			0.99024
23	9852	MS6	-0.84321			-0.84321
24	8862	EBSR4	0.76635			0.76635
25	8861	MC4	0.25879			0.25879
26	8865	HT	0.57228			0.57228
27	8867	MC4	0.18188			0.18188
28	8871	MC4	0.03863			0.03863
29	8805	HT	0.28483			0.28483
30	8807	MC4	0.03819			0.03819
31	8808	MC4	-0.43028			-0.43028
32	8809	HT	0.11628			0.11628
33	8846	EBSR4	0.2237			0.2237
34	8753	ETC	-1.29646			-1.29646
35	8754	ETC	-1.27686			-1.27686
36	8755	ETC	-0.48937	1.65996		0.5853
37	8797	MC4	-0.35939			-0.35939
38	8798	MC4	0.17292			0.17292

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	8796	MC4	0.05932			0.05932
40	8800	MC4	-0.77328			-0.77328
41	8799	MC4	-0.52283			-0.52283
42	8880	MC4	-0.02503			-0.02503
43	8881	MC4	-0.37143			-0.37143
44	8879	MS6	0.52607			0.52607
45	8884	MC4	0.26397			0.26397
46	8794	HT	0.60324			0.60324
47	8783	MC4	-0.68116			-0.68116
48	8792	MC4	-0.21395			-0.21395
49	8784	MC4	-0.25728			-0.25728
50	8781	MC4	0.34351			0.34351
51	8791	MS5	1.31404			1.31404
52	421	MC4	-0.19236			-0.19236
53	422	MS4	1.07426			1.07426
54	8856	EBSR4	0.53849			0.53849
55	8834	MC4	-0.09598			-0.09598
56	8837	MS5	1.07611			1.07611
57	8841	MC4	0.06178			0.06178
58	8843	MS5	0.82014			0.82014
59	8769	ETC	0.3946			0.3946
60	8770	ETC	-1.5716	0.25326		-0.65917
61	8771	ETC	-1.2835			-1.2835

Appendix H.10—Spring 16 Operational Item Parameter Estimates — Grade 3 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10453	MC4	-3.3943			-3.3943
2	10425	MC4	-1.53829			-1.53829
3	10411	MC4	-1.61799			-1.61799
4	10403	MC4	-0.58186			-0.58186
5	8463	MC4	-0.59241			-0.59241
6	9455	EQ	-0.07043			-0.07043
7	10416	MC4	0.31901			0.31901
8	10427	EQ	1.0682			1.0682
9	10448	EQ	1.09544			1.09544
10	10430	EQ	2.32239			2.32239
11	10685	MC4	1.69992			1.69992
12	10470	EQ	2.23998			2.23998
13	10399	EQ	0.88755			0.88755
14	10446	EQ	1.29275			1.29275
15	10398	EQ	0.4498			0.4498
16	10396	EQ	0.38434			0.38434
17	10671	EQ	0.39429			0.39429
18	10400	EQ	-0.25945			-0.25945
19	10434	EQ	-0.75061			-0.75061
20	8461	MC4	-0.60467			-0.60467
21	10395	EQ	-2.04152			-2.04152
22	10438	EQ	-2.34273			-2.34273
23	10443	MC4	-2.32838			-2.32838
24	10409	MC4	-1.87291			-1.87291
25	10384	EQ	-1.35609			-1.35609
26	10439	EQ	-1.01284			-1.01284
27	10436	MC4	-0.50493			-0.50493
28	10391	MC4	-0.4991			-0.4991
29	10466	EQ	0.826			0.826
30	8481	MC4	0.82314			0.82314
31	10433	EQ	1.33708			1.33708
32	10455	EQ	1.85302			1.85302
33	9464	EQ	1.3174			1.3174
34	10454	MI	2.28432			2.28432
35	9469	EQ	1.57178			1.57178
36	11120	EQ	1.1283			1.1283
37	10683	MC4	1.42686			1.42686
38	10392	EQ	0.62557			0.62557

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10687	MC4	0.89578			0.89578
40	10418	EQ	-0.2178			-0.2178
41	10450	GI	0.34171			0.34171
42	8483	MC4	-0.02135			-0.02135
43	10431	EQ	-1.08021			-1.08021
44	9460	EQ	-1.67509			-1.67509
45	10402	EQ	-2.02582			-2.02582

Appendix H.11—Spring 16 Operational Item Parameter Estimates — Grade 4 Mathematics

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
1		9456	GI	-3.07549			-3.07549
2		10766	MS5	-1.46597			-1.46597
3		10771	MS5	-0.98077			-0.98077
4		10826	GI	-1.49232	0.80594		-0.34319
5		10710	MS5	-0.44114			-0.44114
6		9474	EQ	-0.17779			-0.17779
7		8501	MC4	0.50338			0.50338
8		10739	MC4	0.68294			0.68294
9		9465	EQ	0.63804			0.63804
10		10782	MS5	1.46482			1.46482
11		10754	MS5	1.42943			1.42943
12		10780	EQ	2.24981	2.50292		2.37636
13		9475	GI	2.19462			2.19462
14		10728	EQ	1.29344			1.29344
15		10724	MS5	1.13173			1.13173
16		10715	EQ	0.74127			0.74127
17		10752	MC4	0.44252			0.44252
18		9502	EQ	0.32547			0.32547
19		10718	EQ	0.03231			0.03231
20		10827	EQ	-0.86309			-0.86309
21		10783	MC4	-0.58052			-0.58052
22		10731	EQ	-0.7754			-0.7754
23		8505	MC4	-1.77368			-1.77368
24		10772	MC4	-2.60398			-2.60398
25		10729	EQ	-0.97271			-0.97271
26		8497	MC4	-1.38394			-1.38394
27		10730	EQ	-0.92949			-0.92949
28		10750	EQ	-0.58904			-0.58904
29		10705	MC4	0.15075			0.15075
30		10744	EQ	-0.50014			-0.50014
31		9452	EQ	0.34683			0.34683
32		9467	GI	0.92347			0.92347
33		10716	EQ	1.10308			1.10308
34		9470	EQ	1.96056			1.96056
35		10727	EQ	2.15749			2.15749
36		10753	MS5	0.86014			0.86014
37		10779	EQ	0.34347			0.34347
38		10781	MC4	0.68116			0.68116

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
39		9471	EQ	1.32625			1.32625
40		10719	EQ	-0.03793			-0.03793
41		10763	EQ	-0.20366			-0.20366
42		10751	MC4	-0.3404			-0.3404
43		10708	MC4	-0.36231			-0.36231
44		8493	MC4	-0.80101			-0.80101
45		8499	MC4	-2.11291			-2.11291

Appendix H.12—Spring 16 Operational Item Parameter Estimates — Grade 5 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10789	EQ	-2.83917			-2.83917
2	10811	MC4	-1.11389			-1.11389
3	10829	EQ	-1.59476			-1.59476
4	10836	MC4	-0.75042			-0.75042
5	10799	EQ	-0.59379			-0.59379
6	9486	EQ	0.14711			0.14711
7	10803	EQ	-0.2183			-0.2183
8	10839	EQ	0.63393			0.63393
9	10808	EQ	0.95699			0.95699
10	10813	EQ	1.41929			1.41929
11	9476	EQ	1.35515			1.35515
12	10835	EQ	2.12229			2.12229
13	10832	MC4	-0.6078			-0.6078
14	10849	EQ	1.15146			1.15146
15	10788	MS5	0.2804			0.2804
16	10863	MC4	0.62779			0.62779
17	10824	MC4	-0.17851			-0.17851
18	10798	MS5	-0.17942			-0.17942
19	8525	MC4	-0.58567			-0.58567
20	10872	GI	-0.76667	-0.47954		-0.6231
21	8535	MC4	-0.97262			-0.97262
22	8539	MC4	-1.17271			-1.17271
23	9716	MI	-2.30144			-2.30144
24	10848	EQ	-0.81249			-0.81249
25	8521	MC4	-1.17672			-1.17672
26	10790	EQ	0.78506			0.78506
27	11107	MC4	-0.46606			-0.46606
28	10805	EQ	-0.48618			-0.48618
29	10850	EQ	0.19694			0.19694
30	10816	EQ	0.02499			0.02499
31	10851	EQ	0.92342			0.92342
32	10833	EQ	-0.99926			-0.99926
33	9485	GI	1.76276			1.76276
34	10868	EQ	3.6372			3.6372
35	10817	TI	3.1767	1.86902		2.52286
36	10840	EQ	0.96138			0.96138
37	10820	MC4	1.62329			1.62329
38	9487	EQ	1.1116			1.1116

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10796	EQ	0.28237			0.28237
40	10869	MS6	0.21386			0.21386
41	10800	EQ	1.11572			1.11572
42	8527	MC4	-0.68748			-0.68748
43	10791	EQ	0.89984			0.89984
44	10875	MC4	-0.69325			-0.69325
45	10795	EQ	-1.39678			-1.39678

Appendix H.13—Spring 16 Operational Item Parameter Estimates — Grade 6 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10067	EQ	-3.72557			-3.72557
2	9491	GI	-1.70216			-1.70216
3	9492	EQ	-1.41594			-1.41594
4	10053	EQ	-0.7795			-0.7795
5	8567	MC4	-0.65504			-0.65504
6	10051	EQ	-0.35771			-0.35771
7	10120	EQ	0.47249			0.47249
8	10076	TI	0.82913			0.82913
9	10083	MS5	0.91604			0.91604
10	10096	EQ	0.93422			0.93422
11	9512	EQ	1.37691			1.37691
12	9718	MI	2.34031			2.34031
13	10151	GI	1.98277			1.98277
14	10117	MC4	0.89436			0.89436
15	10143	EQ	0.87374			0.87374
16	10095	TI	0.97298			0.97298
17	10071	MC4	0.56468			0.56468
18	10093	MC4	0.03634			0.03634
19	10064	EQ	-0.12482			-0.12482
20	10052	MC4	0.15283			0.15283
21	10115	EQ	-0.21997			-0.21997
22	10113	MC4	-1.43848			-1.43848
23	10137	EQ	-2.35826			-2.35826
24	8549	MC4	-2.53259			-2.53259
25	10106	EQ	-1.86826			-1.86826
26	9719	MI	-1.29034			-1.29034
27	10108	MC4	-0.6664			-0.6664
28	10057	EQ	-1.19929			-1.19929
29	9496	EQ	0.22756			0.22756
30	10049	EQ	-0.09381			-0.09381
31	10048	MC4	0.40685			0.40685
32	10078	EQ	1.05127			1.05127
33	10060	EQ	1.20803			1.20803
34	9513	GI	1.29516			1.29516
35	9498	EQ	2.73279			2.73279
36	10139	MS6	2.24036			2.24036
37	10111	MS5	1.60177			1.60177
38	10070	EQ	0.2958			0.2958

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10094	EQ	1.03964			1.03964
40	10079	EQ	0.82051			0.82051
41	10088	EQ	0.63477			0.63477
42	10150	EQ	-0.04156			-0.04156
43	8555	MC4	-0.17599			-0.17599
44	10148	MC4	-0.4154			-0.4154
45	10103	EQ	-0.93491			-0.93491
46	10107	EQ	-1.82823			-1.82823
47	10129	MC4	-2.30001			-2.30001

Appendix H.14—Spring 16 Operational Item Parameter Estimates — Grade 7 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	8698	MC4	-2.05133			-2.05133
2	10290	MC4	-0.68728			-0.68728
3	10301	EQ	-0.54092			-0.54092
4	8597	MC4	0.07583			0.07583
5	9520	GI	0.27072			0.27072
6	10294	EQ	0.11229			0.11229
7	10701	EQ	1.03009			1.03009
8	10298	EQ	0.12637			0.12637
9	10351	MC4	1.341			1.341
10	10302	EQ	1.16441			1.16441
11	10317	EQ	1.81273			1.81273
12	10347	EQ	1.81767			1.81767
13	10340	GI	1.98187			1.98187
14	10344	MS5	1.05834			1.05834
15	10341	GI	2.17195			2.17195
16	9514	GI	0.89724			0.89724
17	10374	EQ	0.52966			0.52966
18	8609	MC4	0.36499			0.36499
19	10366	EQ	0.38042	0.32945		0.35494
20	9516	EQ	0.28886			0.28886
21	10322	EQ	-0.24088			-0.24088
22	8593	MC4	-0.19898			-0.19898
23	8603	MC4	-0.44189			-0.44189
24	10378	MC4	-2.47548			-2.47548
25	8613	MC4	-1.05036			-1.05036
26	13863	MC4	-1.02788			-1.02788
27	10303	MC4	-1.12478			-1.12478
28	10288	EQ	-0.97446			-0.97446
29	13865	MC4	-0.8631			-0.8631
30	10362	MI	-0.5352			-0.5352
31	13856	MC4	-0.56928			-0.56928
32	10349	MC4	-0.06195			-0.06195
33	13867	MC4	1.17205			1.17205
34	13857	MC4	0.53956			0.53956
35	10375	MC4	1.72141			1.72141
36	10352	MC4	1.07916			1.07916
37	10309	MI	0.97888			0.97888
38	10369	MC4	0.25616			0.25616

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10331	EQ	0.38423			0.38423
40	9508	EQ	1.88124			1.88124
41	13860	MC4	-0.51315			-0.51315
42	10318	MC4	-0.85168			-0.85168
43	10299	EQ	-0.49012			-0.49012
44	10313	MC4	-0.82795			-0.82795
45	10379	GI	0.47078			0.47078
46	10371	MC4	-1.15822			-1.15822
47	13859	MC4	-1.02675			-1.02675

Appendix H.15—Spring 16 Operational Item Parameter Estimates — Grade 8 Mathematics

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10488	TI	-2.19043			-2.19043
2	10530	MC4	-0.83532			-0.83532
3	10567	EQ	-0.90128			-0.90128
4	10561	MC4	-0.22329			-0.22329
5	9532	EQ	0.73485			0.73485
6	10532	EQ	0.16579			0.16579
7	9521	EQ	1.11345			1.11345
8	10541	EQ	0.75865			0.75865
9	10585	EQ	1.6013			1.6013
10	10570	MI	1.10187			1.10187
11	9522	EQ	2.50544			2.50544
12	10579	EQ	3.02045			3.02045
13	9518	EQ	2.17755			2.17755
14	10523	EQ	2.68961			2.68961
15	10542	EQ	1.16308			1.16308
16	10538	EQ	1.41298			1.41298
17	10564	EQ	0.40128			0.40128
18	10580	EQ	0.96886			0.96886
19	9519	GI	-0.04329			-0.04329
20	10587	MC4	0.43524			0.43524
21	10513	MC4	-0.21755			-0.21755
22	10483	EQ	-0.67181			-0.67181
23	8623	MC4	-1.63504			-1.63504
24	10557	MC4	-3.16617			-3.16617
25	10562	MC4	-0.80153			-0.80153
26	10507	MC4	-1.82784			-1.82784
27	10498	MC4	-0.91777			-0.91777
28	10588	MC4	-1.06474			-1.06474
29	10554	MC4	-1.02849			-1.02849
30	10496	GI	-0.34275			-0.34275
31	10548	GI	0.35118			0.35118
32	9527	GI	-0.043			-0.043
33	10494	MS5	-0.08244			-0.08244
34	8651	MC4	-0.12423			-0.12423
35	10518	TI	1.44965			1.44965
36	9525	EQ	0.25872	0.32204		0.29038
37	10510	MC4	0.71001			0.71001
38	10487	MC4	-0.12623			-0.12623

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10527	EQ	0.11829			0.11829
40	10534	GI	0.04773			0.04773
41	10574	MS6	-0.26598			-0.26598
42	8635	MC4	-0.83239			-0.83239
43	10520	GI	-0.95493			-0.95493
44	8631	MC4	-1.22684			-1.22684
45	10525	MC4	-1.64422			-1.64422
46	10581	MC4	-1.92813			-1.92813
47	10528	EQ	-2.45491			-2.45491

Appendix H.16—Spring 16 Operational Item Parameter Estimates — Algebra I

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
1		10882	EQ	0.8083			0.8083
2		10888	EQ	1.89687			1.89687
3		10960	MC4	0.60667			0.60667
4		10951	MC4	0.71118			0.71118
5		11013	MS5	0.09622			0.09622
6		9530	TI	0.69364			0.69364
7		10972	EQ	-0.09826			-0.09826
8		11045	MC4	-0.09418			-0.09418
9		10895	EQ	-0.05275			-0.05275
10		10977	MC4	-0.42252			-0.42252
11		10953	MC4	-0.75521			-0.75521
12		10934	EQ	-0.94776			-0.94776
13		9707	GI	-2.17345			-2.17345
14		10887	MC4	-1.07459			-1.07459
15		10907	MC4	-1.12152			-1.12152
16		9705	GI	-0.96679			-0.96679
17		10974	MC4	-0.58968			-0.58968
18		11058	MC4	0.27251			0.27251
19		10889	MC4	-0.16663			-0.16663
20		10966	MC4	0.15284			0.15284
21		10943	MC4	0.46515			0.46515
22		10990	MC4	0.65596			0.65596
23		10978	MS5	1.8257			1.8257
24		11044	EQ	1.12987			1.12987
25		9546	EQ	0.79976			0.79976
26		9536	GI	1.36497			1.36497
27		11011	EQ	0.32291			0.32291
28		10896	MC4	0.61086			0.61086
29		11004	MC4	0.25768			0.25768
30		10942	MC4	0.09573			0.09573
31		10935	MC4	-0.12131			-0.12131
32		10945	MC4	-0.65134			-0.65134
33		9543	GI	-0.70978	-0.39294		-0.55136
34		10993	MC4	-1.08396			-1.08396
35		10905	MC4	-0.97781			-0.97781
36		10973	MC4	-1.36424			-1.36424
37		11052	MC4	-1.26248			-1.26248
38		10963	MC4	-0.87904			-0.87904

Item		Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
				Step 1	Step 2	Step 3	
39		10877	MC4	-0.39859			-0.39859
40		9535	EQ	-0.04011			-0.04011
41		10880	MC4	1.31529			1.31529
42		10897	MC4	0.32317			0.32317
43		10906	MC4	0.18842			0.18842
44		9533	TI	0.7707			0.7707
45		10988	EQ	0.41304			0.41304
46		10941	MC4	1.52007			1.52007
47		9531	EQ	0.97351			0.97351

Appendix H.17—Spring 16 Operational Item Parameter Estimates — Geometry

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	11037	MC4	-3.17485			-3.17485
2	10919	MC4	-1.19939			-1.19939
3	11018	MC4	-0.61172			-0.61172
4	11072	MC4	-0.56959			-0.56959
5	11089	MS6	-0.7697			-0.7697
6	10924	EQ	-0.20101			-0.20101
7	9556	EQ	-0.21949			-0.21949
8	10912	MC4	0.70057			0.70057
9	10926	EQ	0.4718			0.4718
10	9554	HT	0.94194			0.94194
11	9722	MI	1.01157			1.01157
12	9592	GI	1.50932			1.50932
13	9575	EQ	2.83952			2.83952
14	11017	HT	0.72651			0.72651
15	11074	MS5	0.57924			0.57924
16	9581	HT	0.20433			0.20433
17	11008	MC4	0.53528			0.53528
18	11007	EQ	-0.74951			-0.74951
19	9560	HT	0.32402			0.32402
20	11068	MC4	-0.64576			-0.64576
21	11035	MC4	-0.58779			-0.58779
22	10986	MC4	-1.15355			-1.15355
23	11061	MC4	-1.23202			-1.23202
24	10921	MC4	-1.82488			-1.82488
25	11086	MC4	-1.06633			-1.06633
26	11033	MC4	-1.06679			-1.06679
27	11016	EQ	-0.8117			-0.8117
28	11040	EQ	-0.73442			-0.73442
29	11078	MC4	-0.56802			-0.56802
30	11029	MC4	-0.13274			-0.13274
31	9564	EQ	0.43952			0.43952
32	11085	HT	0.56151			0.56151
33	11032	HT	0.86558			0.86558
34	11065	EQ	1.01365			1.01365
35	10998	EQ	3.68979			3.68979
36	9551	HT	2.06285			2.06285
37	11081	MS5	1.69723			1.69723
38	11092	MS6	0.93109			0.93109

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	11060	MC4	0.47251			0.47251
40	10930	EQ	0.73492			0.73492
41	11034	MC4	-0.04112			-0.04112
42	11114	HT	-0.12484			-0.12484
43	11036	MC4	-0.31518			-0.31518
44	11026	HT	-0.76141			-0.76141
45	9547	EQ	-1.26142			-1.26142
46	10910	MC4	-0.7847			-0.7847
47	11015	EQ	-1.64967			-1.64967

Appendix H.18—Spring 16 Operational Item Parameter Estimates — Algebra II

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
1	10215	MC4	-2.47281			-2.47281
2	10175	MC4	-1.28953			-1.28953
3	10210	MC4	-1.79638			-1.79638
4	10236	MC4	-0.89747			-0.89747
5	11121	MC4	-1.0416			-1.0416
6	9570	EQ	-0.10457			-0.10457
7	10245	EQ	0.65995			0.65995
8	9578	EQ	0.25925			0.25925
9	10164	EQ	0.69032			0.69032
10	9577	EQ	1.22957			1.22957
11	10182	EQ	1.37683			1.37683
12	10249	EQ	1.06528			1.06528
13	10256	GI	1.04876			1.04876
14	9548	EQ	2.2849			2.2849
15	10255	EQ	1.17865			1.17865
16	9567	EQ	0.98222			0.98222
17	10261	EQ	0.50767			0.50767
18	9549	EQ	0.03224			0.03224
19	10187	MC4	-0.17866			-0.17866
20	10233	EQ	-0.81277			-0.81277
21	10160	MC4	-0.65505			-0.65505
22	10192	MC4	-1.21466			-1.21466
23	10214	MC4	-1.65318			-1.65318
24	10206	MC4	-2.46401			-2.46401
25	10200	MC4	-1.46483			-1.46483
26	10204	HT	-1.37366			-1.37366
27	10203	MC4	-1.02449			-1.02449
28	10177	MC4	-0.56882			-0.56882
29	10259	MC4	-0.09362			-0.09362
30	10240	MS5	0.00695			0.00695
31	9568	EQ	0.85211			0.85211
32	10220	MC4	0.5445			0.5445
33	10168	MS5	1.69076			1.69076
34	10176	MS5	0.82007			0.82007
35	9589	EQ	2.2194			2.2194
36	9591	EQ	1.9907			1.9907
37	10180	EQ	2.12488			2.12488
38	10199	HT	0.94861			0.94861

Item	Item ID	Item Type	Item Parameter Estimates			Average Rasch Value
			Step 1	Step 2	Step 3	
39	10223	EQ	0.48221			0.48221
40	10228	MI	0.55259			0.55259
41	9573	EQ	0.31592			0.31592
42	10230	MC4	-0.3605			-0.3605
43	9580	EQ	-0.05533			-0.05533
44	10209	MC4	-0.65893			-0.65893
45	10243	MC4	-0.46516			-0.46516
46	10217	MC4	-1.30014			-1.30014
47	10237	MC4	-1.70996			-1.70996

Appendix I.1 – Standard Errors of Measurement at Performance Level Cuts Spring 2016 – ELA

	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 3 ELA	10.35	9.46	10.15	12.50	10.46
Grade 4 ELA	10.68	10.00	10.41	13.09	10.82
Grade 5 ELA	10.67	10.00	10.50	12.93	10.80
Grade 6 ELA	10.52	9.62	10.34	13.22	10.47
Grade 7ELA	10.39	9.00	10.00	12.97	10.16
Grade 8 ELA	10.01	9.05	10.25	12.61	10.08
Grade 9 ELA	10.22	9.00	9.91	12.31	10.04
Grade 10 ELA	9.74	9.00	9.68	11.86	9.84
Grade 11 ELA	9.93	9.00	9.46	12.00	9.94

Appendix I.2 – Standard Errors of Measurement at Performance Level Cuts Spring 2016– Mathematics

	Minimally Proficient	Partially Proficient	Proficient	Highly Proficient	Overall
Grade 3 Math	11.97	10.74	11.89	17.34	12.70
Grade 4 Math	11.53	10.00	11.07	15.54	11.49
Grade 5 Math	12.32	10.00	10.68	13.72	11.48
Grade 6 Math	12.50	10.00	10.42	13.83	11.67
Grade 7 Math	11.11	10.00	10.18	13.31	10.90
Grade8 Math	11.81	10.00	10.20	12.68	11.24
Algebra I	11.39	9.55	9.55	13.21	10.76
Geometry	13.65	10.39	10.00	12.60	11.90
Algebra II	13.74	11.11	10.00	11.27	12.20

Appendix J.1 Writing Prompt Rater Agreement Report - Grade 3 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,155	4,155	0.9	16.2	**65.8**	16.1	0.9	0	0
	Evidence/Elaboration	4,155	4,155	1	16.5	**65.1**	16.3	1	0	0
	Conventions	4,155	4,155	1.1	14.5	**68.7**	14.6	1.1	0	0
2	Purpose/Organization	4,176	4,176	1.4	18.2	**60.6**	18.3	1.4	0	0
	Evidence/Elaboration	4,176	4,176	0.9	17.4	**63.5**	17.5	0.9	0	0
	Conventions	4,176	4,176	1.9	15.3	**65.5**	15.3	1.9	0	0
3	Purpose/Organization	4,153	4,153	0.7	17.3	**64**	17.2	0.7	0	0
	Evidence/Elaboration	4,153	4,153	0.7	16.5	**65.7**	16.4	0.7	0	0
	Conventions	4,153	4,153	1	13.7	**70.8**	13.6	1	0	0
4	Purpose/Organization	4,186	4,186	0.9	17.3	**63.7**	17.3	0.9	0	0
	Evidence/Elaboration	4,186	4,186	1.4	18.4	**60.3**	18.4	1.4	0	0
	Conventions	4,186	4,186	0.5	14	**71.1**	14	0.5	0	0
5	Purpose/Organization	4,176	4,176	1.1	17.1	**63.4**	17.2	1.1	0	0
	Evidence/Elaboration	4,176	4,176	1.1	17.6	**62.5**	17.6	1.1	0	0
	Conventions	4,176	4,176	0.8	13.5	**71.4**	13.5	0.7	0	0
6	Purpose/Organization	4,175	4,175	1.2	16.8	**64**	16.9	1.2	0	0
	Evidence/Elaboration	4,175	4,175	1.7	17.5	**61.5**	17.6	1.7	0	0
	Conventions	4,175	4,175	0.7	14.2	**70.2**	14.2	0.7	0	0
Paper Essay	Purpose/Organization	28,556	5,526	0.4	3.4	**92.5**	3.4	0.4	0	0
	Evidence/Elaboration	28,556	5,526	0.3	3.6	**92.2**	3.6	0.3	0	0
	Conventions	28,556	5,526	0.2	2.8	**94**	2.8	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.2 Writing Prompt Rater Agreement Report - Grade 4 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,190	4,190	0.3	12.7	**74**	12.7	0.3	0	0
	Evidence/Elaboration	4,190	4,190	0.1	11.3	**77.2**	11.3	0.1	0	0
	Conventions	4,190	4,190	1	15.8	**66.4**	15.8	1	0	0
2	Purpose/Organization	4,189	4,189	0.4	14.2	**70.9**	14.2	0.4	0	0
	Evidence/Elaboration	4,189	4,189	0.2	12.3	**75.1**	12.3	0.2	0	0
	Conventions	4,189	4,189	0.3	17.8	**63.7**	17.8	0.3	0	0
3	Purpose/Organization	4,188	4,188	0.3	14	**71.5**	14	0.3	0	0
	Evidence/Elaboration	4,188	4,188	0.2	13	**73.5**	13	0.2	0	0
	Conventions	4,188	4,188	0.3	16.6	**66.2**	16.6	0.3	0	0
4	Purpose/Organization	4,192	4,192	0.4	13.8	**71.8**	13.8	0.4	0	0
	Evidence/Elaboration	4,192	4,192	0.2	11.5	**76.8**	11.5	0.2	0	0
	Conventions	4,192	4,192	0.2	16.2	**67.1**	16.2	0.2	0	0
5	Purpose/Organization	4,192	4,192	0.4	14.1	**71.1**	14.1	0.4	0	0
	Evidence/Elaboration	4,192	4,192	0.3	11.9	**75.7**	11.9	0.3	0	0
	Conventions	4,192	4,192	0.5	17.3	**64.2**	17.3	0.5	0	0
6	Purpose/Organization	4,182	4,182	0.1	12.7	**74.4**	12.7	0.1	0	0
	Evidence/Elaboration	4,182	4,182	0.2	11	**77.5**	11	0.2	0	0
	Conventions	4,182	4,182	0.3	16.3	**66.8**	16.3	0.3	0	0
Paper Essay	Purpose/Organization	28,026	5,098	0.1	11.7	**76.3**	11.7	0.1	0	0
	Evidence/Elaboration	28,026	5,098	0.2	11	**77.6**	11	0.2	0	0
	Conventions	28,026	5,098	0.2	12.6	**74.4**	12.6	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.3 Writing Prompt Rater Agreement Report - Grade 5 ELA

				Rater Agreement						
Online Prompt	Dimension	Total Reads	Second Reads	Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,186	4,186	0.2	14.5	**70.5**	14.5	0.2	0	0
	Evidence/Elaboration	4,186	4,186	0.2	14	**71.5**	14	0.2	0	0
	Conventions	4,186	4,186	0.2	13.6	**72.3**	13.6	0.2	0	0
2	Purpose/Organization	4,198	4,198	0.1	14.6	**70.5**	14.6	0.1	0	0
	Evidence/Elaboration	4,198	4,198	0.2	14.4	**70.8**	14.4	0.2	0	0
	Conventions	4,198	4,198	0.2	12.8	**74.0**	12.8	0.2	0	0
3	Purpose/Organization	4,190	4,190	0.2	13.9	**71.8**	13.9	0.2	0	0
	Evidence/Elaboration	4,190	4,190	0.3	13.4	**72.6**	13.4	0.3	0	0
	Conventions	4,190	4,190	0.2	13.4	**72.7**	13.4	0.2	0	0
4	Purpose/Organization	4,196	4,196	0.3	14.8	**69.9**	14.8	0.3	0	0
	Evidence/Elaboration	4,196	4,196	0.3	15.1	**69.1**	15.2	0.3	0	0
	Conventions	4,196	4,196	0.2	13.6	**72.3**	13.6	0.2	0	0
5	Purpose/Organization	4,192	4,192	0.2	14.2	**71.1**	14.2	0.2	0	0
	Evidence/Elaboration	4,192	4,192	0.5	16.8	**65.3**	16.8	0.5	0	0
	Conventions	4,192	4,192	0.5	12.9	**73.2**	12.9	0.5	0	0
6	Purpose/Organization	4,194	4,194	0.4	15.1	**69.1**	15.1	0.4	0	0
	Evidence/Elaboration	4,194	4,194	0.5	17	**65.0**	17	0.5	0	0
	Conventions	4,194	4,194	0.1	13.7	**72.4**	13.7	0.1	0	0
Paper Essay	Purpose/Organization	27,955	5,088	0.2	8.3	**83.1**	8.3	0.2	0	0
	Evidence/Elaboration	27,955	5,088	0.2	7.6	**84.6**	7.6	0.2	0	0
	Conventions	27,955	5,088	0.1	6.4	**86.9**	6.4	0.1	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.4 Writing Prompt Rater Agreement Report - Grade 6 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,190	4,190	1	17.8	**62.4**	17.8	1	0	0
	Evidence/Elaboration	4,190	4,190	0.8	15.7	**66.9**	15.7	0.8	0	0
	Conventions	4,190	4,190	0.5	16.4	**66.2**	16.4	0.5	0	0
2	Purpose/Organization	4,194	4,194	0.9	17.4	**63.4**	17.4	0.9	0	0
	Evidence/Elaboration	4,194	4,194	0.9	14.5	**69.1**	14.5	0.9	0	0
	Conventions	4,194	4,194	1.2	17.5	**62.5**	17.5	1.2	0	0
3	Purpose/Organization	4,186	4,186	1.4	18	**61.2**	18	1.4	0	0
	Evidence/Elaboration	4,186	4,186	1.5	16.1	**64.8**	16.1	1.5	0	0
	Conventions	4,186	4,186	0.6	15.2	**68.2**	15.2	0.6	0	0
4	Purpose/Organization	4,194	4,194	1.3	18	**61.5**	18	1.3	0	0
	Evidence/Elaboration	4,194	4,194	1.4	17.2	**62.8**	17.2	1.4	0	0
	Conventions	4,194	4,194	0.5	16.2	**66.6**	16.2	0.5	0	0
5	Purpose/Organization	4,190	4,190	0.6	16.7	**65.4**	16.7	0.6	0	0
	Evidence/Elaboration	4,190	4,190	0.2	14.8	**69.8**	14.8	0.2	0	0
	Conventions	4,190	4,190	0.7	16.1	**66.4**	16.1	0.7	0	0
6	Purpose/Organization	4,194	4,194	0.6	17.5	**63.8**	17.5	0.6	0	0
	Evidence/Elaboration	4,194	4,194	0.5	17.3	**64.5**	17.3	0.5	0	0
	Conventions	4,194	4,194	0.6	13.9	**71.1**	13.9	0.6	0	0
Paper Essay	Purpose/Organization	30,154	5,695	0.3	10.6	**78.1**	10.6	0.3	0	0
	Evidence/Elaboration	30,154	5,695	0.1	8.8	**82.2**	8.8	0.1	0	0
	Conventions	30,154	5,695	0.4	9.3	**80.6**	9.3	0.4	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.5 Writing Prompt Rater Agreement Report - Grade 7 ELA

Online Prompt	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
1	Purpose/Organization	4,164	4,164	0.6	16.7	**65.4**	16.7	0.6	0	0
	Evidence/Elaboration	4,164	4,164	0.6	16.5	**65.7**	16.5	0.6	0	0
	Conventions	4,164	4,164	0.2	14.9	**69.7**	14.9	0.2	0	0
2	Purpose/Organization	4,168	4,168	0.2	13.5	**72.6**	13.5	0.2	0	0
	Evidence/Elaboration	4,168	4,168	0.2	13.9	**71.8**	13.9	0.2	0	0
	Conventions	4,168	4,168	0.2	10.5	**78.5**	10.5	0.2	0	0
3	Purpose/Organization	4,176	4,176	0.1	15	**69.8**	15	0.1	0	0
	Evidence/Elaboration	4,176	4,176	0.2	13.3	**73**	13.3	0.2	0	0
	Conventions	4,176	4,176	0.2	15.3	**68.9**	15.3	0.2	0	0
4	Purpose/Organization	4,170	4,170	0.5	15.3	**68.4**	15.3	0.5	0	0
	Evidence/Elaboration	4,170	4,170	0.4	13.3	**72.6**	13.3	0.4	0	0
	Conventions	4,170	4,170	0.1	11.6	**76.7**	11.6	0.1	0	0
5	Purpose/Organization	4,186	4,186	0.1	13.4	**73.1**	13.4	0.1	0	0
	Evidence/Elaboration	4,186	4,186	0.2	12.7	**74.2**	12.7	0.2	0	0
	Conventions	4,186	4,186	0.1	12.4	**75.1**	12.4	0.1	0	0
6	Purpose/Organization	4,180	4,180	0.1	14.4	**71**	14.4	0.1	0	0
	Evidence/Elaboration	4,180	4,180	0.2	13.3	**73**	13.3	0.2	0	0
	Conventions	4,180	4,180	0.1	14.1	**71.5**	14.1	0.1	0	0
Paper Essay	Purpose/Organization	29,075	5,353	0.1	10.9	**78.1**	10.9	0.1	0	0
	Evidence/Elaboration	29,075	5,353	0.1	10.1	**79.6**	10.1	0.1	0	0
	Conventions	29,075	5,353	0.2	9.5	**80.7**	9.5	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.6 Writing Prompt Rater Agreement Report - Grade 8 ELA

	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
Paper Essay	Purpose/Organization	28,253	5,136	0.8	15.3	**67.8**	15.3	0.8	0	0.1
	Evidence/Elaboration	28,253	5,136	0.7	15.3	**67.8**	15.3	0.7	0	0.1
	Conventions	28,253	5,136	0.5	8.3	**82.3**	8.3	0.5	0	0.1

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.7 ELA Writing Prompt Rater Agreement Report - Grade 9 ELA

	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
Paper Essay	Purpose/Organization	33,077	6,026	0.4	13.2	**72.8**	13.2	0.4	0	0
	Evidence/Elaboration	33,077	6,026	0.7	12.7	**73.2**	12.7	0.7	0	0
	Conventions	33,077	6,026	0.1	6.8	**86.2**	6.8	0.1	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.8 ELA Writing Prompt Rater Agreement Report - Grade 10 ELA

	Dimension	Total Reads	Second Reads	Rater Agreement						
				Low	Adj Low	Equal	Adj High	High	Mismatch CC	MM CC/Score
Paper Essay	Purpose/Organization	30,323	5,520	0.3	11.9	**75.5**	11.9	0.3	0	0
	Evidence/Elaboration	30,323	5,520	0.5	11.7	**75.6**	11.7	0.5	0	0
	Conventions	30,323	5,520	0.2	6.5	**86.6**	6.5	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix J.9 Writing Prompt Rater Agreement Report - Grade 11 ELA

				Rater Agreement						
Online Prompt	**Dimension**	**Total Reads**	**Second Reads**	**Low**	**Adj Low**	**Equal**	**Adj High**	**High**	**Mismatch CC**	**MM CC/Score**
1	Purpose/Organization	4,190	4,190	0.5	17.4	**64.3**	17.4	0.5	0	0
	Evidence/Elaboration	4,190	4,190	0.8	18.4	**61.6**	18.4	0.8	0	0
	Conventions	4,190	4,190	0.2	12.2	**75.2**	12.2	0.2	0	0
2	Purpose/Organization	4,166	4,166	0.6	16.5	**65.8**	16.5	0.6	0	0
	Evidence/Elaboration	4,166	4,166	0.9	17.2	**63.8**	17.2	0.9	0	0
	Conventions	4,166	4,166	0.2	14.5	**70.7**	14.5	0.2	0	0
3	Purpose/Organization	4,176	4,176	0.9	15.7	**66.8**	15.8	0.9	0	0
	Evidence/Elaboration	4,176	4,176	1.4	17.6	**62**	17.6	1.4	0	0
	Conventions	4,176	4,176	0.4	12.5	**74.2**	12.5	0.4	0	0
4	Purpose/Organization	4,172	4,172	0.7	17.4	**63.7**	17.5	0.7	0	0
	Evidence/Elaboration	4,172	4,172	1	17.4	**63.2**	17.4	1	0	0
	Conventions	4,172	4,172	0.2	13.2	**73.2**	13.3	0.2	0	0
5	Purpose/Organization	4,180	4,180	0.6	16.6	**65.5**	16.6	0.6	0	0
	Evidence/Elaboration	4,180	4,180	1.1	19.1	**59.5**	19.1	1.1	0	0
	Conventions	4,180	4,180	0.2	12.6	**74.4**	12.6	0.2	0	0
6	Purpose/Organization	4,180	4,180	0.6	16.6	**65.8**	16.6	0.6	0	0
	Evidence/Elaboration	4,180	4,180	1.1	17.8	**62.3**	17.8	1.1	0	0
	Conventions	4,180	4,180	0.1	10.9	**78**	10.9	0.1	0	0
Paper Essay	Purpose/Organization	26,893	4,886	0.4	8.8	**81.6**	8.8	0.4	0	0
	Evidence/Elaboration	26,893	4,886	0.7	9.2	**80.2**	9.2	0.7	0	0
	Conventions	26,893	4,886	0.2	4.5	**90.5**	4.5	0.2	0	0

Note: Perfect Agreement = Equal; Adjacent Agreement = Adj Low or Adj High; Nonadjacent Agreement = Low or High; Mismatched Scores = Mismatch CC; Nonscorable/Scorable = MM CC/Score

Appendix K.1—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 3 ELA

No. Item	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	835	-1.77	-1.36	-1.55	-0.19
2	2	837	-0.41	-0.14	-0.18	-0.04
3	3	8708	-2.08	-1.37	-1.85	-0.49
4	4	8710	-1.51	-1.13	-1.28	-0.15
5	5	8711	-0.33	-0.14	-0.11	0.04
6	6	9330	-0.49	-0.12	-0.26	-0.15
7	7	9331	0.40	0.54	0.63	0.09
8	8	9333	0.08	0.45	0.31	-0.14
9	9	9335	-0.87	-0.21	-0.65	-0.44
10	10	9336	-0.32	0.05	-0.10	-0.14
11	11	9347	0.05	0.16	0.28	0.12
12	12	9349	0.59	1.17	0.81	-0.35
13	13	9353	-0.12	0.10	0.10	0.01
14	14	9355	-0.59	-0.24	-0.36	-0.12
15	15	9356	1.08	0.96	1.30	0.34
16	16	9357	0.60	0.62	0.82	0.20
17	17	9359	0.05	0.20	0.28	0.08
18	18	9377	-0.04	0.12	0.19	0.07
19	19	9378	-1.31	-1.33	-1.09	0.24
19	20	9378	-0.58	-0.23	-0.36	-0.13
20	21	9379	-1.26	-0.68	-1.03	-0.35
21	22	9380	-1.11	-0.91	-0.88	0.03
22	23	9398	-0.28	0.22	-0.06	-0.27
23	24	9400	0.27	0.44	0.50	0.05
24	25	9401	0.87	1.23	1.09	-0.14
25	26	9402	1.43	1.71	1.66	-0.05
26	27	9404	-0.49	-0.34	-0.26	0.08
27	28	9414	-0.26	-0.42	-0.03	0.39
28	29	9418	0.48	0.51	0.71	0.20
29	30	9422	-1.10	-0.87	-0.88	-0.01
30	31	9687	1.16	1.38	1.38	0.00
31	32	9690	-1.44	-1.36	-1.21	0.15
32	33	9691	-0.60	-0.64	-0.38	0.27
33	34	9692	-0.83	-0.57	-0.60	-0.03
34	35	9694	1.48	1.36	1.71	0.35
35	36	9697	-0.49	-0.27	-0.27	0.00
36	37	9698	-1.34	-1.31	-1.12	0.19
37	38	9699	-0.53	-0.05	-0.30	-0.25
38	39	9700	1.01	0.87	1.24	0.36
39	40	10268	-0.03	0.44	0.20	-0.24
40	41	10632	0.25	0.54	0.48	-0.06
41	42	10634	0.95	0.72	1.18	0.46
Mean			**-0.23**	**0.00**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.23**	

Appendix K.2—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 4 ELA

No. Item	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	1122	0.41	1.09	0.41	-0.68
2	2	1124	-1.62	-0.21	-1.62	-1.41
3	3	9382	-2.14	-0.96	-2.14	-1.18
4	4	9386	-0.49	0.74	-0.49	-1.23
5	5	9387	0.52	1.21	0.52	-0.69
6	6	9388	-1.29	-0.24	-1.29	-1.05
7	7	9389	-0.92	0.23	-0.92	-1.15
8	8	9390	-1.85	-0.48	-1.85	-1.37
9	9	9391	-0.69	0.38	-0.69	-1.07
10	10	9392	-1.64	-0.36	-1.64	-1.28
11	11	9395	-1.24	-0.16	-1.24	-1.08
12	12	9397	-0.15	0.98	-0.15	-1.12
13	13	9425	-1.60	-0.60	-1.60	-1.00
14	14	9426	-0.03	1.05	-0.03	-1.08
15	15	9428	-0.16	0.68	-0.16	-0.84
16	16	9429	-3.17	-2.16	-3.17	-1.01
16	17	9429	-0.38	0.38	-0.38	-0.76
17	18	9431	-0.48	0.56	-0.48	-1.05
18	19	9437	-0.99	-0.19	-0.99	-0.80
19	20	9438	0.80	1.18	0.80	-0.38
20	21	9439	1.13	1.50	1.13	-0.37
21	22	9446	-1.73	-1.01	-1.73	-0.72
22	23	9450	-1.13	-0.24	-1.13	-0.89
23	24	9451	0.22	0.48	0.22	-0.26
24	25	9595	-1.38	-0.86	-1.38	-0.52
25	26	9596	-0.31	0.61	-0.31	-0.91
26	27	9597	-1.05	-0.16	-1.05	-0.89
27	28	9598	-0.97	-0.18	-0.97	-0.79
28	29	9603	1.77	2.60	1.77	-0.83
29	30	9616	-1.15	0.09	-1.15	-1.24
30	31	9899	-0.43	0.47	-0.43	-0.90
31	32	9900	1.77	2.27	1.77	-0.50
32	33	9902	-2.17	-1.22	-2.17	-0.95
33	34	9903	-0.40	0.42	-0.40	-0.81
34	35	9906	-1.50	-0.53	-1.50	-0.96
35	36	10263	-1.68	-0.61	-1.68	-1.07
36	37	10266	-1.70	-0.61	-1.70	-1.10
37	38	10277	0.16	0.85	0.16	-0.69
38	39	10278	1.17	1.78	1.17	-0.61
39	40	10644	-1.01	-0.09	-1.01	-0.92
40	41	10645	-3.05	-0.90	-3.05	-2.15
40	42	10645	-0.84	0.98	-0.84	-1.82
41	43	10647	-1.81	-0.82	-1.81	-1.00
Mean			**-0.77**	**0.18**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.96**	

Appendix K.3—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 5 ELA

No. Item	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	9290	-1.38	-0.74	-0.71	0.03
2	2	9291	-1.82	-1.07	-1.14	-0.07
2	3	9291	0.02	0.37	0.70	0.32
3	4	9292	-1.76	-1.15	-1.08	0.07
3	5	9292	-0.11	0.38	0.57	0.19
4	6	9294	-1.72	-1.16	-1.04	0.12
5	7	9298	-1.50	-0.40	-0.82	-0.42
6	8	9299	-0.81	-0.09	-0.14	-0.05
7	9	9302	-0.28	0.11	0.40	0.28
8	10	9303	-0.53	-0.18	0.14	0.33
9	11	9304	-0.15	0.27	0.53	0.26
10	12	9305	-1.74	-1.23	-1.07	0.16
11	13	9306	-1.01	-0.13	-0.34	-0.20
12	14	9307	-1.35	-0.45	-0.68	-0.23
13	15	9308	-1.00	0.28	-0.32	-0.61
14	16	9309	2.72	3.37	3.40	0.03
15	17	9311	1.43	2.27	2.11	-0.16
16	18	9312	2.05	3.51	2.72	-0.78
17	19	9320	0.50	1.44	1.18	-0.26
18	20	9599	-0.56	0.02	0.12	0.10
19	21	9600	-0.38	0.74	0.29	-0.45
20	22	9601	-0.88	-0.39	-0.20	0.19
21	23	9751	-0.91	-0.17	-0.24	-0.07
22	24	9752	0.52	1.41	1.20	-0.21
23	25	9753	0.14	0.90	0.82	-0.08
24	26	9754	2.56	3.10	3.24	0.13
25	27	9755	0.57	1.32	1.25	-0.07
26	28	9757	-0.76	0.05	-0.08	-0.14
27	29	9758	-1.16	-0.78	-0.48	0.30
28	30	9762	-0.62	-0.20	0.05	0.25
29	31	9763	-0.20	-0.06	0.48	0.54
30	32	9765	0.73	1.07	1.41	0.34
31	33	9767	-0.61	0.25	0.07	-0.18
32	34	9783	-1.54	-0.77	-0.86	-0.09
33	35	9784	0.06	0.44	0.74	0.30
34	36	9808	-0.48	0.24	0.20	-0.04
35	37	9809	-0.43	0.23	0.24	0.01
36	38	9833	0.13	0.70	0.81	0.11
37	39	9842	-0.05	0.45	0.63	0.18
38	40	10264	1.95	2.61	2.63	0.02
39	41	10659	-1.50	-0.97	-0.83	0.14
40	42	10661	-2.88	-2.09	-2.20	-0.11
40	43	10661	-1.34	-0.43	-0.67	-0.24
41	44	10662	-1.69	-1.09	-1.01	0.07
Mean			**-0.40**	**0.27**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.68**	

Appendix K.4—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 6 ELA

No. Item	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	9083	-2.86	-2.27	-2.37	-0.10
1	2	9083	0.21	0.49	0.69	0.20
2	3	9084	-1.01	-0.52	-0.53	-0.01
2	4	9084	1.46	1.80	1.95	0.15
3	5	9108	-1.80	-1.19	-1.32	-0.13
3	6	9108	0.76	1.12	1.24	0.12
4	7	9109	-1.98	-1.50	-1.49	0.01
4	8	9109	0.14	0.22	0.62	0.40
5	9	9130	-0.94	-0.93	-0.45	0.48
6	10	9131	0.39	0.72	0.88	0.16
7	11	9134	-0.73	-0.06	-0.24	-0.18
8	12	9135	0.04	0.45	0.53	0.08
9	13	9138	-1.96	-1.66	-1.48	0.18
10	14	9139	-1.39	-0.50	-0.91	-0.41
11	15	9140	-0.19	0.47	0.29	-0.17
12	16	9142	1.11	1.72	1.59	-0.13
13	17	9143	-0.10	0.53	0.39	-0.14
14	18	9145	-1.44	-0.50	-0.96	-0.46
15	19	9153	-0.90	-0.68	-0.42	0.27
16	20	9168	-0.13	0.14	0.35	0.21
17	21	9169	0.94	0.99	1.43	0.44
18	22	9262	0.70	1.28	1.19	-0.09
19	23	9263	-1.00	-0.14	-0.51	-0.37
20	24	9264	0.53	0.81	1.01	0.20
21	25	9266	-1.53	-0.90	-1.04	-0.14
22	26	9267	-1.08	-0.48	-0.59	-0.11
23	27	9268	1.60	2.10	2.08	-0.02
24	28	9272	-0.49	-0.16	-0.01	0.15
25	29	9797	0.49	1.00	0.98	-0.02
26	30	9798	-0.37	-0.44	0.11	0.55
27	31	9799	-0.65	-0.15	-0.17	-0.02
28	32	9800	0.73	1.44	1.21	-0.23
29	33	9802	-0.06	0.35	0.42	0.07
30	34	9803	-2.25	-1.75	-1.77	-0.02
31	35	9804	-0.98	-0.37	-0.50	-0.13
32	36	9807	0.68	1.33	1.17	-0.16
33	37	9865	0.17	1.65	0.65	-1.00
34	38	9866	-0.41	0.03	0.07	0.04
35	39	9867	0.40	0.59	0.88	0.30
36	40	9869	-1.77	-1.52	-1.28	0.24
37	41	9870	-0.67	-0.21	-0.19	0.02
38	42	9872	1.39	1.75	1.88	0.13
39	43	9889	-0.44	0.10	0.05	-0.06
40	44	10265	-1.53	-0.74	-1.05	-0.31
41	45	10280	-0.34	0.13	0.15	0.01
Mean			**-0.38**	**0.10**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.48**	

Appendix K.5—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 7 ELA

No. Item	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	9103	-0.25	0.13	0.33	0.20
2	2	9104	-1.80	-1.26	-1.22	0.04
2	3	9104	0.04	0.29	0.62	0.33
3	4	9105	-0.72	-0.33	-0.14	0.18
4	5	9128	0.70	1.44	1.29	-0.15
5	6	9146	-1.78	-1.18	-1.20	-0.01
6	7	9147	-1.13	-0.16	-0.54	-0.38
7	8	9152	-0.83	-0.20	-0.25	-0.05
8	9	9176	0.85	1.69	1.43	-0.26
9	10	9177	-0.04	0.73	0.55	-0.18
10	11	9189	0.55	0.79	1.13	0.35
11	12	9220	-0.81	-0.34	-0.23	0.12
12	13	9610	0.33	0.55	0.91	0.36
13	14	9611	-0.14	0.82	0.44	-0.38
14	15	9614	-1.60	-0.85	-1.02	-0.17
15	16	9709	-0.83	0.76	-0.24	-1.01
16	17	9711	-0.29	-0.12	0.30	0.42
17	18	9713	-0.35	0.47	0.23	-0.24
18	19	9750	-0.76	0.51	-0.18	-0.68
19	20	9777	0.19	0.64	0.77	0.13
20	21	9778	-0.16	0.34	0.43	0.09
21	22	9779	-0.26	0.16	0.32	0.16
22	23	9781	-0.36	0.09	0.23	0.14
23	24	9785	0.78	1.13	1.36	0.22
24	25	9786	0.45	1.07	1.03	-0.03
25	26	9787	0.00	0.41	0.59	0.18
26	27	9789	-0.83	-0.45	-0.25	0.20
27	28	9791	-0.32	-0.14	0.27	0.41
28	29	9793	1.04	0.73	1.63	0.90
29	30	9811	-0.96	-0.43	-0.38	0.05
30	31	9817	-0.69	0.01	-0.10	-0.12
31	32	9962	0.01	0.53	0.59	0.06
32	33	9982	-0.21	0.62	0.37	-0.25
33	34	10274	1.12	1.48	1.71	0.22
34	35	10613	0.59	1.53	1.18	-0.35
35	36	10619	-0.69	0.31	-0.11	-0.41
36	37	10621	0.45	1.01	1.04	0.03
37	38	10623	-0.28	0.48	0.30	-0.18
38	39	10656	-0.58	0.05	0.01	-0.04
39	40	10657	-1.89	-1.51	-1.31	0.21
39	41	10657	-0.70	-0.21	-0.12	0.09
40	42	10658	-1.34	-0.86	-0.76	0.11
40	43	10658	1.24	1.58	1.82	0.24
41	44	10695	-1.10	0.02	-0.52	-0.53
Mean			**-0.30**	**0.28**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.58**	

Appendix K.6—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 8 ELA

No. Item	No. Step parameters	Item ID	Item Parameters			4. Parameter Drift
			1. Spring 2016 Operational	2. AIR bank	3. Post-equated	
1	1	8965	-1.72	-1.48	-1.59	-0.11
2	2	8967	-0.90	-0.36	-0.76	-0.40
3	3	8971	1.02	0.75	1.16	0.41
4	4	8973	-0.35	0.13	-0.22	-0.35
5	5	9014	-1.13	-0.91	-0.99	-0.08
6	6	9015	1.08	1.08	1.21	0.13
7	7	9017	-0.23	-0.44	-0.09	0.35
8	8	9018	0.68	0.56	0.81	0.25
9	9	9019	-0.09	-0.27	0.05	0.32
10	10	9020	0.96	1.45	1.09	-0.35
11	11	9022	0.86	1.22	0.99	-0.23
12	12	9023	0.82	0.69	0.96	0.26
13	13	9024	-0.67	-0.55	-0.54	0.01
14	14	9026	0.37	0.32	0.51	0.19
15	15	9028	0.04	-0.34	0.18	0.52
16	16	9044	0.76	0.94	0.89	-0.05
17	17	9046	-0.96	-1.12	-0.82	0.29
18	18	9076	-0.78	-0.28	-0.64	-0.37
19	19	9077	-1.93	-1.65	-1.80	-0.15
19	20	9077	0.20	0.28	0.33	0.05
20	21	9078	-1.07	-0.58	-0.94	-0.36
20	22	9078	0.67	0.99	0.81	-0.18
21	23	9113	-0.32	-0.30	-0.18	0.11
22	24	9115	0.61	0.62	0.75	0.13
23	25	9116	-0.42	-0.20	-0.28	-0.08
24	26	9170	0.37	0.65	0.50	-0.14
25	27	9171	1.71	1.74	1.84	0.10
26	28	9172	0.37	0.49	0.51	0.01
27	29	9230	-1.29	-1.16	-1.15	0.00
28	30	9240	-0.49	-0.25	-0.35	-0.10
29	31	9253	-0.76	-0.86	-0.62	0.24
30	32	9254	-0.46	-0.66	-0.32	0.34
31	33	9255	0.57	0.88	0.70	-0.18
32	34	9256	-0.38	0.08	-0.25	-0.33
33	35	9259	0.51	0.51	0.65	0.13
34	36	9728	-1.79	-1.75	-1.66	0.09
34	37	9728	-0.34	-0.25	-0.20	0.05
35	38	9729	-1.86	-1.61	-1.73	-0.12
35	39	9729	0.13	0.26	0.27	0.01
36	40	9880	-0.18	-0.09	-0.05	0.04
37	41	9881	0.43	0.50	0.56	0.06
38	42	9883	-1.62	-1.27	-1.48	-0.22
39	43	9885	-0.81	-0.46	-0.68	-0.22
40	44	10626	1.13	1.04	1.26	0.23
41	45	10639	2.78	3.21	2.92	-0.30
Mean			**-0.10**	**0.04**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.13**	

Appendix K.7—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 9 ELA

No. Item	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	1444	-0.65	-0.15	-0.42	-0.26
2	2	1446	-0.26	0.28	-0.03	-0.31
3	3	8948	-0.05	0.11	0.19	0.08
4	4	8951	-1.12	-0.68	-0.89	-0.21
4	5	8951	1.21	1.34	1.45	0.11
5	6	8987	-0.81	-0.88	-0.57	0.30
6	7	8989	-1.24	-0.80	-1.01	-0.21
7	8	8990	0.28	0.11	0.51	0.40
8	9	9003	0.07	-0.09	0.30	0.40
9	10	9004	-0.23	-0.12	0.01	0.13
10	11	9005	0.20	0.59	0.43	-0.16
11	12	9006	-0.16	0.22	0.07	-0.14
12	13	9007	-0.38	-0.08	-0.14	-0.07
13	14	9012	0.16	0.42	0.40	-0.02
14	15	9032	0.58	0.67	0.82	0.15
15	16	9033	-0.39	-0.26	-0.15	0.10
16	17	9034	0.95	0.96	1.18	0.22
17	18	9038	0.32	0.49	0.55	0.07
18	19	9040	0.77	0.40	1.00	0.60
19	20	9041	1.61	1.65	1.84	0.19
20	21	9042	0.36	0.34	0.59	0.25
21	22	9043	0.14	0.26	0.37	0.12
22	23	9047	0.81	1.43	1.04	-0.39
23	24	9048	-0.62	-0.32	-0.39	-0.07
24	25	9050	1.15	1.01	1.38	0.37
25	26	9051	-0.03	0.22	0.20	-0.01
26	27	9053	-1.07	-0.65	-0.84	-0.19
27	28	9058	0.13	0.72	0.36	-0.37
28	29	9060	1.04	1.58	1.28	-0.31
29	30	9063	-0.65	-0.28	-0.42	-0.14
30	31	9064	0.98	0.80	1.21	0.41
31	32	9065	0.06	0.46	0.29	-0.17
32	33	9066	0.14	0.40	0.37	-0.03
33	34	9069	0.79	0.83	1.02	0.19
34	35	9734	-0.62	-0.66	-0.39	0.28
35	36	9735	-1.25	-1.05	-1.01	0.03
35	37	9735	1.35	1.26	1.58	0.32
36	38	9736	-0.52	-0.44	-0.29	0.15
37	39	9737	-1.64	-1.52	-1.41	0.11
38	40	10595	-1.24	-0.57	-1.01	-0.44
39	41	10596	-1.86	-1.03	-1.63	-0.60
40	42	10597	-1.54	-0.56	-1.31	-0.74
41	43	10603	0.14	0.53	0.37	-0.16
42	44	11097	-0.68	-0.29	-0.45	-0.16
43	45	11098	0.18	0.22	0.42	0.20
Mean			**-0.08**	**0.15**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.23**	

Appendix K.8—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 10 ELA

No. Items	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	8757	-1.50	-1.10	-1.37	-0.27
2	2	8758	0.20	0.36	0.33	-0.03
3	3	8759	-1.30	-1.33	-1.18	0.16
3	4	8759	-0.35	-0.25	-0.22	0.04
4	5	8760	0.74	1.03	0.86	-0.16
5	6	8761	-0.23	0.13	-0.10	-0.23
6	7	8762	0.45	0.51	0.58	0.07
7	8	8764	-0.80	-0.43	-0.67	-0.24
7	9	8764	0.99	1.07	1.12	0.05
8	10	8785	-1.37	-0.99	-1.24	-0.25
9	11	8788	-0.92	-0.60	-0.80	-0.20
10	12	8795	1.40	1.59	1.53	-0.06
11	13	8815	0.43	0.68	0.56	-0.12
12	14	8816	0.52	0.51	0.65	0.14
13	15	8818	0.28	0.82	0.41	-0.41
14	16	8819	0.50	0.97	0.62	-0.34
15	17	8821	0.06	-0.24	0.19	0.42
16	18	8844	2.96	2.67	3.08	0.42
17	19	8925	-0.49	0.10	-0.36	-0.46
18	20	8926	0.30	0.64	0.42	-0.21
19	21	8972	0.33	0.41	0.46	0.05
20	22	9623	-0.46	-0.24	-0.33	-0.09
21	23	9624	0.84	1.15	0.97	-0.18
22	24	9626	0.27	0.40	0.40	0.00
23	25	9627	-0.11	-0.10	0.02	0.11
24	26	9630	0.41	0.47	0.54	0.07
25	27	9769	-0.11	-0.24	0.02	0.26
26	28	9771	-0.57	-0.32	-0.44	-0.12
27	29	9772	-1.40	-1.81	-1.27	0.54
28	30	9773	0.73	0.69	0.86	0.16
29	31	9774	-0.21	-0.07	-0.08	-0.01
30	32	9821	-0.09	-0.07	0.03	0.10
31	33	9822	-0.68	-1.19	-0.55	0.65
32	34	9824	0.70	1.01	0.83	-0.18
33	35	9825	-0.65	-0.37	-0.53	-0.16
34	36	9826	-0.57	-0.34	-0.44	-0.10
35	37	9827	0.21	0.42	0.33	-0.08
36	38	9829	-0.15	-0.30	-0.02	0.28
37	39	9830	1.69	1.91	1.82	-0.09
38	40	9831	-1.26	-1.27	-1.14	0.13
39	41	9837	-0.51	-0.32	-0.39	-0.07
40	42	9839	-0.20	-0.06	-0.07	0.00
41	43	9840	0.37	0.25	0.49	0.24
42	44	9887	1.86	1.98	1.99	0.01
43	45	9888	-0.16	-0.20	-0.03	0.17
Mean			**0.05**	**0.18**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**0.13**	

Appendix K.9—Linking Spring 16 Operational Parameters to the Bank Scale—Grade 11 ELA

No. Items	No. Step parameters	Item ID	Item Parameters: 1. Spring 2016 Operational	2. AIR bank	3. Post-equated	4. Parameter Drift
1	1	421	-0.11	-0.04	-0.19	-0.15
2	2	422	1.15	1.42	1.07	-0.35
3	3	8753	-1.22	-1.40	-1.30	0.10
4	4	8754	-1.20	-1.30	-1.28	0.03
5	5	8755	-0.41	-0.31	-0.49	-0.18
5	6	8755	1.74	1.75	1.66	-0.09
6	7	8769	0.47	0.39	0.39	0.01
7	8	8770	-1.49	-1.41	-1.57	-0.16
7	9	8770	0.33	0.32	0.25	-0.07
8	10	8771	-1.21	-1.37	-1.28	0.09
9	11	8781	0.42	0.76	0.34	-0.41
10	12	8783	-0.60	-1.12	-0.68	0.44
11	13	8784	-0.18	-0.69	-0.26	0.44
12	14	8791	1.39	0.81	1.31	0.50
13	15	8792	-0.14	-0.33	-0.21	0.11
14	16	8794	0.68	0.78	0.60	-0.17
15	17	8796	0.14	-0.04	0.06	0.10
16	18	8797	-0.28	-0.29	-0.36	-0.07
17	19	8798	0.25	0.18	0.17	-0.01
18	20	8799	-0.44	-0.48	-0.52	-0.04
19	21	8800	-0.70	-0.86	-0.77	0.09
20	22	8805	0.36	-0.09	0.28	0.37
21	23	8807	0.12	0.06	0.04	-0.03
22	24	8808	-0.35	-0.60	-0.43	0.17
23	25	8809	0.19	0.01	0.12	0.10
24	26	8834	-0.02	-0.47	-0.10	0.38
25	27	8837	1.15	1.16	1.08	-0.08
26	28	8841	0.14	-0.13	0.06	0.19
27	29	8843	0.90	0.91	0.82	-0.09
28	30	8846	0.30	-0.06	0.22	0.28
29	31	8856	0.62	0.22	0.54	0.32
30	32	8861	0.34	0.32	0.26	-0.06
31	33	8862	0.84	0.76	0.77	0.00
32	34	8865	0.65	0.44	0.57	0.13
33	35	8867	0.26	0.13	0.18	0.05
34	36	8871	0.12	0.02	0.04	0.02
35	37	8879	0.60	1.10	0.53	-0.58
36	38	8880	0.05	-0.24	-0.03	0.21
37	39	8881	-0.29	-0.69	-0.37	0.32
38	40	8884	0.34	0.31	0.26	-0.04
39	41	9852	-0.77	-0.25	-0.84	-0.59
40	42	9853	-0.67	-0.18	-0.75	-0.57
41	43	9856	-0.05	-0.02	-0.13	-0.11
42	44	9858	1.07	1.04	0.99	-0.05
43	45	9860	-0.26	0.21	-0.34	-0.55
Mean			**0.09**	**0.02**		
Constant	mean(2. AIR bank)-mean(1. Spring 2016 Operational)				**-0.08**	

Appendix K.10 — Field Test Analysis: Linking Operational Parameters to the Bank Scale—Grade 3 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10384	-2.07	-1.18	-1.36	-0.18
10391	-1.21	-0.23	-0.50	-0.26
10392	-0.08	0.63	0.63	-0.01
10395	-2.75	-1.62	-2.04	-0.42
10396	-0.33	0.10	0.38	0.28
10398	-0.26	0.81	0.45	-0.36
10399	0.18	1.28	0.89	-0.40
10400	-0.97	-0.63	-0.26	0.37
10402	-2.74	-1.93	-2.03	-0.10
10403	-1.29	-0.64	-0.58	0.06
10409	-2.58	-1.75	-1.87	-0.12
10411	-2.33	-1.43	-1.62	-0.19
10416	-0.39	0.72	0.32	-0.40
10418	-0.93	-0.36	-0.22	0.14
10425	-2.25	-1.52	-1.54	-0.01
10427	0.36	0.91	1.07	0.16
10430	1.61	2.22	2.32	0.11
10431	-1.79	-1.08	-1.08	0.00
10433	0.63	1.21	1.34	0.13
10434	-1.46	-0.86	-0.75	0.11
10436	-1.21	-0.16	-0.50	-0.35
10438	-3.05	-1.82	-2.34	-0.53
10439	-1.72	-0.67	-1.01	-0.35
10443	-3.04	-2.22	-2.33	-0.11
10446	0.58	1.07	1.29	0.22
10448	0.39	1.01	1.10	0.08
10450	-0.37	-0.06	0.34	0.40
10453	-4.10	-2.99	-3.39	-0.40
10454	1.57	2.12	2.28	0.17
10455	1.14	1.71	1.85	0.15
10466	0.12	0.78	0.83	0.05
10470	1.53	2.08	2.24	0.16
10671	-0.32	0.27	0.39	0.12
10683	0.72	1.28	1.43	0.15
10685	0.99	1.50	1.70	0.20
10687	0.19	0.79	0.90	0.10
11120	0.42	1.12	1.13	0.01
8461	-1.31	-0.87	-0.60	0.27
8463	-1.30	-0.57	-0.59	-0.02
8481	0.11	0.72	0.82	0.10
8483	-0.73	-0.85	-0.02	0.83
9455	-0.78	-0.11	-0.07	0.04
9460	-2.39	-1.49	-1.68	-0.18
9464	0.61	1.20	1.32	0.12
9469	0.86	1.71	1.57	-0.14
mean	-0.71	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.71	

Appendix K.11 — Field Test Analysis: Linking Operational Parameters to the Bank Scale—Grade 4 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10705	-0.48	0.07	0.15	0.09
10708	-1.00	-0.73	-0.36	0.37
10710	-1.07	-0.18	-0.44	-0.26
10715	0.11	0.43	0.74	0.31
10716	0.47	1.13	1.10	-0.03
10718	-0.60	-0.12	0.03	0.15
10719	-0.67	0.01	-0.04	-0.04
10724	0.50	1.02	1.13	0.11
10727	1.52	2.13	2.16	0.03
10728	0.66	1.07	1.29	0.22
10729	-1.61	-0.93	-0.97	-0.04
10730	-1.56	-0.50	-0.93	-0.43
10731	-1.41	-1.09	-0.78	0.31
10739	0.05	0.47	0.68	0.21
10744	-1.13	0.11	-0.50	-0.61
10750	-1.22	-0.28	-0.59	-0.31
10751	-0.97	-0.46	-0.34	0.12
10752	-0.19	0.41	0.44	0.03
10753	0.23	1.15	0.86	-0.29
10754	0.80	1.36	1.43	0.07
10763	-0.84	-0.27	-0.20	0.07
10766	-2.10	-1.15	-1.47	-0.32
10771	-1.61	-0.77	-0.98	-0.21
10772	-3.24	-2.20	-2.60	-0.40
10779	-0.29	0.21	0.34	0.14
10780	1.62	2.22	2.25	0.03
10780	1.87	2.19	2.50	0.31
10781	0.05	0.44	0.68	0.24
10782	0.83	1.67	1.46	-0.21
10783	-1.21	-0.74	-0.58	0.16
10826	-2.13	0.59	-1.49	-2.09
10826	0.17	-1.20	0.81	2.01
10827	-1.50	-0.66	-0.86	-0.21
8493	-1.43	-1.00	-0.80	0.20
8497	-2.02	-1.07	-1.38	-0.31
8499	-2.75	-1.91	-2.11	-0.21
8501	-0.13	0.11	0.50	0.39
8505	-2.41	-1.65	-1.77	-0.12
9452	-0.29	0.49	0.35	-0.15
9456	-3.71	-2.94	-3.08	-0.14
9465	0.00	0.63	0.64	0.00
9467	0.29	0.79	0.92	0.14
9470	1.33	1.88	1.96	0.08
9471	0.69	1.10	1.33	0.23
9474	-0.81	0.01	-0.18	-0.19
9475	1.56	2.10	2.19	0.09
9502	-0.31	-0.11	0.33	0.44
mean	-0.55	0.08		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.63	

Appendix K.12— Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 5 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10788	-0.26	0.37	0.28	-0.09
10789	-3.37	-2.75	-2.84	-0.09
10790	0.25	1.20	0.79	-0.42
10791	0.36	1.18	0.90	-0.28
10795	-1.93	-1.35	-1.40	-0.04
10796	-0.25	0.01	0.28	0.27
10798	-0.72	-0.34	-0.18	0.16
10799	-1.13	-0.43	-0.59	-0.16
10800	0.58	1.29	1.12	-0.18
10803	-0.75	-0.05	-0.22	-0.17
10805	-1.02	-0.26	-0.49	-0.23
10808	0.42	1.00	0.96	-0.04
10811	-1.65	-1.07	-1.11	-0.05
10813	0.88	1.58	1.42	-0.16
10816	-0.51	0.67	0.02	-0.65
10817	2.64	1.98	3.18	1.19
10817	1.33	3.21	1.87	-1.34
10820	1.09	1.16	1.62	0.46
10824	-0.71	-0.31	-0.18	0.13
10829	-2.13	-1.20	-1.59	-0.40
10832	-1.14	-0.66	-0.61	0.05
10833	-1.53	-0.95	-1.00	-0.05
10835	1.59	1.82	2.12	0.30
10836	-1.29	-0.63	-0.75	-0.12
10839	0.10	0.66	0.63	-0.03
10840	0.43	1.11	0.96	-0.15
10848	-1.35	-1.03	-0.81	0.21
10849	0.62	1.10	1.15	0.05
10850	-0.34	0.38	0.20	-0.18
10851	0.39	1.02	0.92	-0.09
10863	0.09	0.38	0.63	0.24
10868	3.10	3.74	3.64	-0.10
10869	-0.32	0.24	0.21	-0.02
10872	-1.30	-0.89	-0.77	0.12
10872	-1.02	-1.00	-0.48	0.52
10875	-1.23	-1.09	-0.69	0.40
11107	-1.00	-0.43	-0.47	-0.04
8521	-1.71	-1.11	-1.18	-0.07
8525	-1.12	-0.93	-0.59	0.35
8527	-1.22	-0.87	-0.69	0.18
8535	-1.51	-1.22	-0.97	0.25
8539	-1.71	-1.34	-1.17	0.17
9476	0.82	1.56	1.36	-0.21
9485	1.23	1.66	1.76	0.10
9486	-0.39	0.35	0.15	-0.20
9487	0.58	0.66	1.11	0.46
9716	-2.84	-2.23	-2.30	-0.07
mean	-0.40	0.13		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.54	

Appendix K.13 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 6 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10048	-0.08	0.58	0.41	-0.18
10049	-0.58	0.22	-0.09	-0.32
10051	-0.85	-0.05	-0.36	-0.31
10052	-0.34	-0.69	0.15	0.84
10053	-1.27	-0.52	-0.78	-0.26
10057	-1.69	-0.23	-1.20	-0.97
10060	0.72	1.14	1.21	0.06
10064	-0.61	-0.12	-0.12	0.00
10067	-4.21	-3.16	-3.73	-0.56
10070	-0.19	1.07	0.30	-0.78
10071	0.08	0.50	0.56	0.06
10076	0.34	0.80	0.83	0.03
10078	0.56	0.85	1.05	0.21
10079	0.33	0.56	0.82	0.26
10083	0.43	0.82	0.92	0.09
10088	0.15	0.47	0.63	0.17
10093	-0.45	-0.08	0.04	0.11
10094	0.55	0.90	1.04	0.14
10095	0.48	0.49	0.97	0.48
10096	0.44	1.12	0.93	-0.19
10103	-1.42	-1.03	-0.93	0.10
10106	-2.36	-1.48	-1.87	-0.39
10107	-2.32	-1.73	-1.83	-0.10
10108	-1.16	-0.70	-0.67	0.04
10111	1.11	1.34	1.60	0.26
10113	-1.93	-1.75	-1.44	0.31
10115	-0.71	-0.83	-0.22	0.61
10117	0.40	0.87	0.89	0.03
10120	-0.02	0.41	0.47	0.06
10129	-2.79	-2.24	-2.30	-0.06
10137	-2.85	-2.11	-2.36	-0.25
10139	1.75	2.28	2.24	-0.04
10143	0.38	0.90	0.87	-0.03
10148	-0.90	-0.61	-0.42	0.19
10150	-0.53	-0.06	-0.04	0.01
10151	1.49	1.57	1.98	0.42
8549	-3.02	-2.29	-2.53	-0.24
8555	-0.67	-0.28	-0.18	0.11
8567	-1.14	-0.32	-0.66	-0.34
9491	-2.19	-1.65	-1.70	-0.05
9492	-1.91	-1.29	-1.42	-0.13
9496	-0.26	0.36	0.23	-0.14
9498	2.24	2.54	2.73	0.19
9512	0.89	1.25	1.38	0.12
9513	0.81	1.17	1.30	0.12
9718	1.85	1.98	2.34	0.36
9719	-1.78	-1.23	-1.29	-0.06
mean	-0.49	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.49	

Appendix K.14— Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 7 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10288	-1.96	-0.77	-0.97	-0.20
10290	-1.67	-0.49	-0.69	-0.20
10294	-0.87	0.36	0.11	-0.24
10298	-0.86	0.47	0.13	-0.34
10299	-1.48	-0.69	-0.49	0.20
10301	-1.53	-0.34	-0.54	-0.20
10302	0.18	1.28	1.16	-0.11
10303	-2.11	-0.78	-1.12	-0.34
10309	-0.01	0.71	0.98	0.27
10313	-1.81	-0.82	-0.83	-0.01
10317	0.83	2.01	1.81	-0.20
10318	-1.84	-0.63	-0.85	-0.22
10322	-1.23	-0.29	-0.24	0.05
10331	-0.60	0.15	0.38	0.24
10340	1.00	2.10	1.98	-0.12
10341	1.19	1.62	2.17	0.55
10344	0.07	1.18	1.06	-0.12
10347	0.83	1.92	1.82	-0.11
10349	-1.05	-0.12	-0.06	0.06
10351	0.35	1.25	1.34	0.09
10352	0.09	1.62	1.08	-0.54
10362	-1.52	-0.57	-0.54	0.03
10366	-0.61	0.41	0.38	-0.03
10366	-0.66	0.18	0.33	0.15
10369	-0.73	0.08	0.26	0.17
10371	-2.15	-1.55	-1.16	0.39
10374	-0.46	0.54	0.53	-0.01
10375	0.73	1.68	1.72	0.04
10378	-3.46	-2.14	-2.48	-0.33
10379	-0.52	0.17	0.47	0.30
10701	0.04	1.07	1.03	-0.04
8593	-1.19	-0.25	-0.20	0.05
8597	-0.91	0.06	0.08	0.02
8603	-1.43	-0.88	-0.44	0.44
8609	-0.62	0.32	0.36	0.05
8613	-2.04	-1.08	-1.05	0.03
8698	-3.04	-1.85	-2.05	-0.20
9508	0.89	1.81	1.88	0.07
9514	-0.09	0.81	0.90	0.09
9516	-0.70	0.01	0.29	0.28
9520	-0.72	0.27	0.27	0.00
mean	-0.77	0.21		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.99	

Appendix K.15 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Grade 8 Mathematics

Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10483	-1.15	-0.97	-0.67	0.30
10487	-0.60	-0.05	-0.13	-0.08
10488	-2.67	-2.00	-2.19	-0.19
10494	-0.56	0.10	-0.08	-0.19
10496	-0.82	-0.55	-0.34	0.21
10498	-1.39	-0.83	-0.92	-0.09
10507	-2.30	-1.61	-1.83	-0.22
10510	0.23	0.71	0.71	0.00
10513	-0.69	-0.72	-0.22	0.50
10518	0.97	1.56	1.45	-0.11
10520	-1.43	-1.02	-0.95	0.06
10523	2.21	2.94	2.69	-0.25
10525	-2.12	-1.58	-1.64	-0.06
10527	-0.36	-0.19	0.12	0.31
10528	-2.93	-2.49	-2.45	0.03
10530	-1.31	-0.77	-0.84	-0.06
10532	-0.31	0.17	0.17	-0.01
10534	-0.43	-0.28	0.05	0.32
10538	0.94	1.74	1.41	-0.33
10541	0.28	0.88	0.76	-0.12
10542	0.69	1.22	1.16	-0.06
10548	-0.12	-0.08	0.35	0.43
10554	-1.50	-0.78	-1.03	-0.24
10557	-3.64	-2.55	-3.17	-0.61
10561	-0.70	-0.24	-0.22	0.02
10562	-1.28	-1.01	-0.80	0.21
10564	-0.07	0.40	0.40	0.00
10567	-1.38	-0.72	-0.90	-0.18
10570	0.63	1.38	1.10	-0.28
10574	-0.74	-0.33	-0.27	0.07
10579	2.55	3.28	3.02	-0.26
10580	0.49	0.96	0.97	0.01
10581	-2.40	-1.86	-1.93	-0.07
10585	1.13	1.83	1.60	-0.22
10587	-0.04	0.42	0.44	0.01
10588	-1.54	-1.19	-1.06	0.12
8623	-2.11	-1.81	-1.64	0.17
8631	-1.70	-1.31	-1.23	0.08
8635	-1.31	-1.18	-0.83	0.35
8651	-0.60	0.03	-0.12	-0.16
9518	1.70	2.22	2.18	-0.04
9519	-0.52	-0.37	-0.04	0.33
9521	0.64	1.05	1.11	0.06
9522	2.03	2.62	2.51	-0.11
9525	-0.22	0.09	0.26	0.17
9525	-0.15	0.29	0.32	0.03
9527	-0.52	-0.23	-0.04	0.19
9532	0.26	0.78	0.73	-0.05
mean	-0.52	-0.04		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.48	

Appendix K.16 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Algebra I

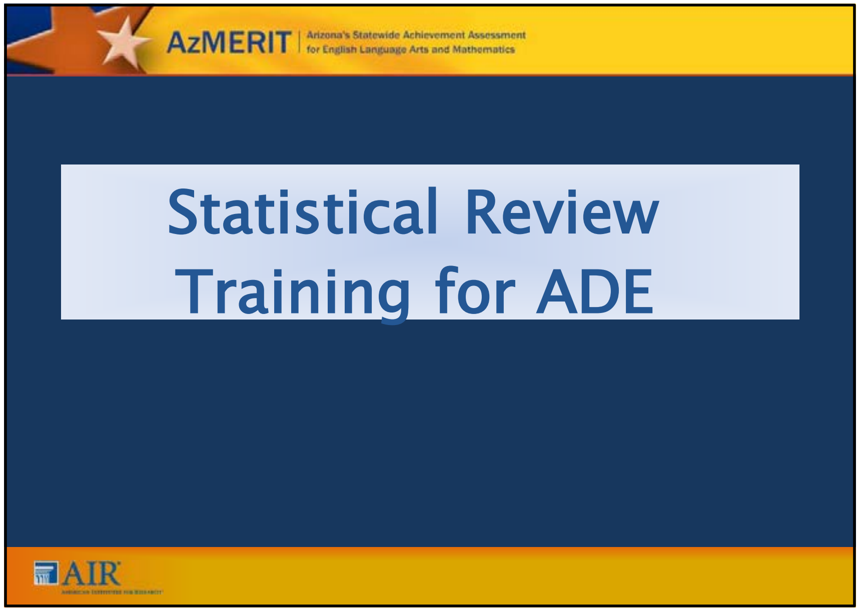
Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10877	-1.76	-0.44	-0.40	0.04
10880	-0.04	1.22	1.32	0.09
10882	-0.55	0.86	0.81	-0.05
10887	-2.43	-1.05	-1.07	-0.02
10888	0.54	2.08	1.90	-0.18
10889	-1.53	-0.08	-0.17	-0.09
10895	-1.41	-0.49	-0.05	0.43
10896	-0.75	0.59	0.61	0.02
10897	-1.04	0.23	0.32	0.09
10905	-2.34	-1.19	-0.98	0.21
10906	-1.17	0.26	0.19	-0.07
10907	-2.48	-0.84	-1.12	-0.28
10934	-2.31	-1.27	-0.95	0.32
10935	-1.48	-0.12	-0.12	0.00
10941	0.16	1.51	1.52	0.01
10942	-1.26	0.11	0.10	-0.01
10943	-0.89	0.33	0.47	0.13
10945	-2.01	-0.70	-0.65	0.05
10951	-0.65	0.68	0.71	0.03
10953	-2.12	-0.94	-0.76	0.18
10960	-0.75	0.72	0.61	-0.11
10963	-2.24	-0.75	-0.88	-0.13
10966	-1.21	0.30	0.15	-0.15
10972	-1.46	-0.05	-0.10	-0.05
10973	-2.72	-1.42	-1.36	0.06
10974	-1.95	-0.27	-0.59	-0.32
10977	-1.78	-0.77	-0.42	0.35
10978	0.47	1.48	1.83	0.35
10988	-0.95	0.79	0.41	-0.37
10990	-0.70	0.81	0.66	-0.16
10993	-2.44	-1.24	-1.08	0.15
11004	-1.10	0.20	0.26	0.06
11011	-1.04	0.46	0.32	-0.14
11013	-1.26	0.21	0.10	-0.11
11044	-0.23	1.33	1.13	-0.20
11045	-1.45	-0.12	-0.09	0.03
11052	-2.62	-1.12	-1.26	-0.14
11058	-1.09	0.36	0.27	-0.08
9530	-0.67	0.43	0.69	0.26
9531	-0.39	1.14	0.97	-0.16
9533	-0.59	0.84	0.77	-0.07
9535	-1.40	-0.17	-0.04	0.13
9536	0.00	1.01	1.36	0.36
9543	-2.07	-0.67	-0.71	-0.04
9543	-1.75	-0.77	-0.39	0.37
9546	-0.56	0.82	0.80	-0.02
9705	-2.33	-0.48	-0.97	-0.49
9707	-3.53	-1.90	-2.17	-0.27
mean	-1.32	0.04		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		1.36	

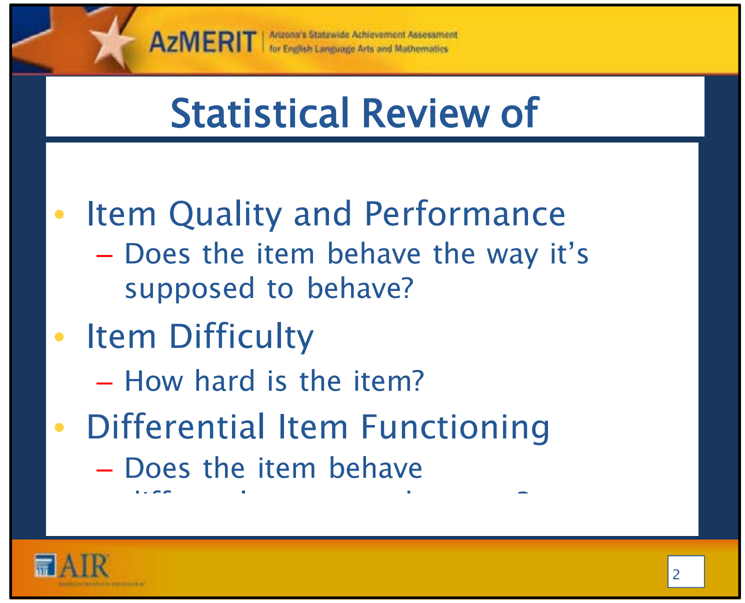
Appendix K.17 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Geometry

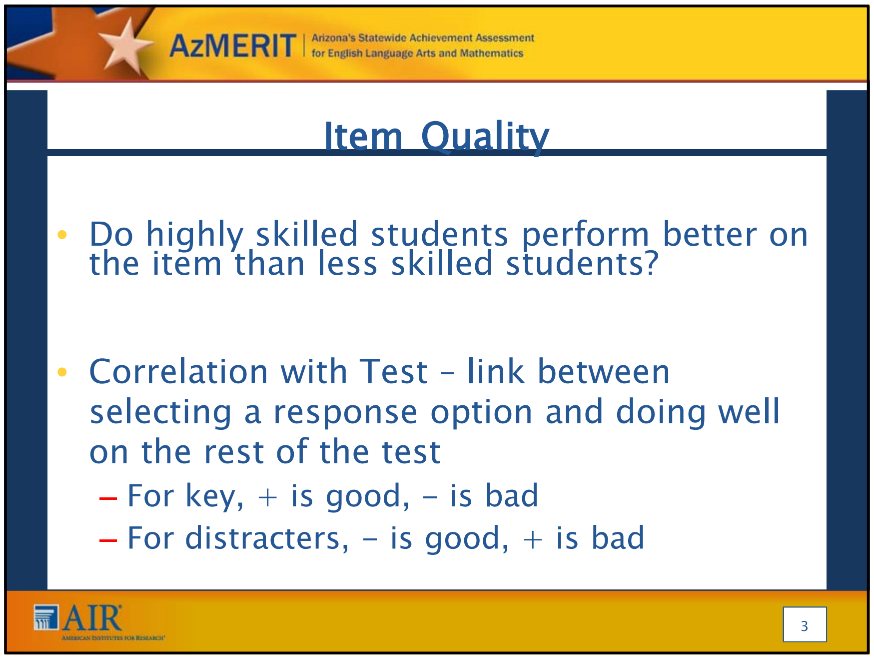
Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10910	-1.51	-1.16	-0.78	0.37
10912	-0.02	0.48	0.70	0.22
10919	-1.92	-1.12	-1.20	-0.08
10921	-2.55	-1.64	-1.82	-0.19
10924	-0.93	-0.19	-0.20	-0.01
10926	-0.25	0.71	0.47	-0.24
10930	0.01	0.46	0.73	0.27
10986	-1.88	-1.30	-1.15	0.14
10998	2.96	3.51	3.69	0.18
11007	-1.48	-0.32	-0.75	-0.43
11008	-0.19	0.26	0.54	0.28
11015	-2.38	-1.44	-1.65	-0.21
11016	-1.54	-0.67	-0.81	-0.14
11017	0.00	0.73	0.73	0.00
11018	-1.34	-0.87	-0.61	0.25
11026	-1.49	-0.62	-0.76	-0.15
11029	-0.86	-0.14	-0.13	0.01
11032	0.14	1.04	0.87	-0.17
11033	-1.79	-1.02	-1.07	-0.05
11034	-0.77	-0.08	-0.04	0.04
11035	-1.31	-0.67	-0.59	0.09
11036	-1.04	-0.42	-0.32	0.11
11037	-3.90	-3.37	-3.17	0.20
11040	-1.46	-0.44	-0.73	-0.29
11060	-0.25	0.63	0.47	-0.16
11061	-1.96	-1.39	-1.23	0.16
11065	0.29	0.94	1.01	0.07
11068	-1.37	-0.79	-0.65	0.14
11072	-1.30	-0.58	-0.57	0.01
11074	-0.15	0.64	0.58	-0.06
11078	-1.29	-0.26	-0.57	-0.30
11081	0.97	1.49	1.70	0.21
11085	-0.16	0.62	0.56	-0.06
11086	-1.79	-1.12	-1.07	0.05
11089	-1.50	-0.42	-0.77	-0.35
11092	0.21	0.87	0.93	0.06
11114	-0.85	-0.30	-0.12	0.18
9547	-1.99	-0.82	-1.26	-0.44
9551	1.34	2.21	2.06	-0.15
9554	0.22	0.75	0.94	0.19
9556	-0.95	-0.19	-0.22	-0.03
9560	-0.40	-0.27	0.32	0.59
9564	-0.29	0.48	0.44	-0.04
9575	2.11	2.96	2.84	-0.12
9581	-0.52	0.09	0.20	0.11
9592	0.78	1.56	1.51	-0.06
9722	0.29	1.20	1.01	-0.19
mean	-0.72	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.73	

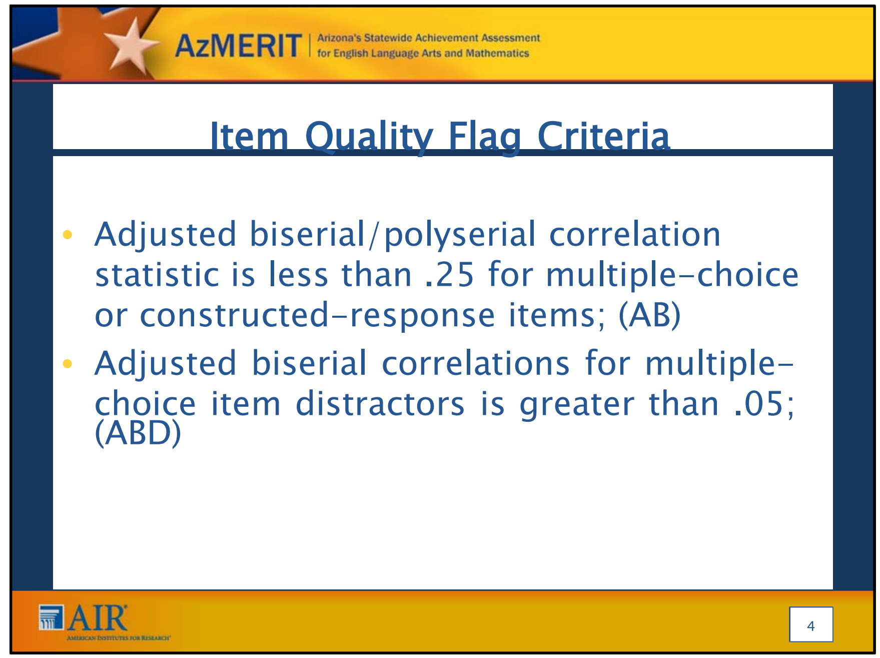
Appendix K.18 — Field Test Analysis: Linking Operational Parameters to the Bank Scale —Algebra II

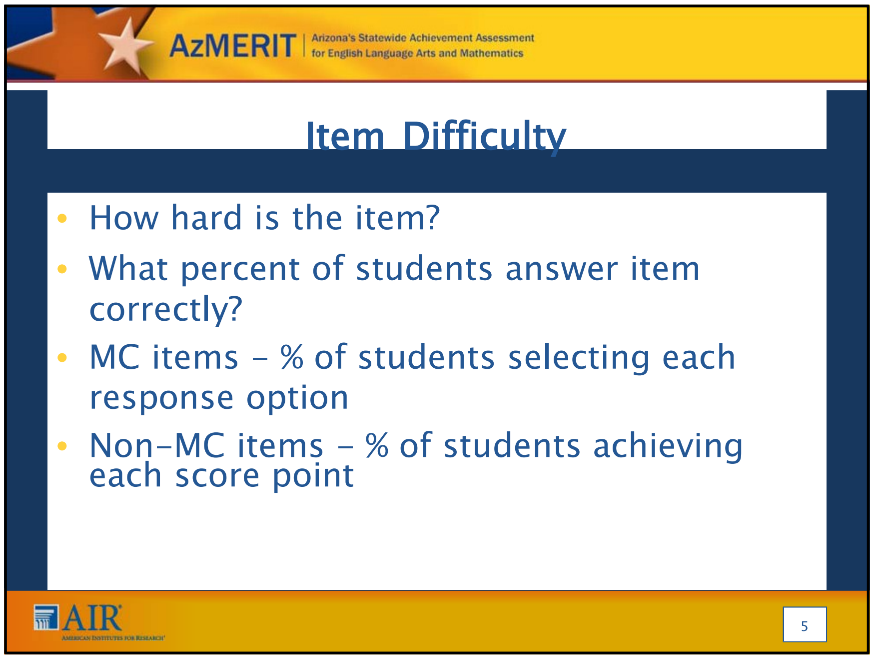
Item ID	Item Parameters			4. Parameter Drift
	1. Spring 16 Operational	2. AIR bank	3. Post-equated	
10160	-1.45	-0.96	-0.66	0.30
10164	-0.11	0.70	0.69	-0.01
10168	0.89	1.54	1.69	0.15
10175	-2.09	-1.11	-1.29	-0.17
10176	0.02	1.21	0.82	-0.39
10177	-1.37	-0.67	-0.57	0.10
10180	1.33	2.15	2.12	-0.03
10182	0.58	1.38	1.38	-0.01
10187	-0.98	-0.52	-0.18	0.34
10192	-2.01	-1.25	-1.21	0.04
10199	0.15	1.20	0.95	-0.25
10200	-2.26	-1.20	-1.46	-0.26
10203	-1.82	-0.85	-1.02	-0.18
10204	-2.17	-1.42	-1.37	0.05
10206	-3.26	-2.25	-2.46	-0.21
10209	-1.46	-0.75	-0.66	0.09
10210	-2.59	-1.48	-1.80	-0.31
10214	-2.45	-1.96	-1.65	0.30
10215	-3.27	-2.31	-2.47	-0.16
10217	-2.10	-1.49	-1.30	0.19
10220	-0.25	0.55	0.54	-0.01
10223	-0.31	0.79	0.48	-0.30
10228	-0.24	0.50	0.55	0.05
10230	-1.16	-0.40	-0.36	0.04
10233	-1.61	-0.88	-0.81	0.07
10236	-1.69	-0.67	-0.90	-0.22
10237	-2.51	-1.98	-1.71	0.27
10240	-0.79	0.24	0.01	-0.23
10243	-1.26	-0.81	-0.47	0.35
10245	-0.14	0.75	0.66	-0.09
10249	0.27	1.26	1.07	-0.19
10255	0.38	1.53	1.18	-0.35
10256	0.25	0.88	1.05	0.17
10259	-0.89	-0.21	-0.09	0.12
10261	-0.29	0.33	0.51	0.18
11121	-1.84	-0.70	-1.04	-0.34
9548	1.49	1.89	2.28	0.39
9549	-0.76	-0.32	0.03	0.35
9567	0.19	0.83	0.98	0.15
9568	0.06	0.53	0.85	0.33
9570	-0.90	0.13	-0.10	-0.24
9573	-0.48	0.26	0.32	0.05
9577	0.43	1.57	1.23	-0.34
9578	-0.54	0.45	0.26	-0.19
9580	-0.85	-0.46	-0.06	0.41
9589	1.42	2.33	2.22	-0.11
9591	1.19	1.90	1.99	0.09
mean	-0.79	0.00		
constant	mean(AIR bank)-mean(Spring 2016 Operational)		0.80	

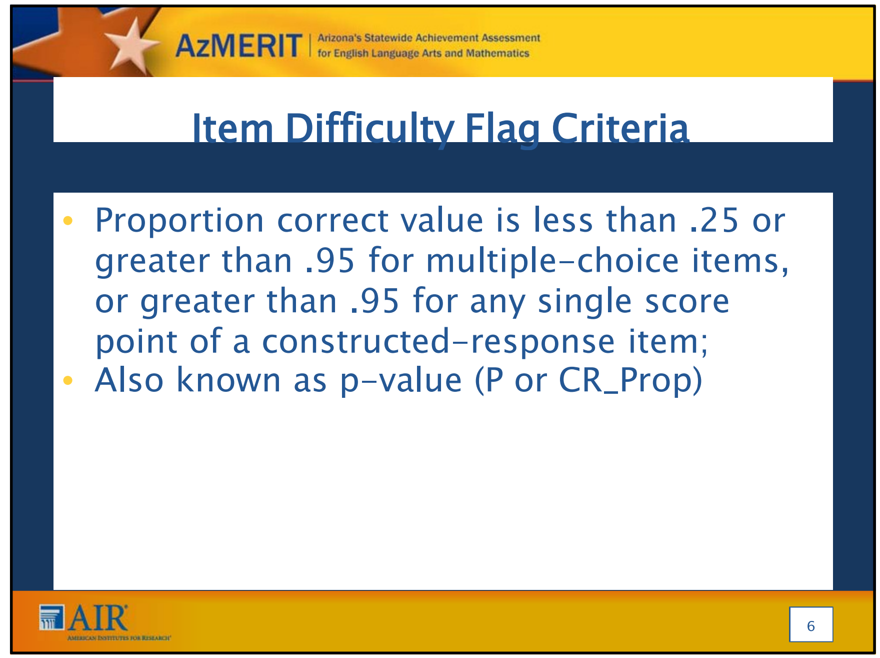

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Statistical Review Training for ADE
AIR

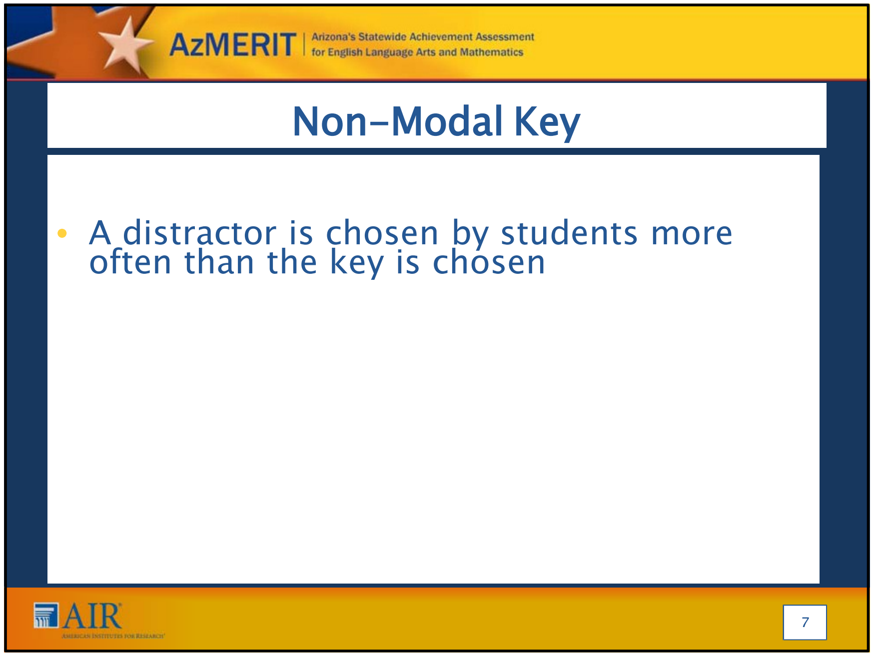

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Statistical Review of
Item Quality and Performance
– Does the item behave the way it's supposed to behave?
Item Difficulty
– How hard is the item?
Differential Item Functioning
– Does the item behave
AIR
2

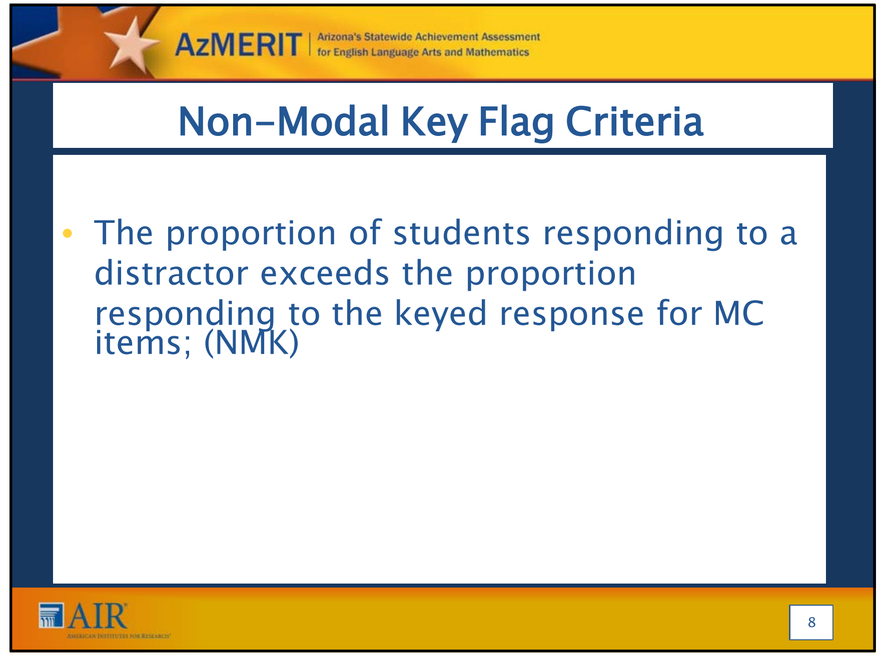

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Quality
Do highly skilled students perform better on the item than less skilled students?
Correlation with Test – link between selecting a response option and doing well on the rest of the test
– For key, + is good, – is bad
– For distracters, – is good, + is bad
AIR
3

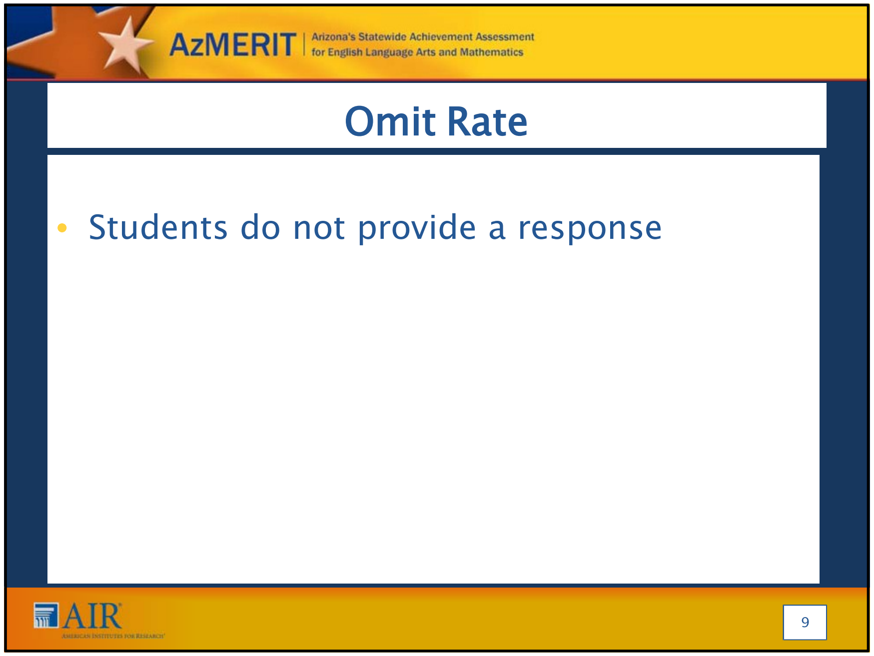

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Quality Flag Criteria
Adjusted biserial/polyserial correlation statistic is less than .25 for multiple-choice or constructed-response items; (AB)
Adjusted biserial correlations for multiple-choice item distractors is greater than .05; (ABD)
AIR
4

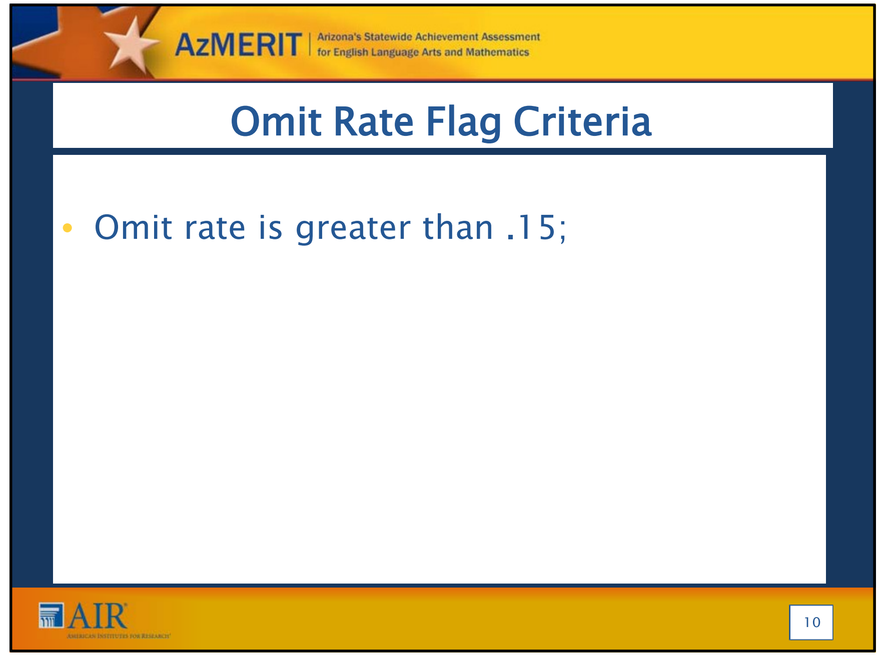

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Difficulty
How hard is the item?
What percent of students answer item correctly?
MC items – % of students selecting each response option
Non-MC items – % of students achieving each score point
AIR
5


AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Item Difficulty Flag Criteria
Proportion correct value is less than .25 or greater than .95 for multiple-choice items, or greater than .95 for any single score point of a constructed-response item;
Also known as p-value (P or CR_Prop)
AIR
6


AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Non–Modal Key
• A distractor is chosen by students more often than the key is chosen
AIR
7


AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Non–Modal Key Flag Criteria
• The proportion of students responding to a distractor exceeds the proportion responding to the keyed response for MC items; (NMK)
AIR
8


AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Omit Rate
• Students do not provide a response
AIR
9


AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Omit Rate Flag Criteria
• Omit rate is greater than .15;
AIR
10

AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

Differential Item Functioning

* Fair Items behave similarly across groups

* Probability of answering correctly is the same for all students of similar ability regardless of group membership

Subgroup Comparisons:
- Female/Male
- Non-Hispanic / Hispanic, Latino or Spanish origin
- Black, African American / White
- American Indian or Alaskan Native / White
- Asian / White
- Native Hawaiian or Other Pacific Islander / White
- Multiple ethnicities selected / White

AIR American Institutes for Research

11

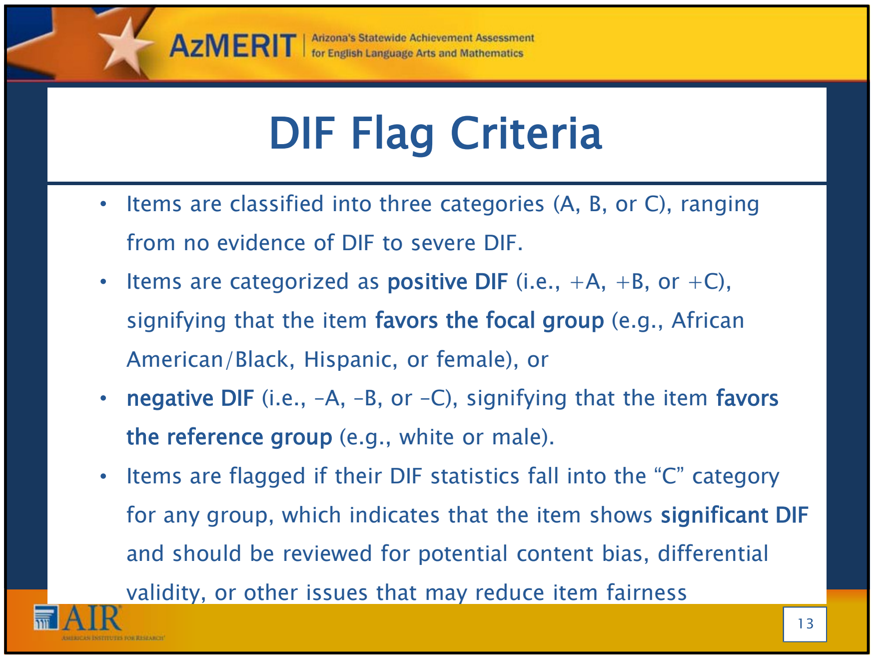
AzMERIT | Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics

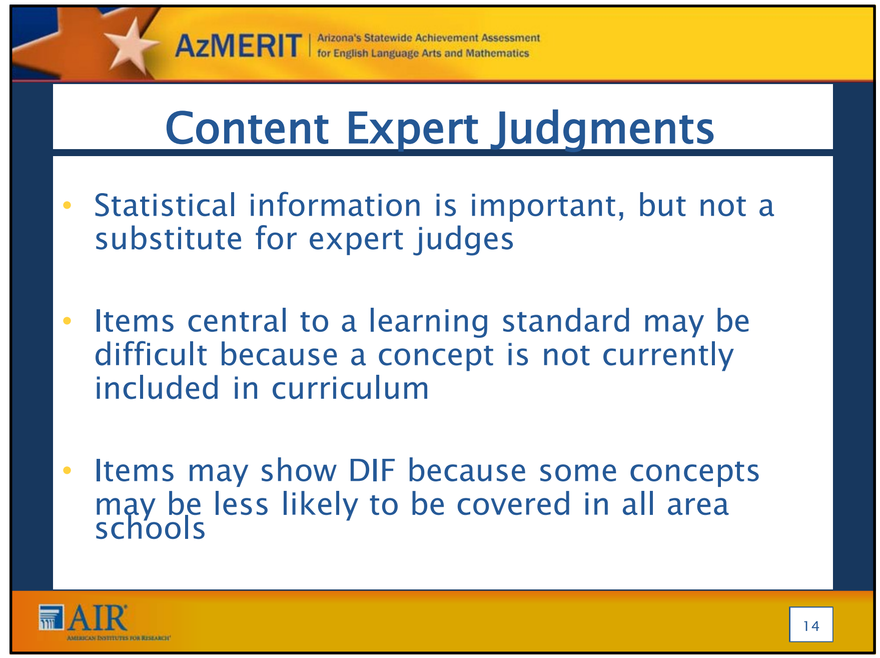
Differential Item Functioning (DIF)

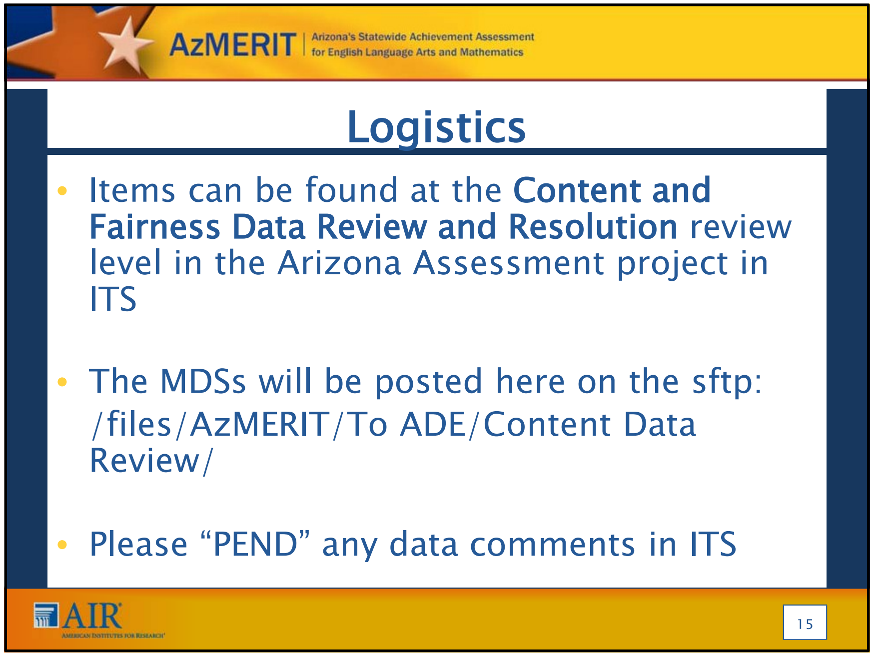
- Direction of possible bias
  - "–" item favors reference groups
  - "+" item favors focal group
- Severity of possible bias
  - "A" No statistical evidence of DIF
  - "B" Evidence for potential mild DIF
  - "C" Evidence for potential severe DIF
- "C" indicates that the item is more difficult for one group and should be reviewed carefully for bias

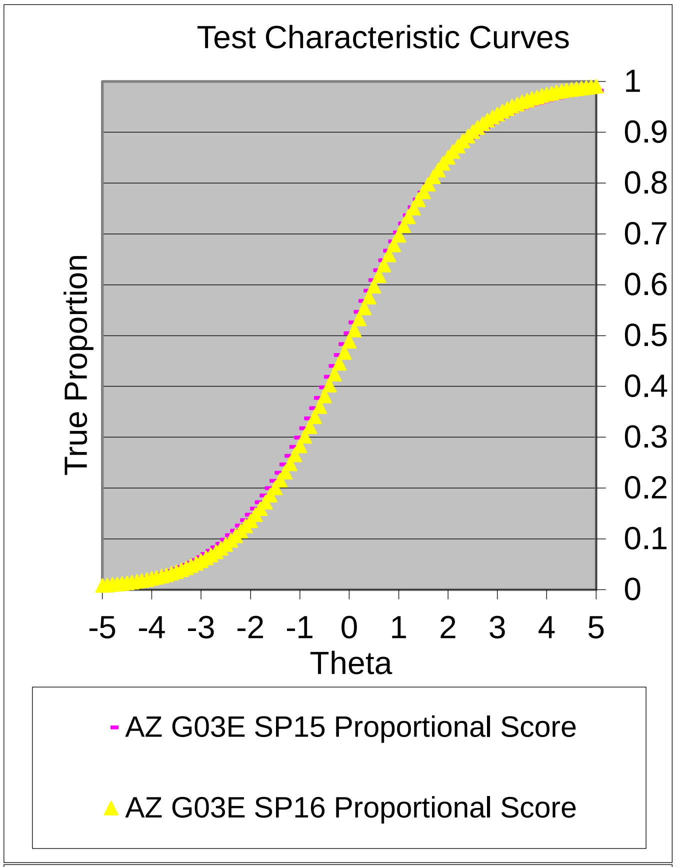
AIR American Institutes for Research

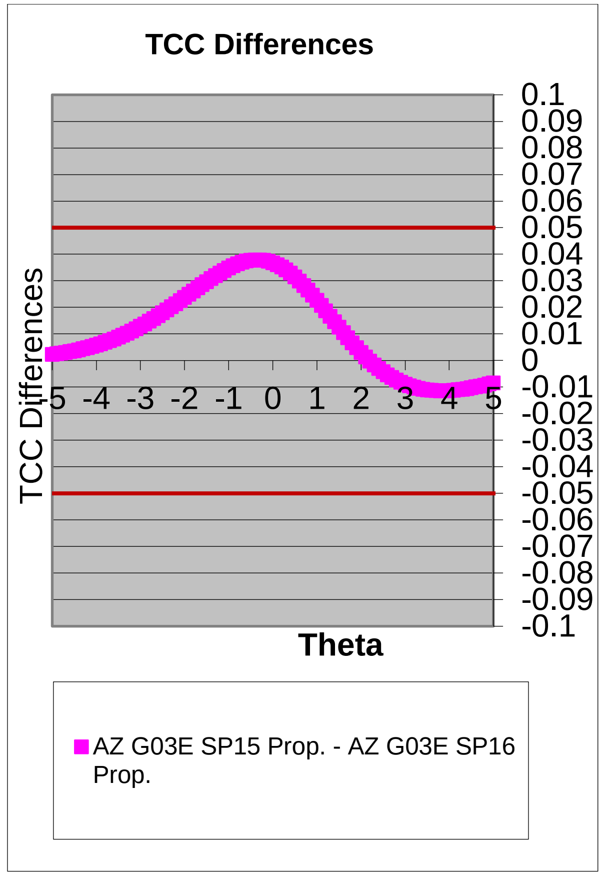
12

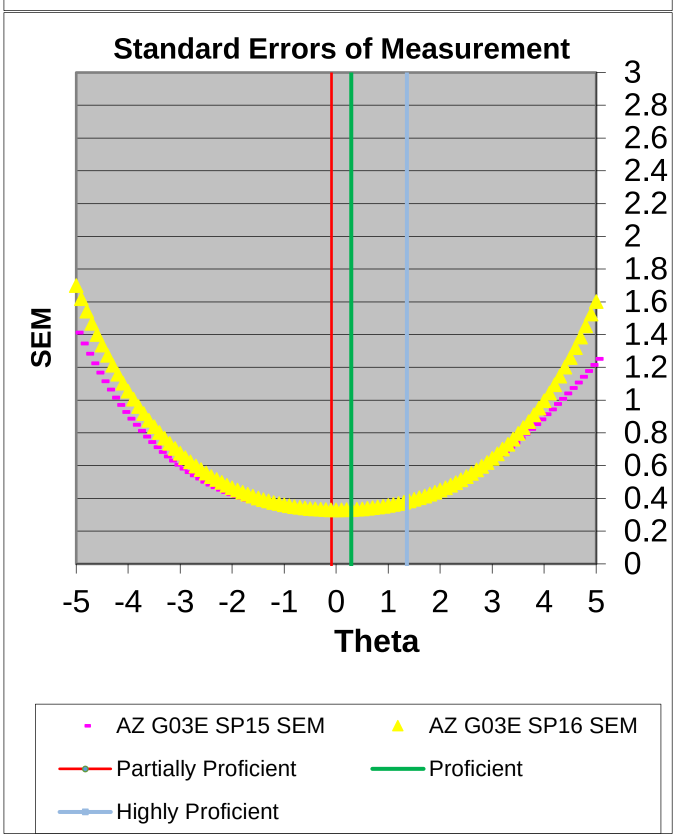

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
DIF Flag Criteria
• Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF.
• Items are categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, or female), or
• negative DIF (i.e., –A, –B, or –C), signifying that the item favors the reference group (e.g., white or male).
• Items are flagged if their DIF statistics fall into the "C" category for any group, which indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness
AIR
13

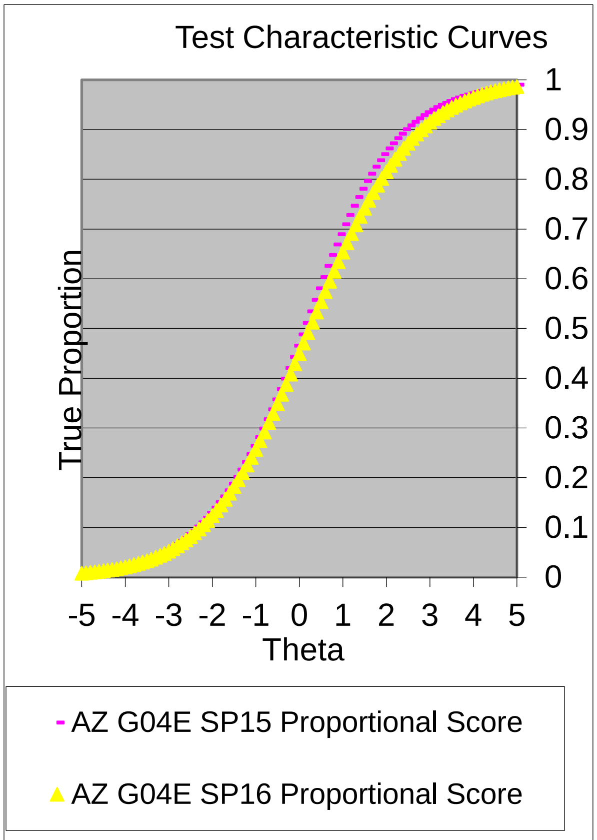

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Content Expert Judgments
• Statistical information is important, but not a substitute for expert judges
• Items central to a learning standard may be difficult because a concept is not currently included in curriculum
• Items may show DIF because some concepts may be less likely to be covered in all area schools
AIR
14

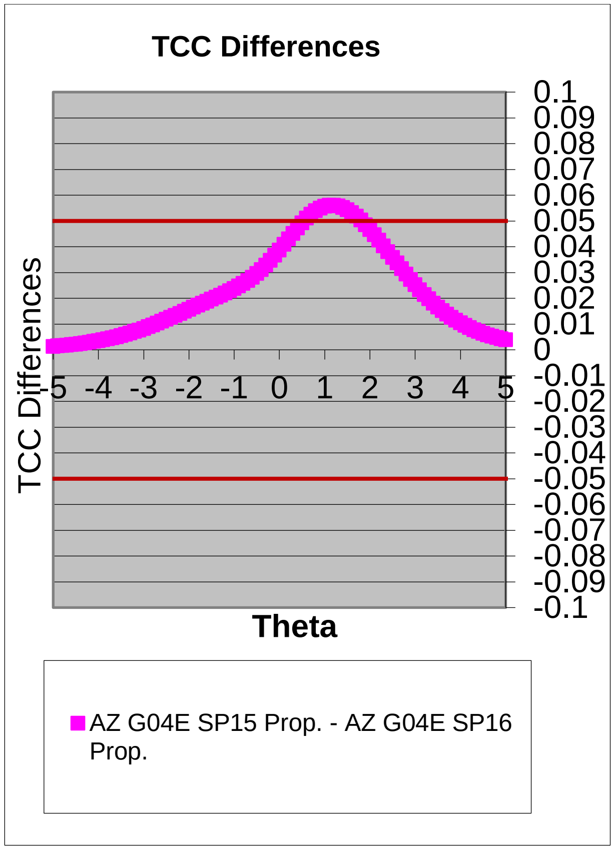

AzMERIT
Arizona's Statewide Achievement Assessment for English Language Arts and Mathematics
Logistics
Items can be found at the Content and Fairness Data Review and Resolution review level in the Arizona Assessment project in ITS
The MDSs will be posted here on the sftp: /files/AzMERIT/To ADE/Content Data Review/
Please "PEND" any data comments in ITS
AIR
15

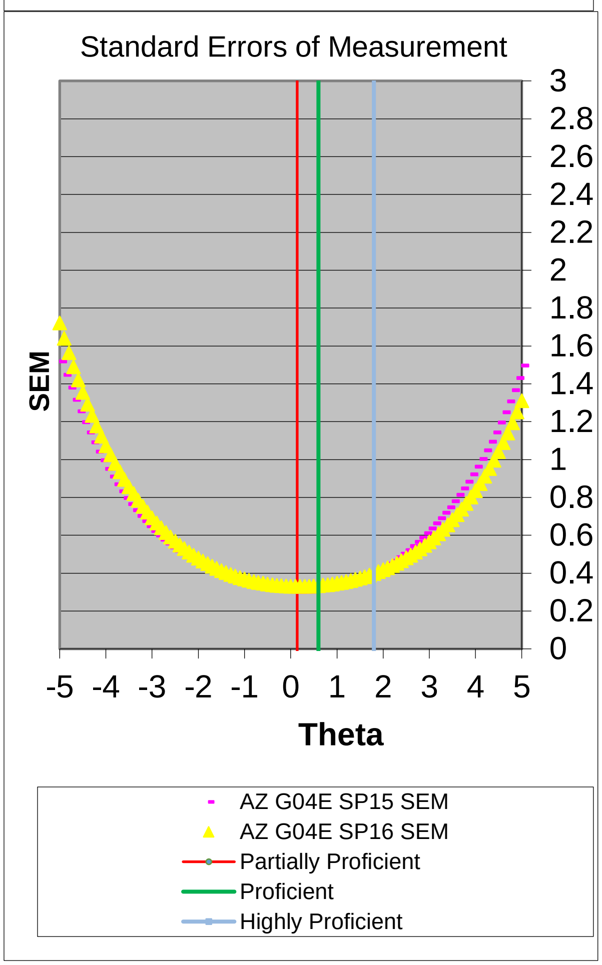

Test Characteristic Curves
True Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G03E SP15 Proportional Score
AZ G03E SP16 Proportional Score


TCC Differences
TCC Differences
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.09
-0.1
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G03E SP15 Prop. - AZ G03E SP16 Prop.

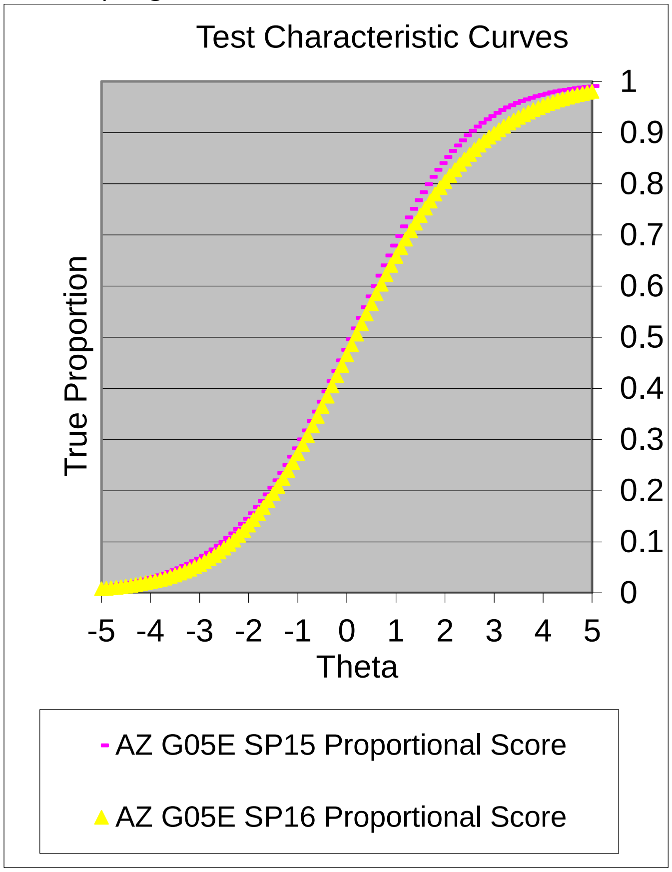

Standard Errors of Measurement
SEM
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G03E SP15 SEM
AZ G03E SP16 SEM
Partially Proficient
Proficient
Highly Proficient

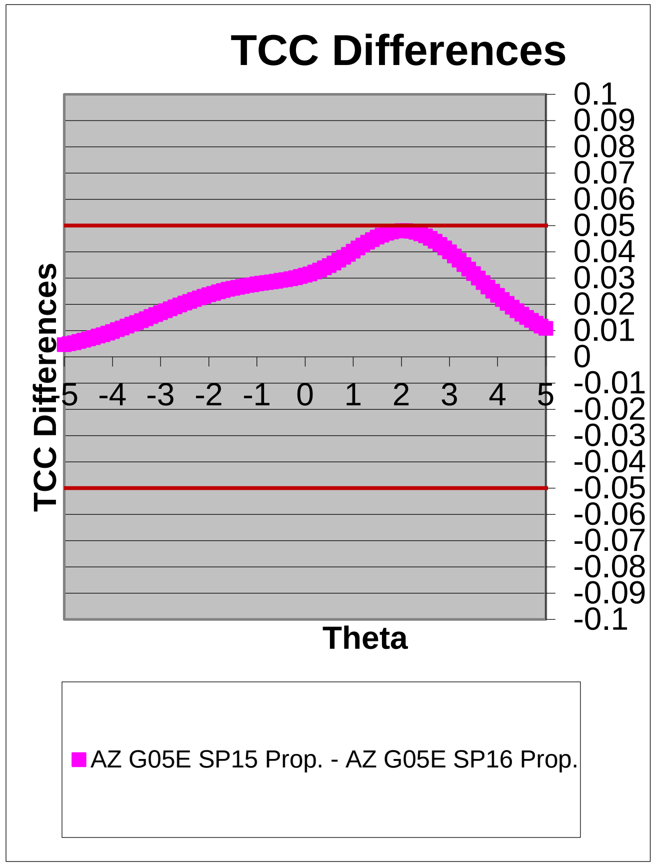

Test Characteristic Curves
True Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G04E SP15 Proportional Score
AZ G04E SP16 Proportional Score

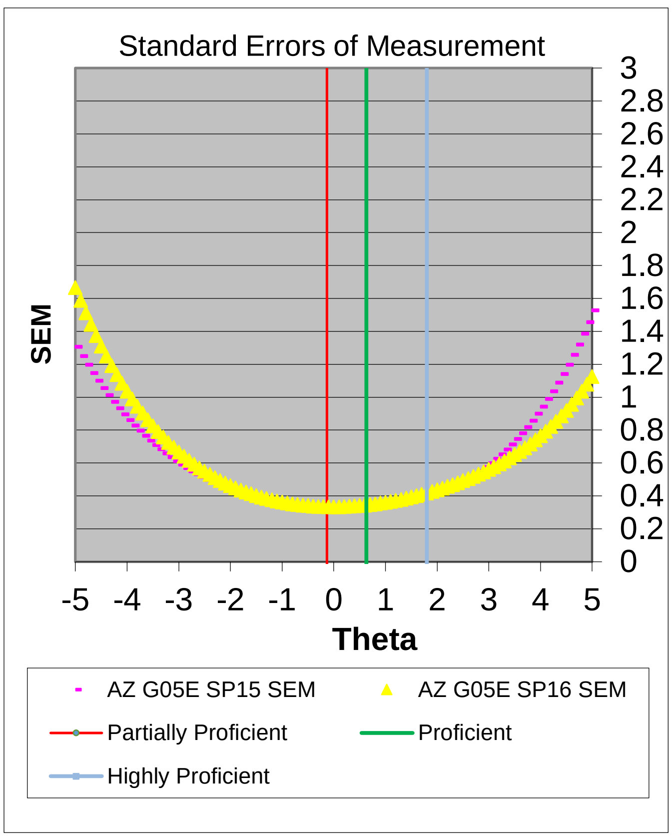

TCC Differences
TCC Differences
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.09
-0.1
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G04E SP15 Prop. - AZ G04E SP16 Prop.


Standard Errors of Measurement
SEM
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G04E SP15 SEM
AZ G04E SP16 SEM
Partially Proficient
Proficient
Highly Proficient

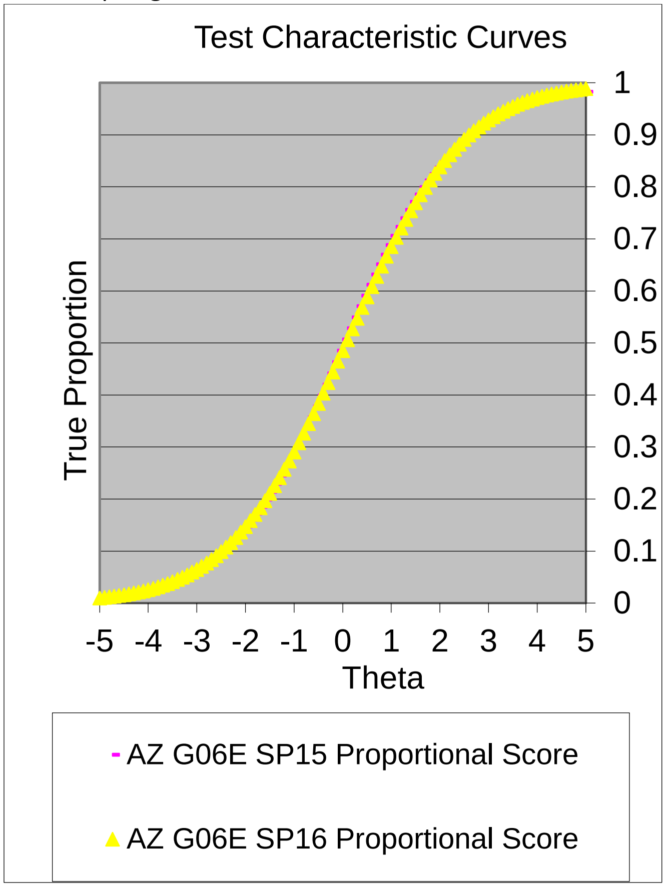
G5ELA Spring 2016 AZ

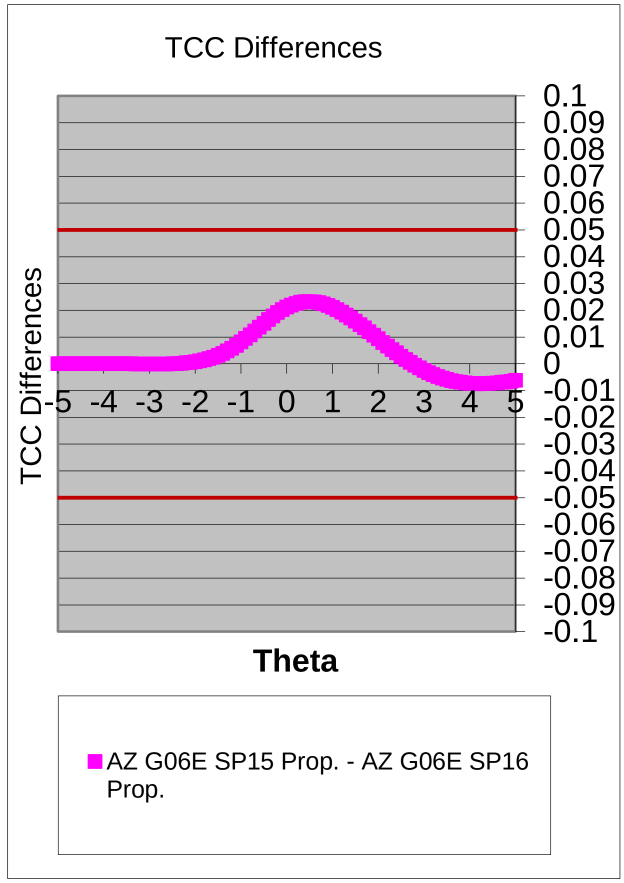


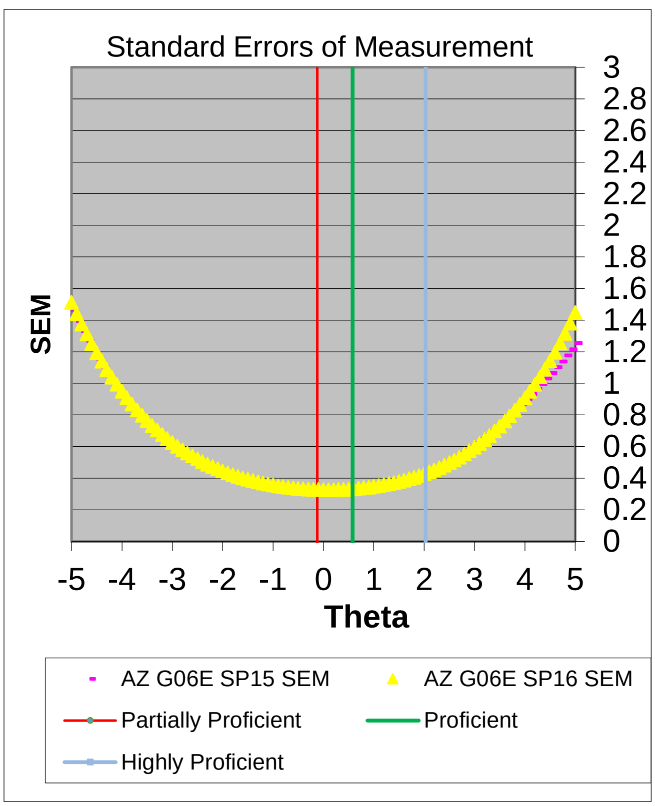




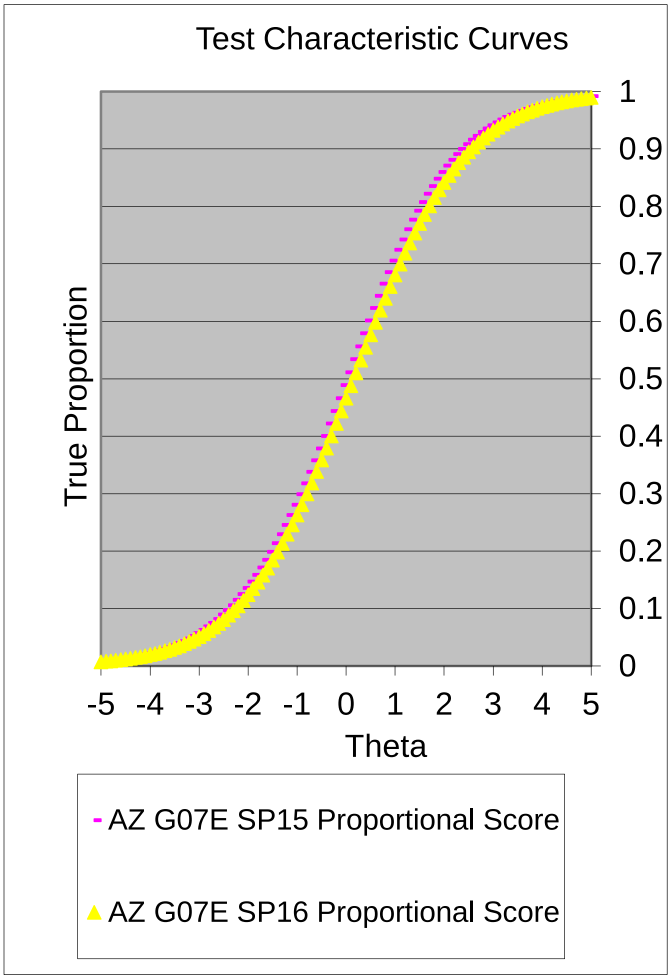
G6ELA Spring 2016 ELA

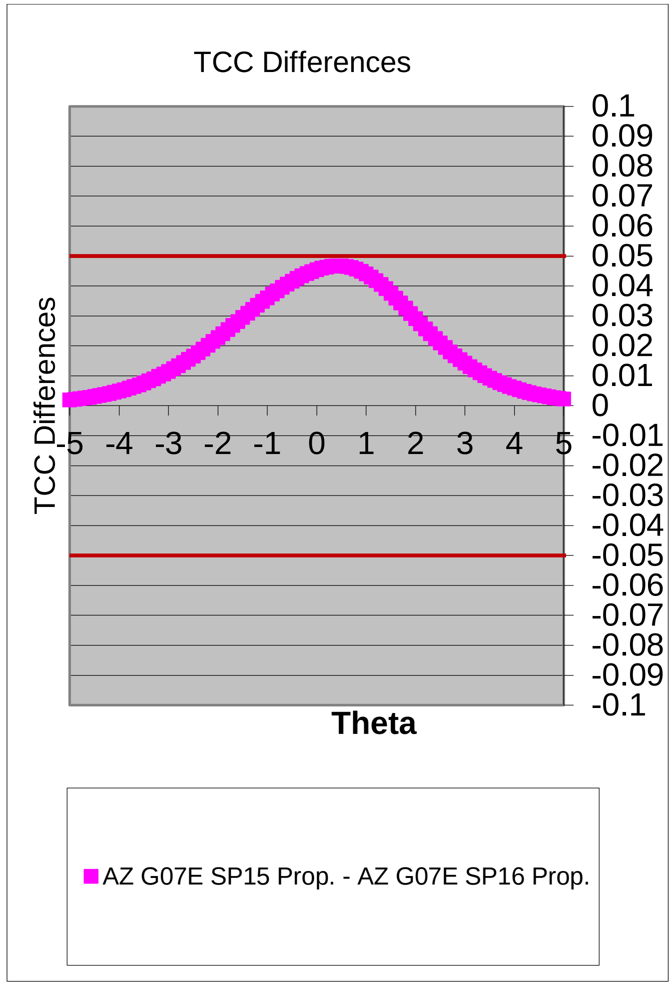


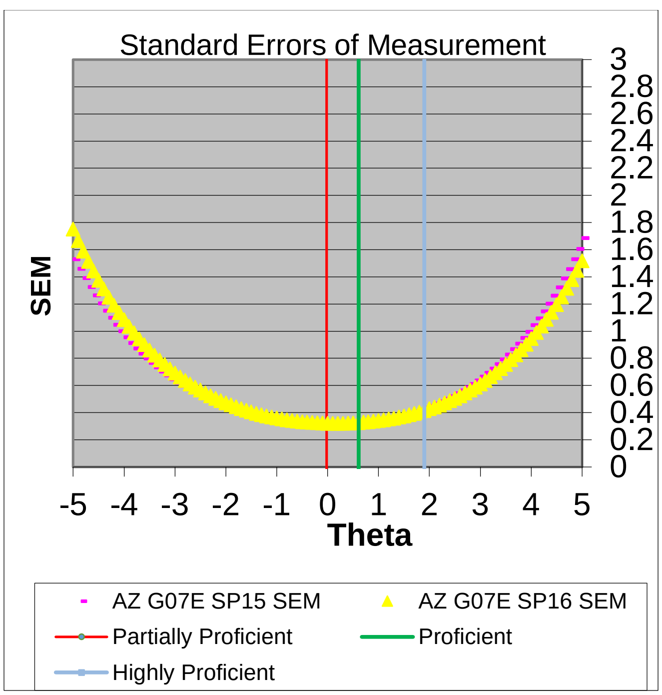




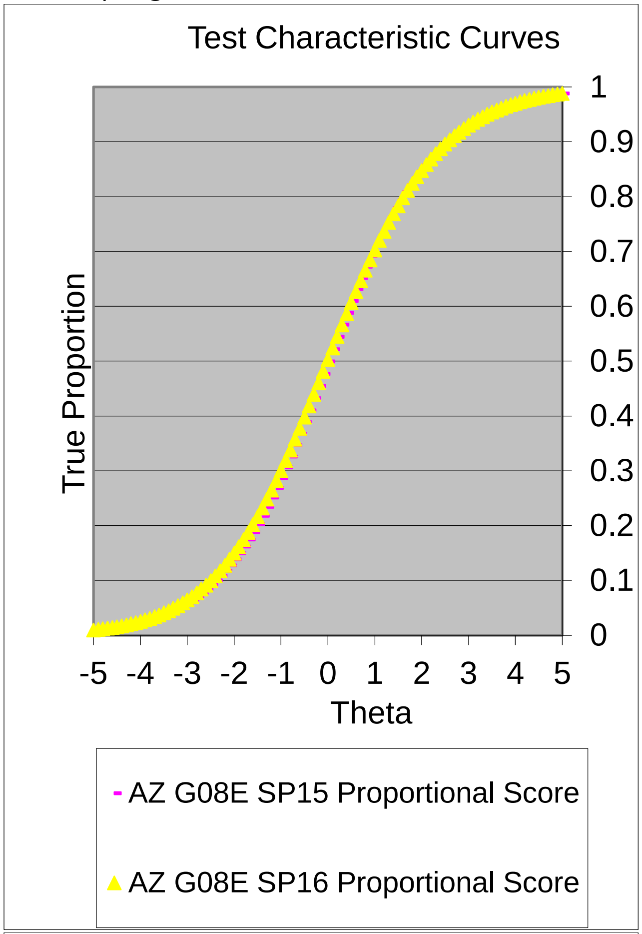
G7ELA Spring 2016 AZ

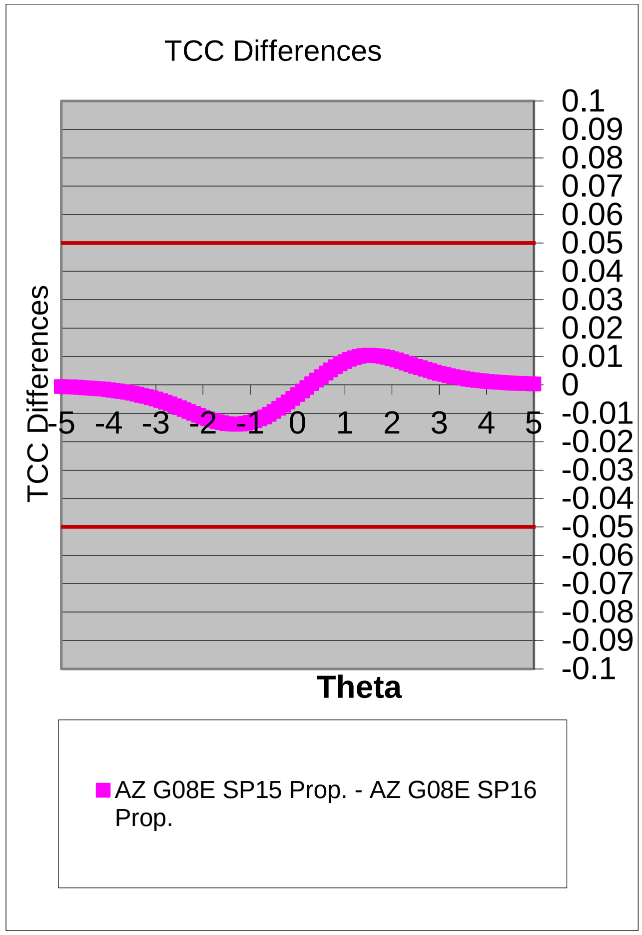


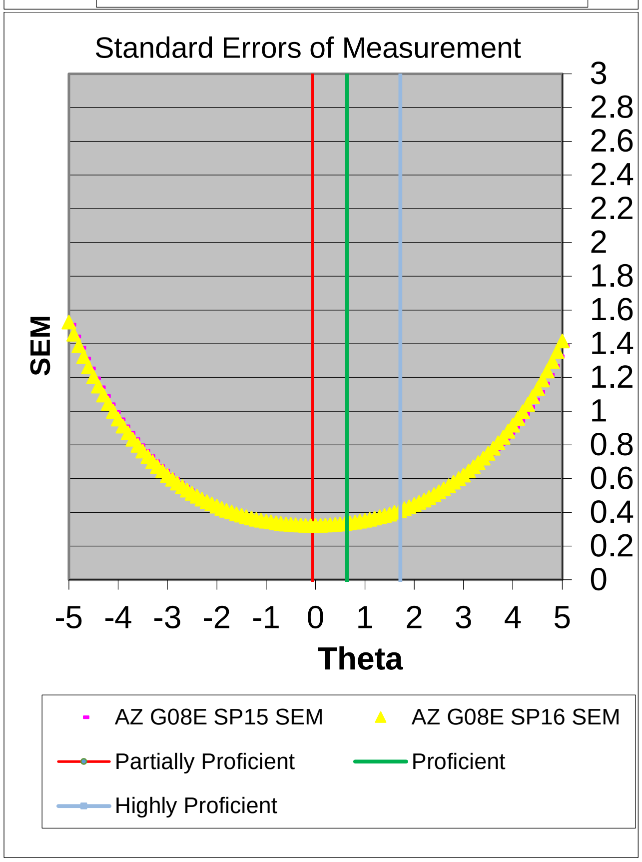




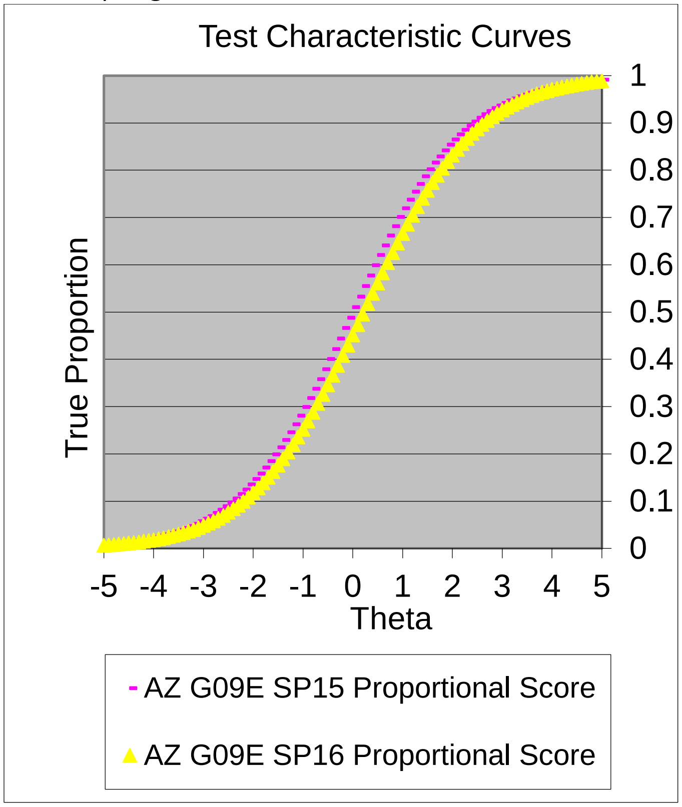
G8ELA Spring 2016 AZ

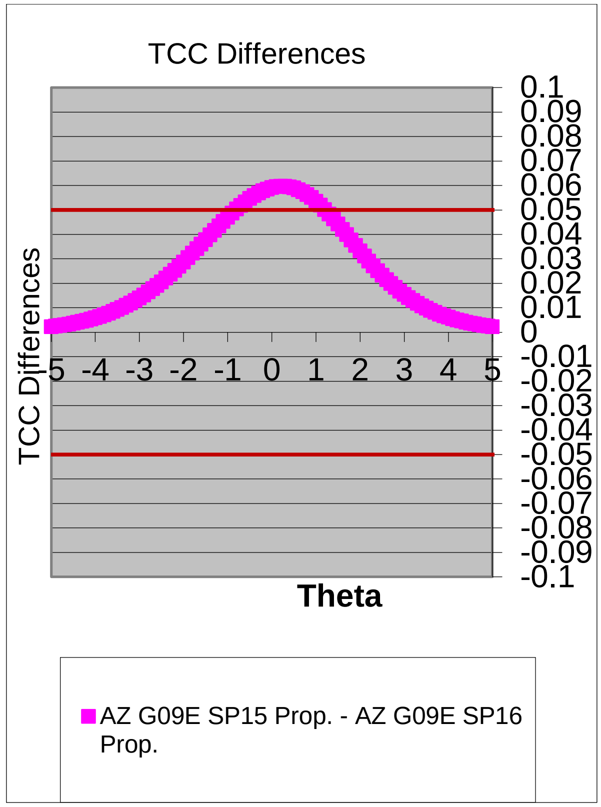


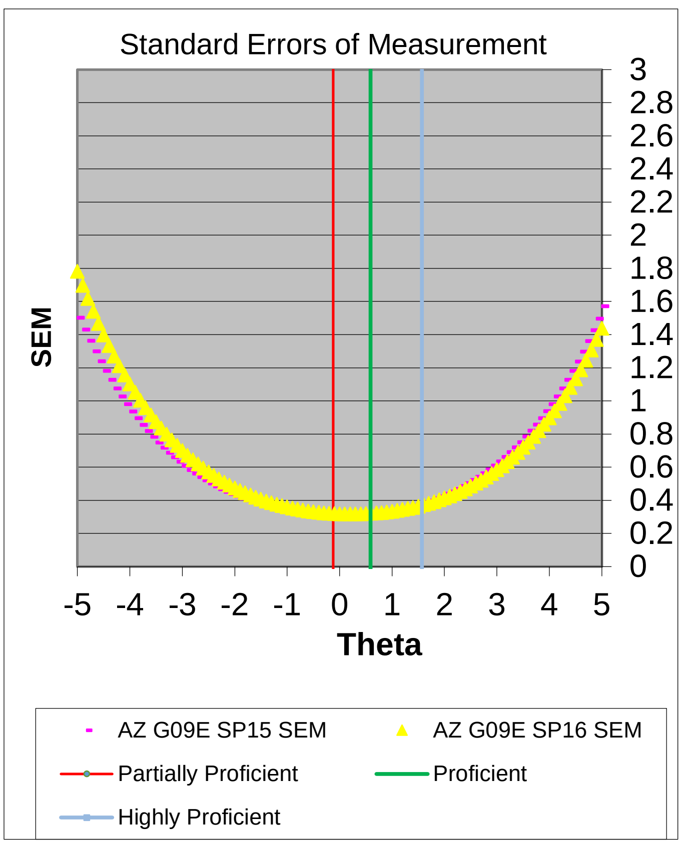




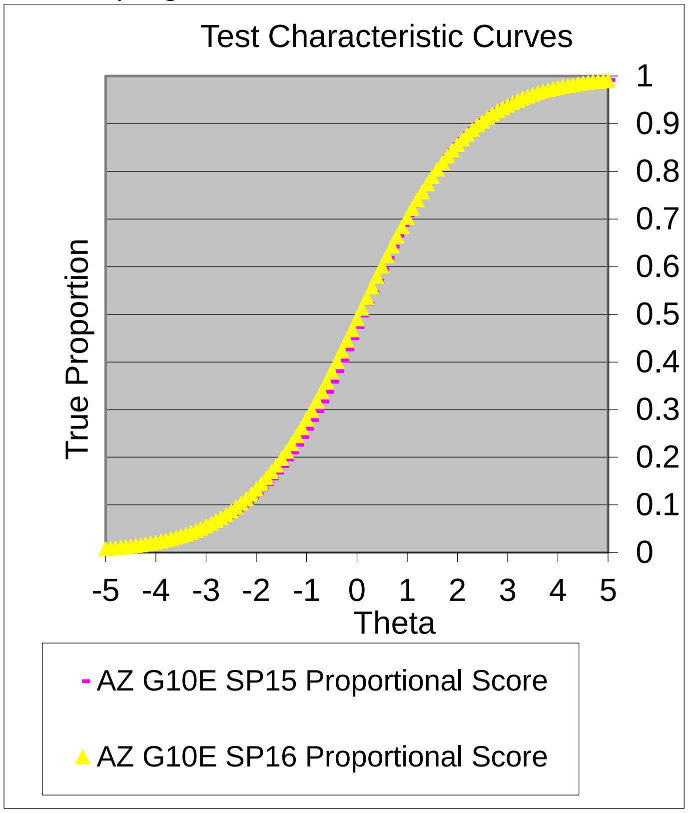
G9ELA Spring 2016 AZ

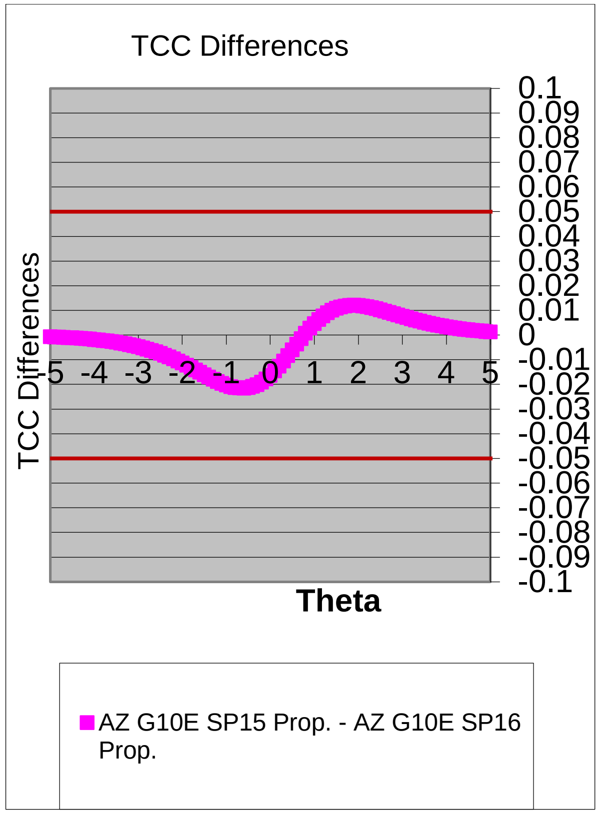


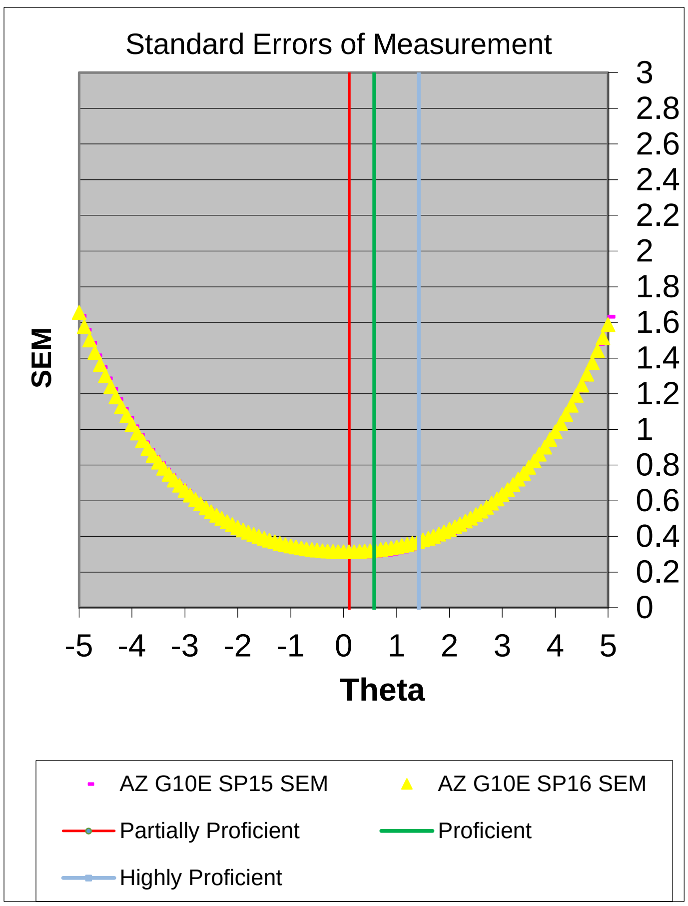




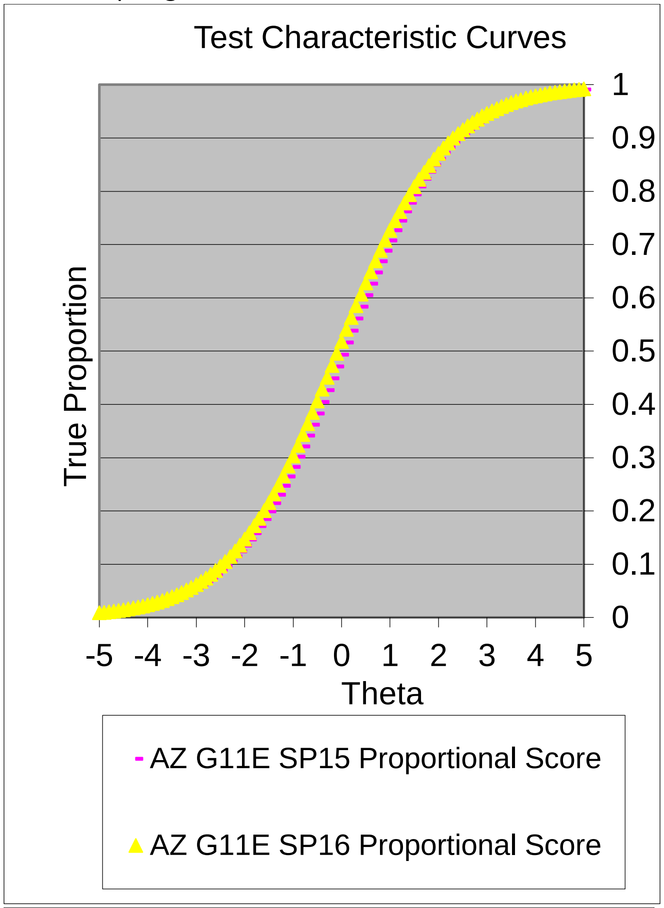
G10ELA Spring 2016 AZ

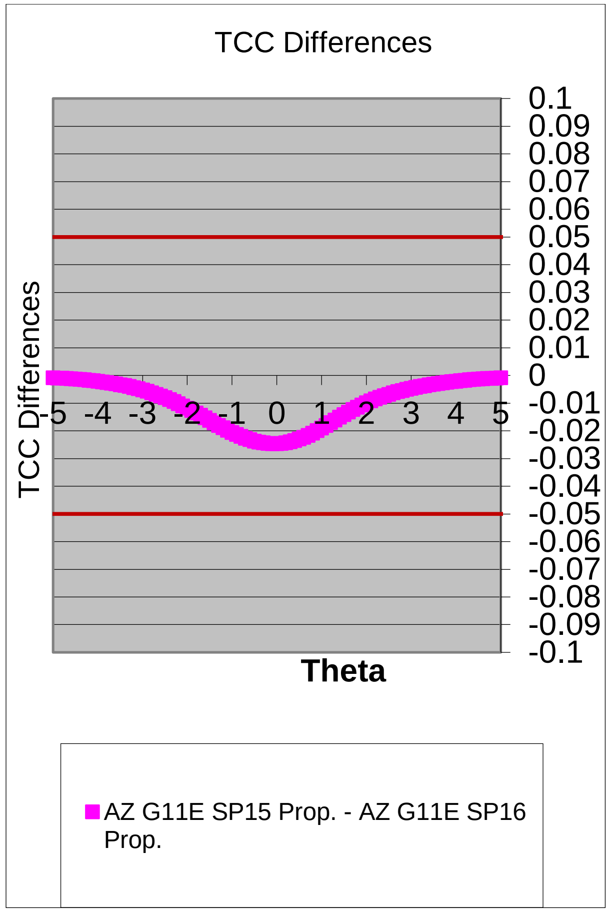


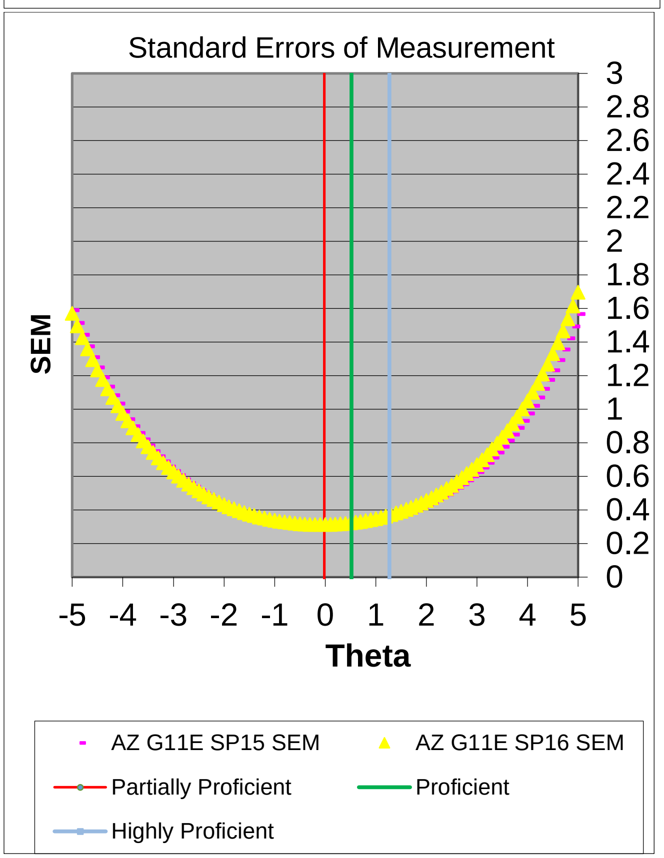




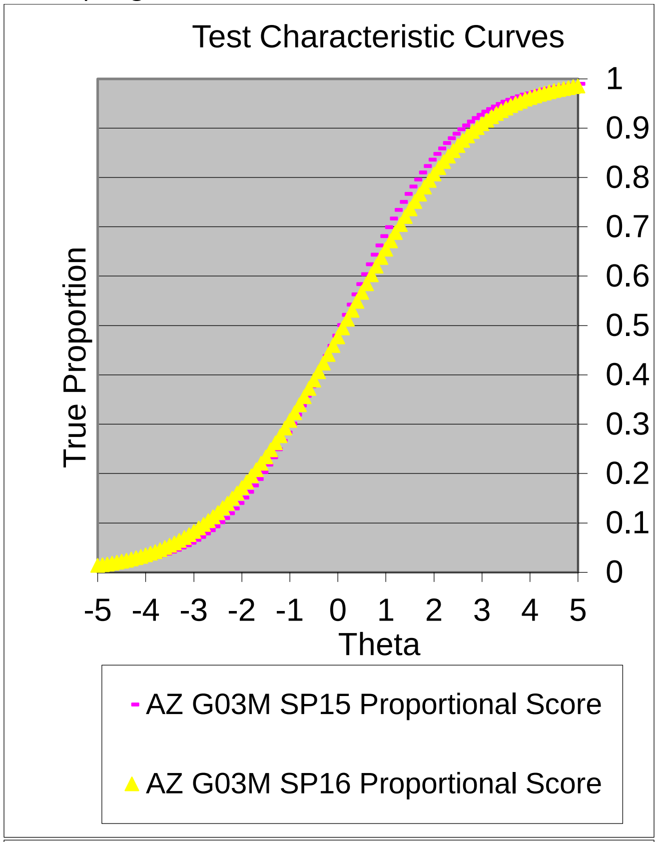
G11ELA Spring 2016 AZ

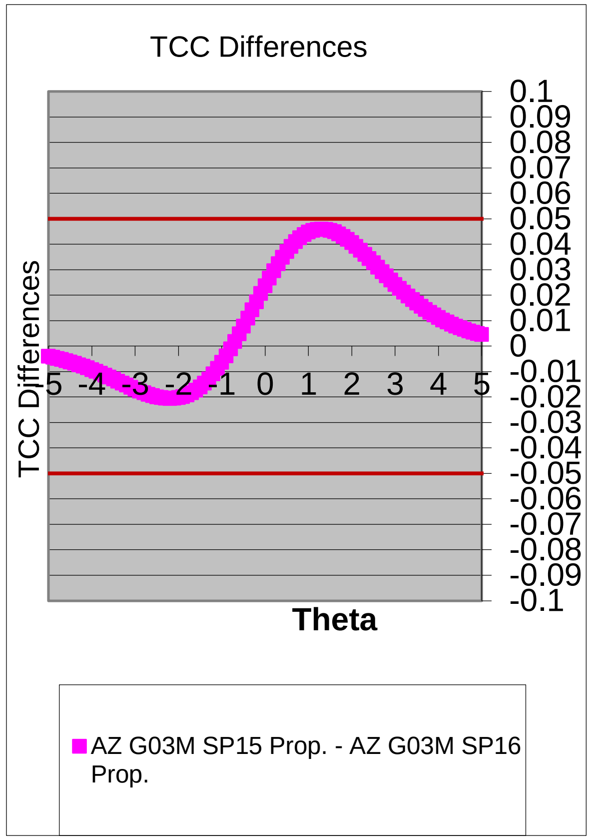


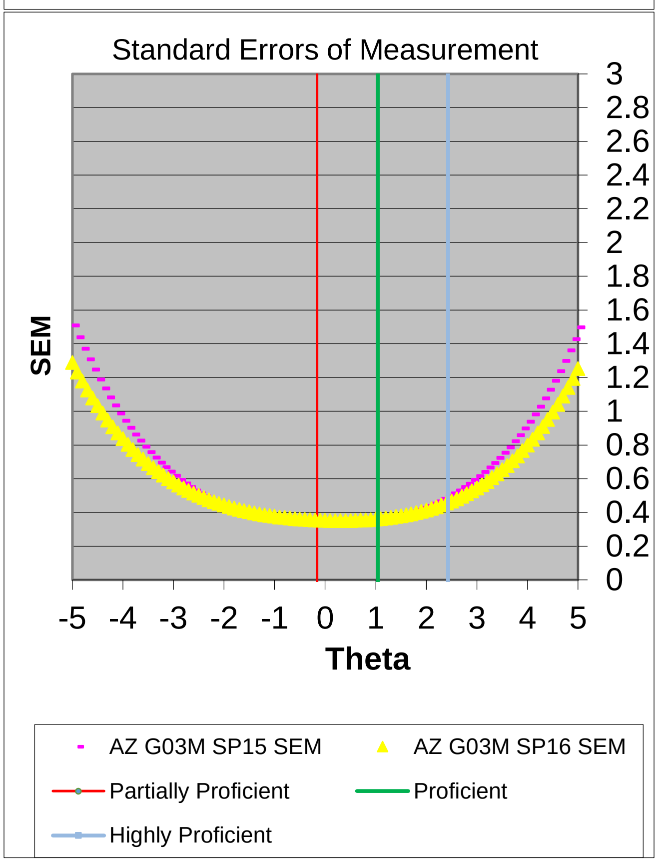




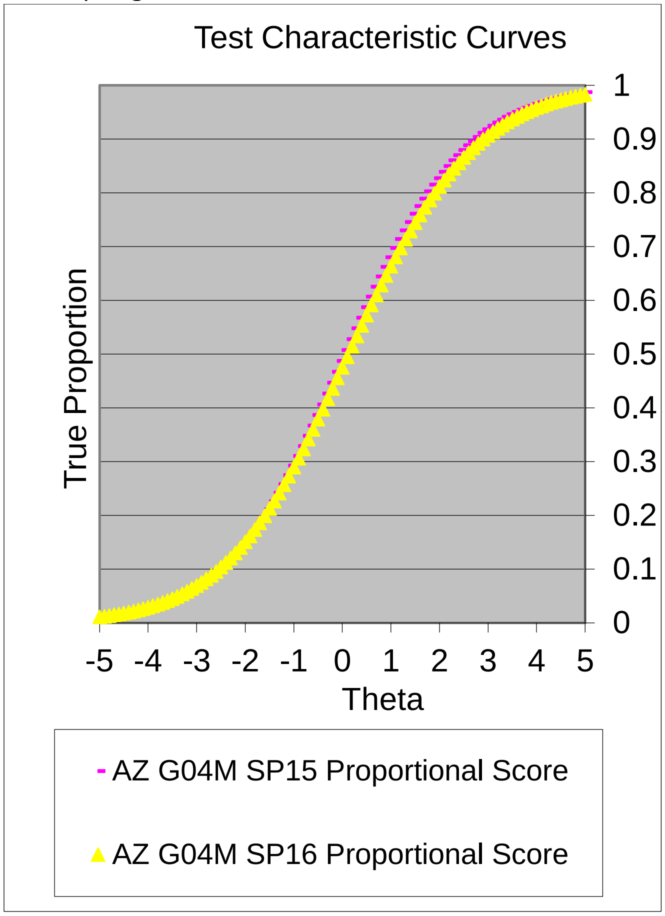
G3M Spring 2016 AZ

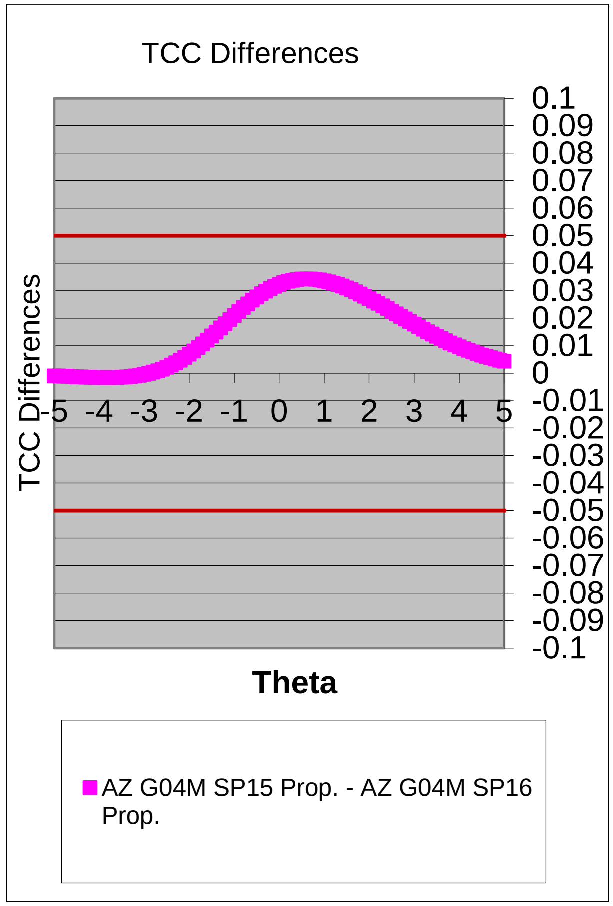


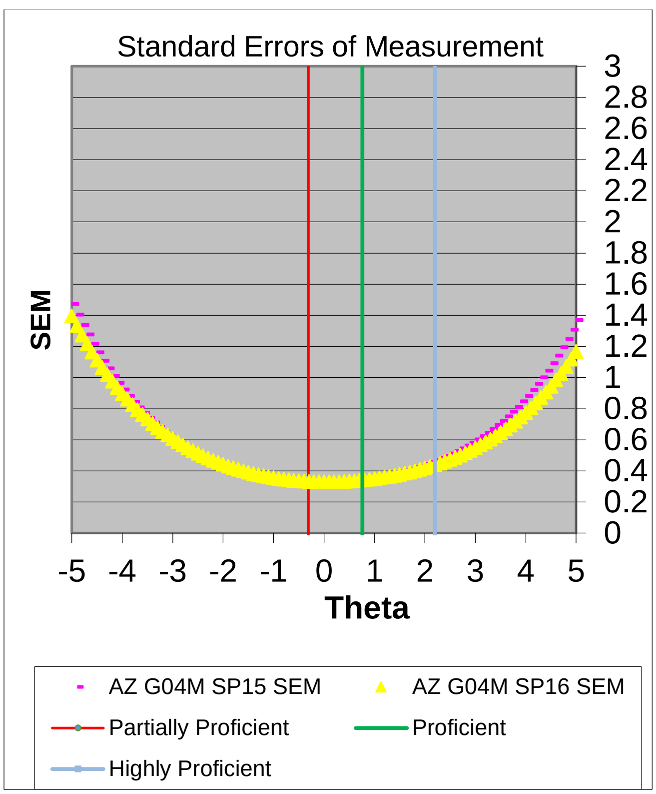


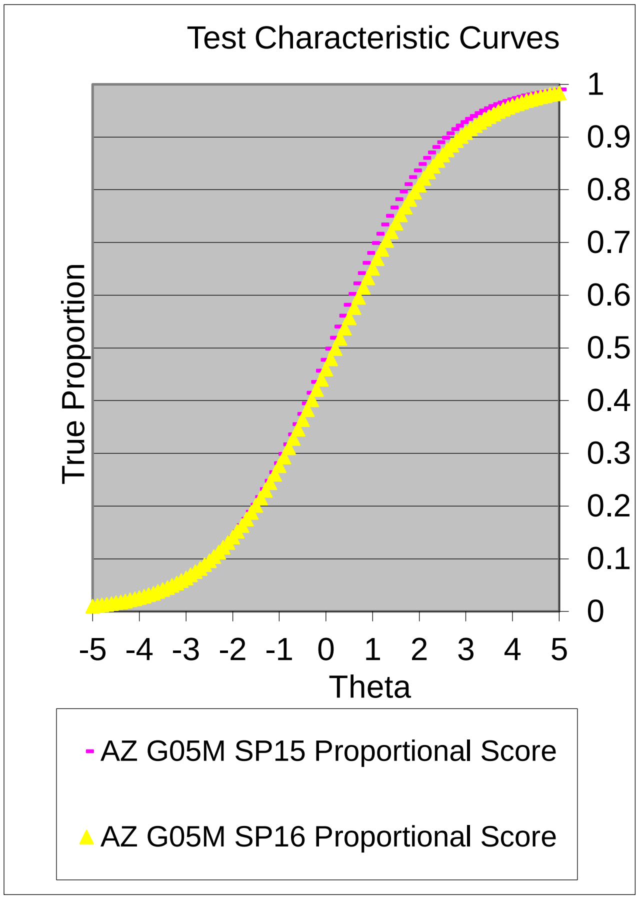


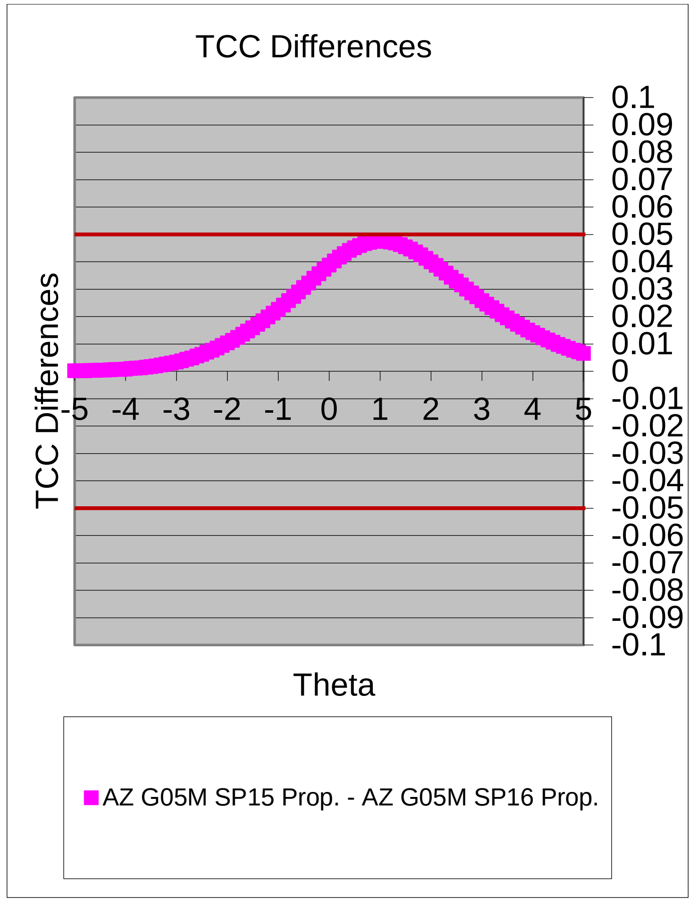
G4M Spring 2016 AZ

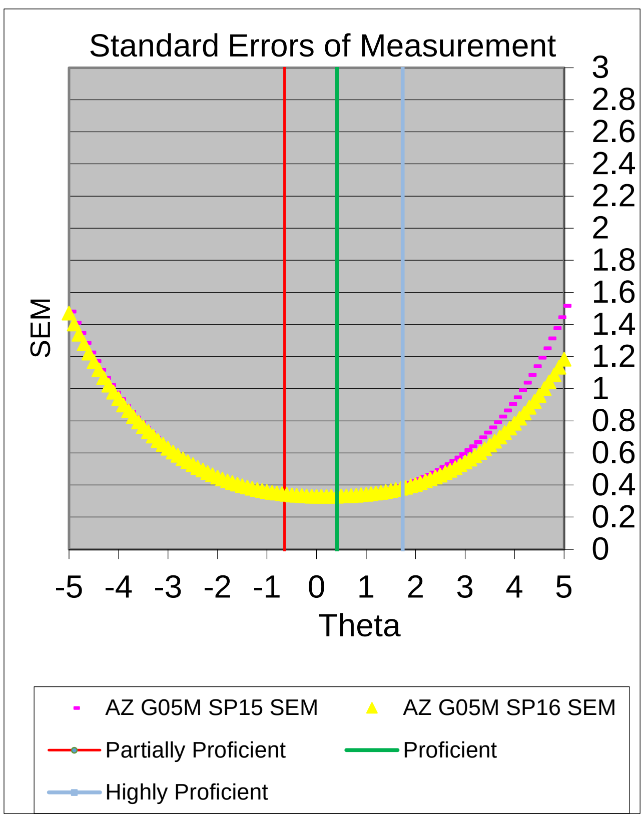


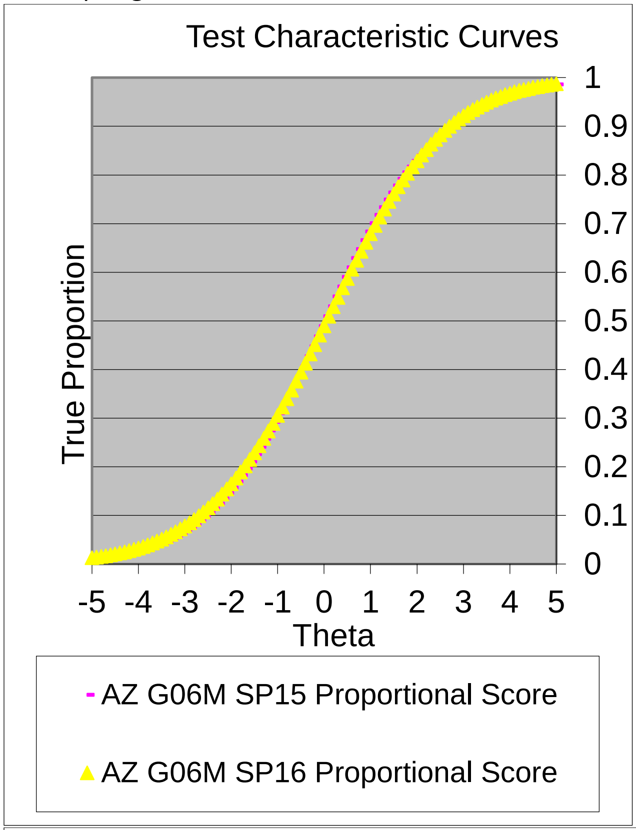


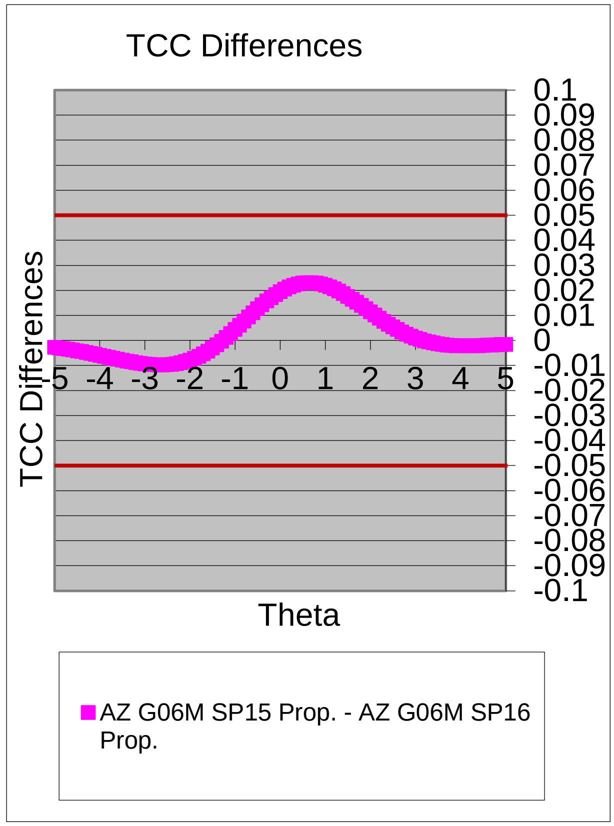


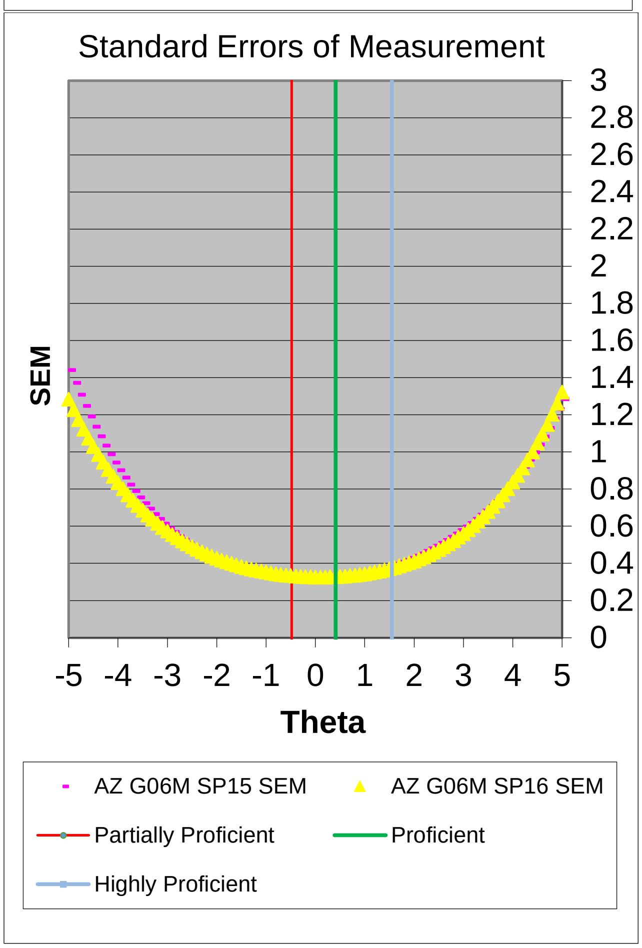

Test Characteristic Curves
True Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G05M SP15 Proportional Score
AZ G05M SP16 Proportional Score


TCC Differences
TCC Differences
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.09
-0.1
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G05M SP15 Prop. - AZ G05M SP16 Prop.

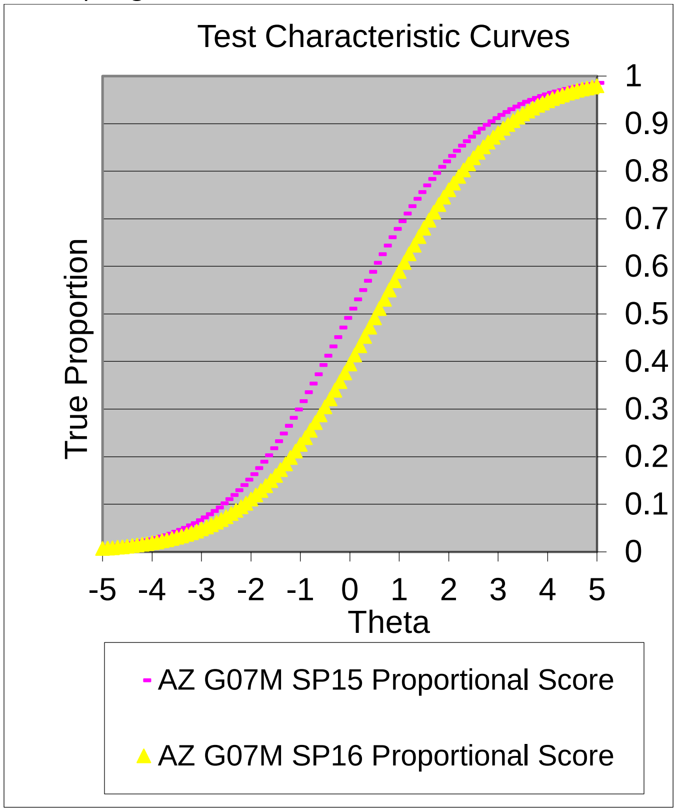

Standard Errors of Measurement
SEM
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G05M SP15 SEM
AZ G05M SP16 SEM
Partially Proficient
Proficient
Highly Proficient

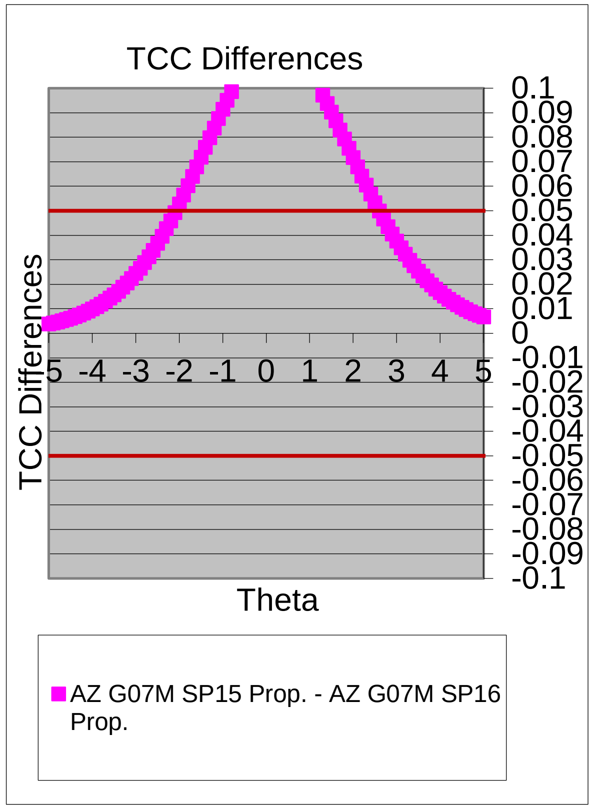

Test Characteristic Curves
True Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G06M SP15 Proportional Score
AZ G06M SP16 Proportional Score

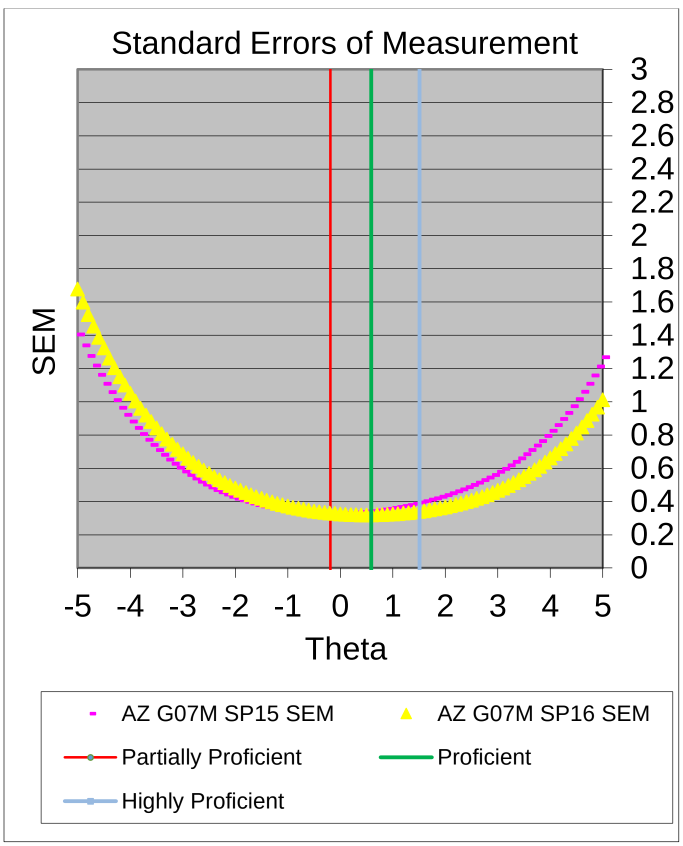

TCC Differences
TCC Differences
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.09
-0.1
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G06M SP15 Prop. - AZ G06M SP16 Prop.


Standard Errors of Measurement
SEM
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
AZ G06M SP15 SEM
AZ G06M SP16 SEM
Partially Proficient
Proficient
Highly Proficient

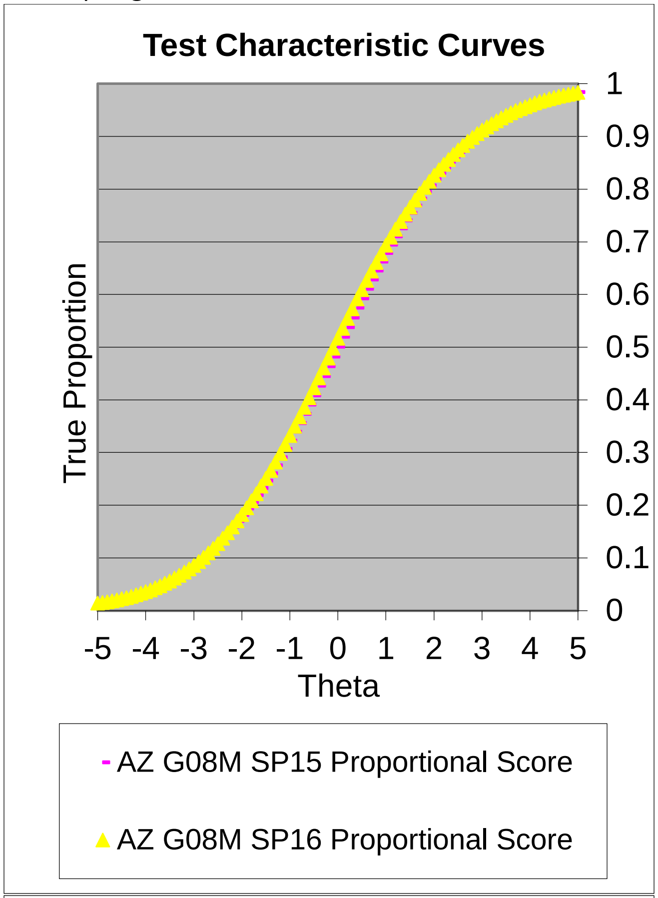
G7M Spring 2016 AZ

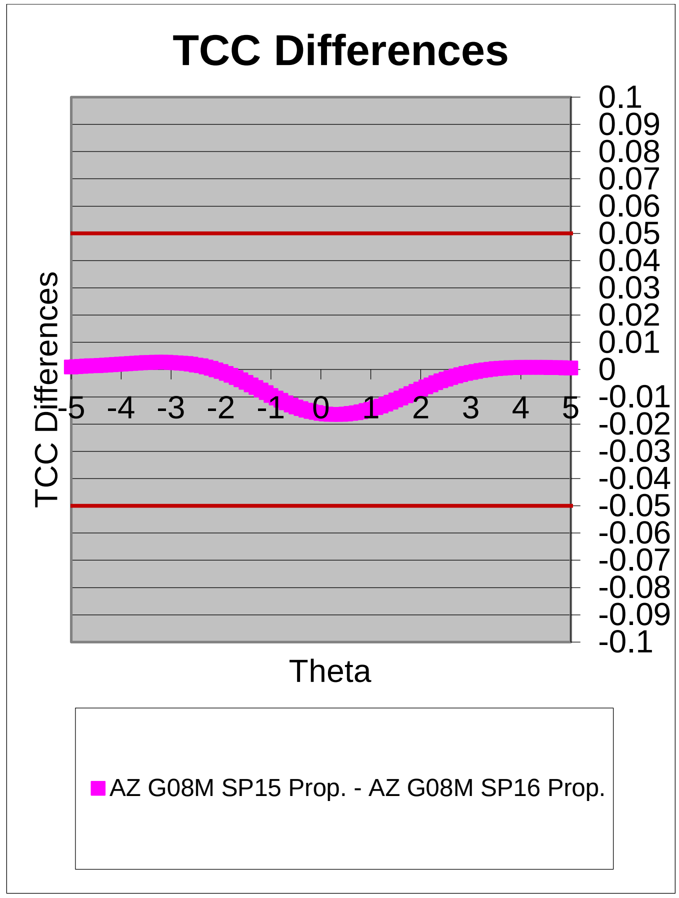


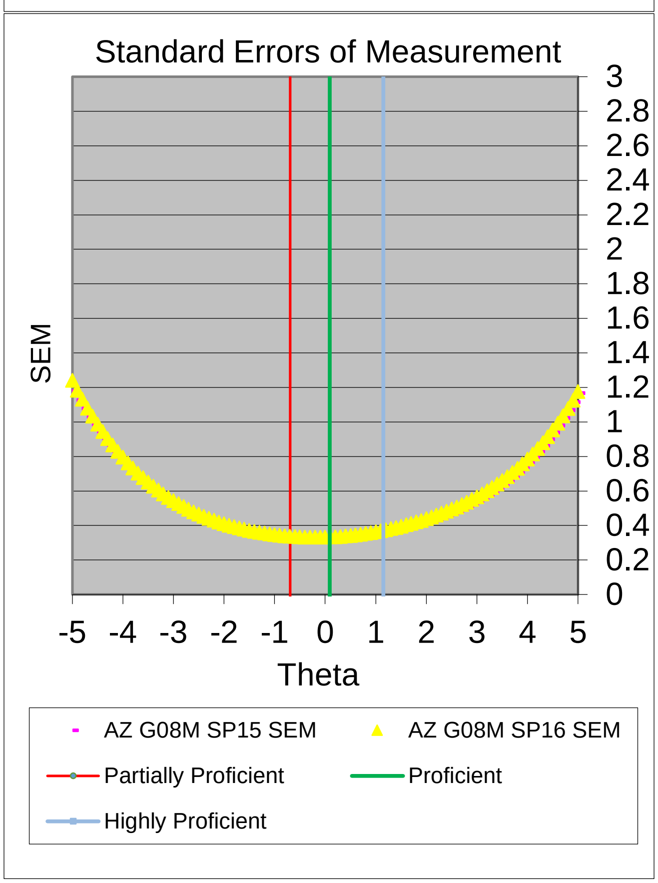




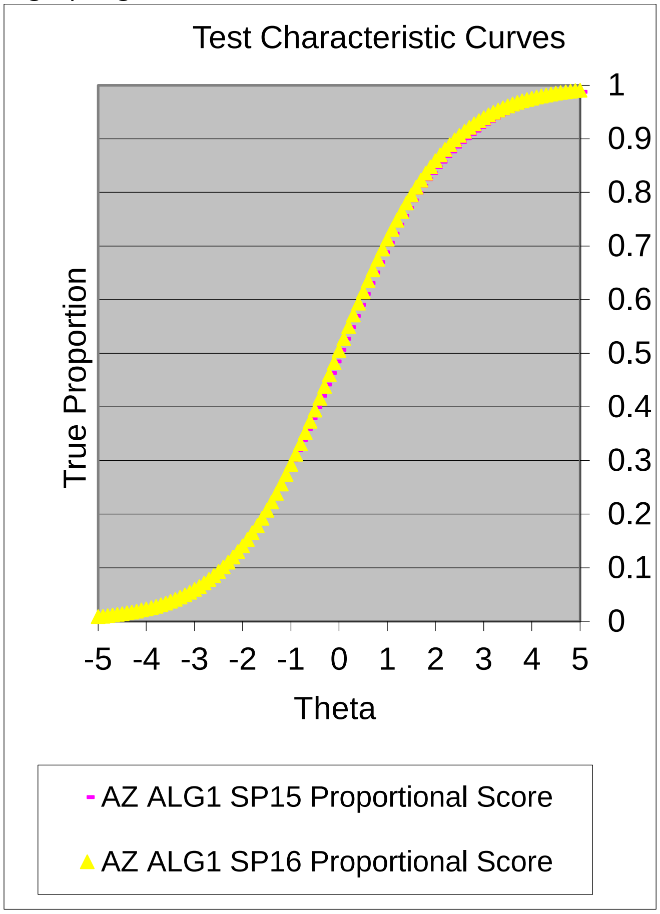
G8M Spring 2016 AZ

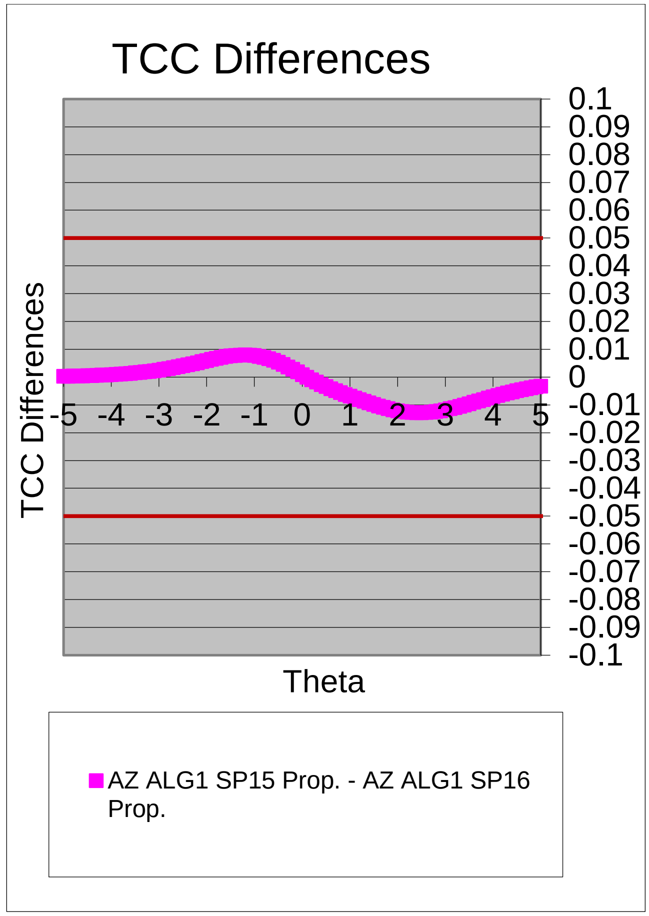


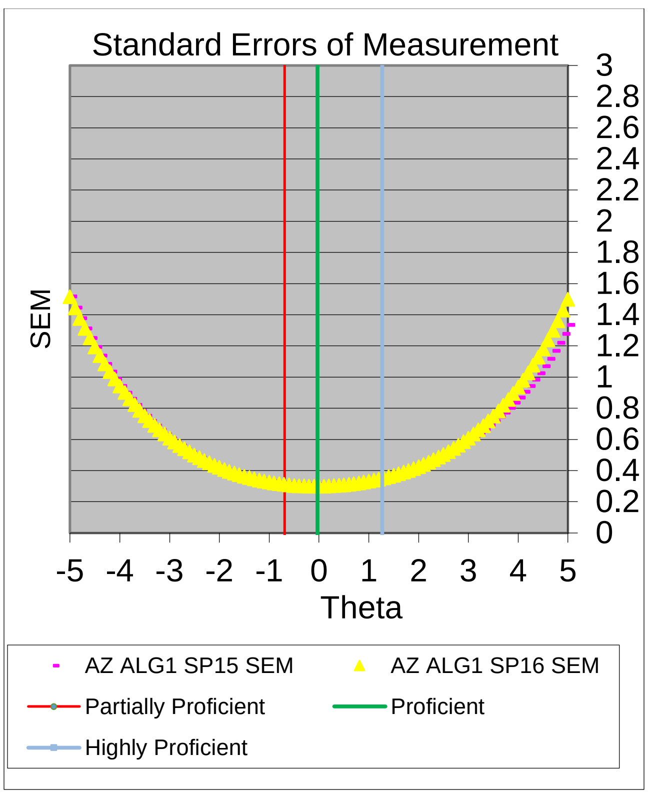




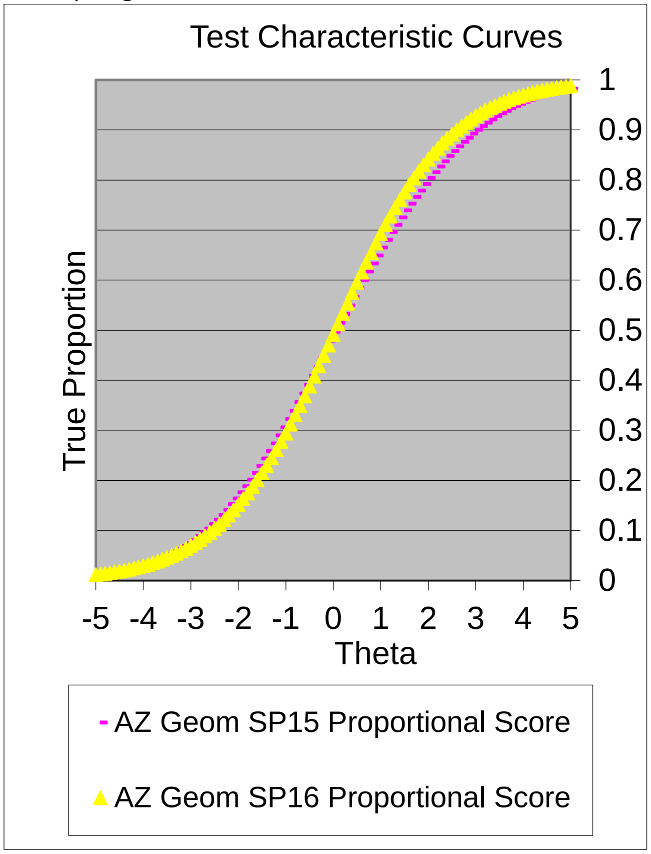
AlgI Spring 2016 AZ

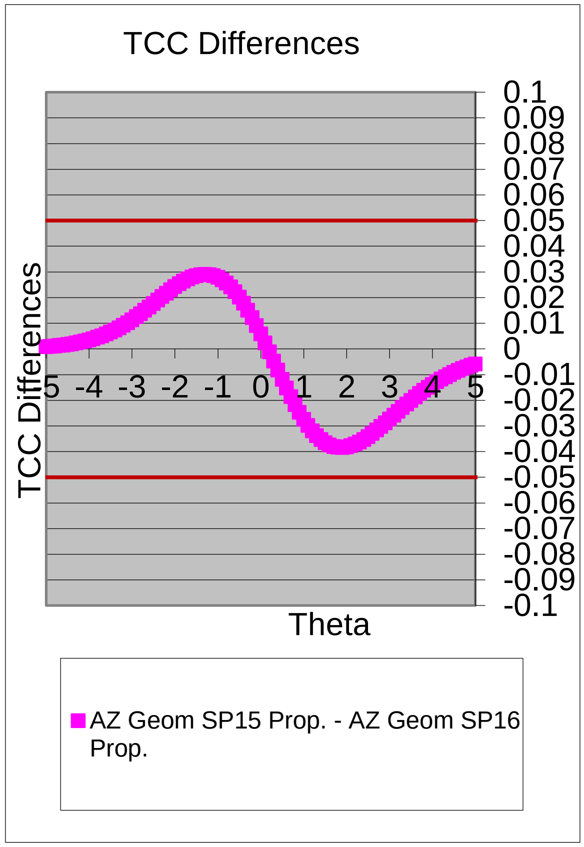


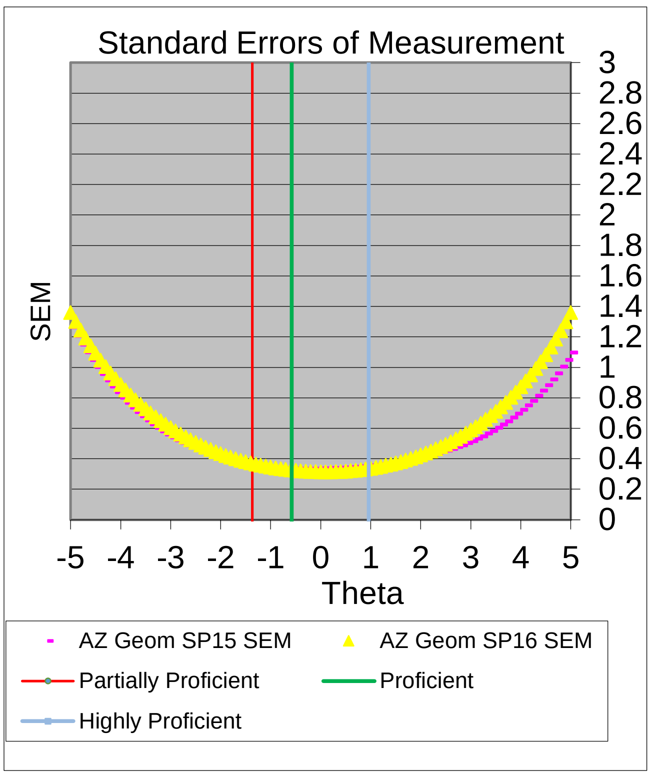




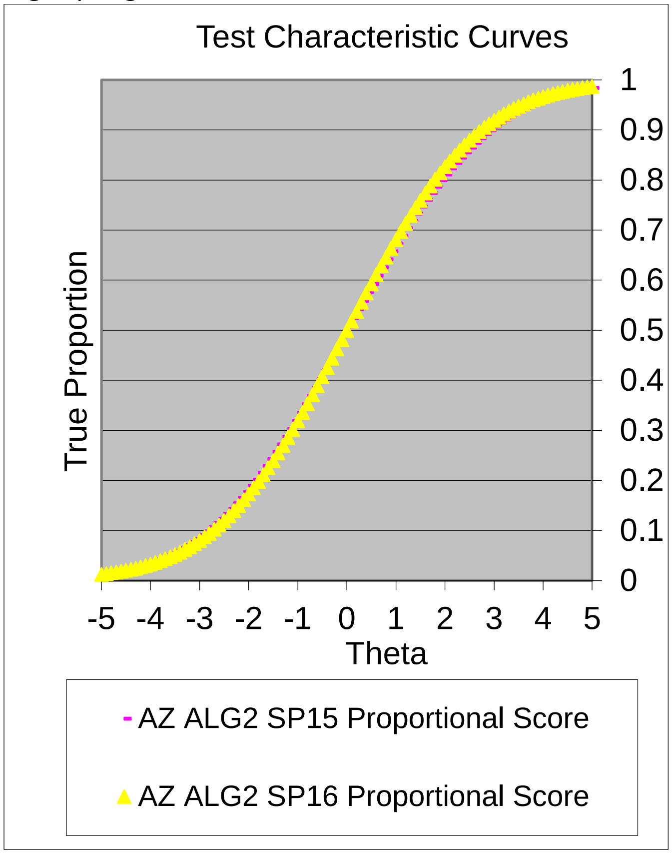
Geo Spring 2016 AZ

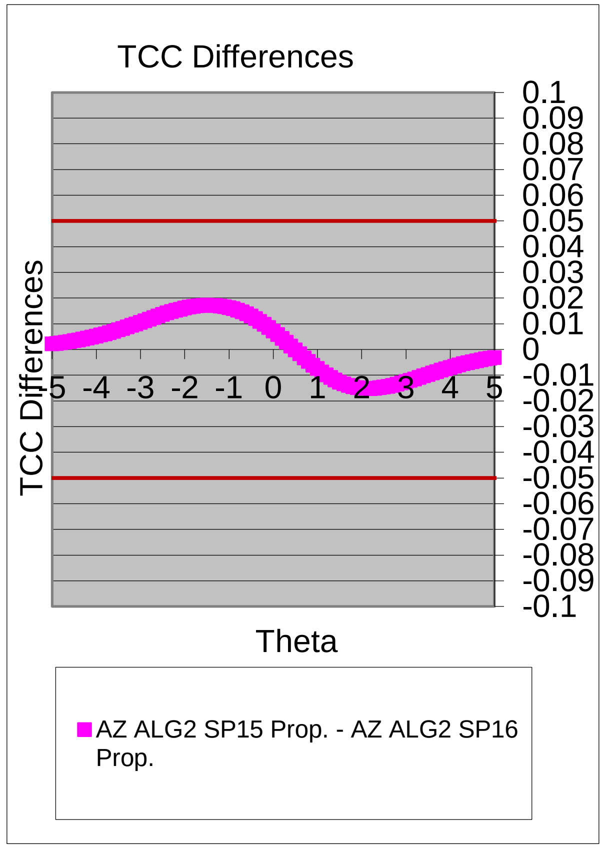


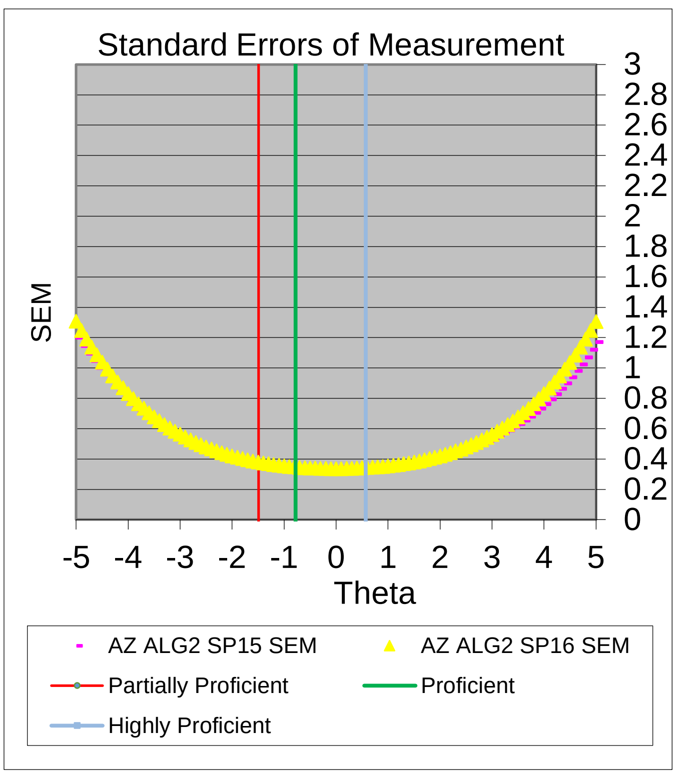




AlgII Spring 2016 AZ







Appendix H:
Sample Issues Log

Sample Issues Log

ID	Date Identified	Entered By	Method of Report	Issue Description	Project Impact	Priority (H, M, L)	Assigned to Owner	Impact Summary	Action Steps	Expected Resolution Date	Escalation Required	Approved By	Current Status	Actual Resolution Date
1	7/8/1c	Kevin Murphy	Weekly check-in meeting	EXAMPLE: A separate meeting needs to be scheduled between appropriate AIR and FDOE staff to discuss the adaptive algorithm and how the dynamic selection of reading items will work	Schedule, resources	M	AIR/FDOE	Meeting scheduled for 7/20/1c	Set up meeting			Yes	Resolved	7/9/1c
2	4/21/1c	Kevin Murphy	Weekly check-in meeting	Example: Alerts for items that have not been uploaded correctly			AIR/FDOE	4/21: AIR is looking into the possibility of adding a feature that will alert content staff when an item has not been uploaded correctly					Open	

Appendix I:
Sample Communications Plan

Communications Plan

Describe the communication types, content categories, ownership and dates of communications in the following matrix.

Phase	Deliverable/ Description	Stakeholder (Audience)	Delivery Method	Timing/ Frequency	Author	Responsible Party (Sender)	Status
Project Management	Schedule updates	NH DOE and AIR		Weekly	AIR Project Director	AIR Project Director	Occurs weekly – on time
Online Assessment System	Updates to key information posted to the online Portal	NH DOE and AIR	Posted to NH DOE ftp site	Based on upcoming milestones and issue	AIR Project Director/ NH DOE Project Director	AIR Project Director	Ongoing

Appendix J:
Sample Agenda: Kickoff Meeting

New Hampshire Statewide Assessments
Project Kickoff Meeting
Agenda
Dates TBD

Location: TBD

Call-In Information: [AIR will provide call-in information upon award of the contract.]

Day 1 GoToMeeting Link: [AIR will provide GoToMeeting link upon award of the contract.]

Day 2 GoToMeeting Link: [AIR will provide GoToMeeting link upon award of the contract.]

Attendees: New Hampshire Department of Education (NH DOE)
American Institutes for Research (AIR)

Day 1: Date TBD

9:00 a.m. – 9:15 a.m.: Introductions and roles on the project

9:15 a.m. – 9:45 a.m.: Overview of the New Hampshire Statewide Assessments; summary of AIR experience with the Smarter Balanced assessments – *New Hampshire and AIR*

9:45 a.m. – 10:15 a.m.: Review of the preliminary project schedule, work plan, and high-level scope of work – *Project Manager, AIR*

a) KnowledgeTree
b) SFTP site for secure student information
c) General review of project management
    I. Weekly meetings agenda and official notes
    II. Issues log
    III. Contact list
d) Program schedule (high-level dates to be presented throughout the meeting)
e) Confirmation of testing window dates (summative and interim assessments)

10:15 a.m. – 10:30 a.m.: Break

10:30 a.m. – 11:00 a.m.: AIR NH DOE Administrative Portal – *Design Team, AIR*

11:00 a.m. – 11:30 a.m.: Training Plan – *Project Manager, AIR*

a) Training modules
b) Online Test Administrator Certification course (optional)
c) AIR webinars
d) Additional resources (posted to the portal)
e) Training schedule

11:30 a.m. – 12:00 p.m.: AIR Help Desk – *Chief Operating Officer, AIR*

a) Overview
b) Reporting capabilities
c) Script and FAQ review/approval process

12:00 p.m. – 1:00 p.m.: Lunch

1:00 p.m. – 2:30 p.m.: Test Information Distribution Engine (TIDE) – *Technical Director, AIR*

a) High-level orientation to the flow of data through AIR systems
b) Student file, user file, institutions file, accommodations file
c) Process for delivery to AIR
d) Layouts
e) TIDE demonstration
f) TIDE configuration specifications

2:30 p.m. – 4:00 p.m.: Online Test Delivery System (TDS) – *Project Manager, AIR*

a) TDS configuration specifications
b) Summative assessments
c) Interim assessments

4:00 p.m. – 4:30 p.m.: Student Result Data Files – *Technical Director, AIR*

a) Layout
b) Process and timing for delivery to New Hampshire
c) Timing

Day 2: Date TBD

9:00 a.m. – 11:00 a.m.: Paper-Pencil Testing and Response Entry into the DEI – *Project Manager, AIR*

a) Materials
b) Distribution and receipt of materials
c) Entry into the DEI
d) Scoring process

11:00 a.m. – 12:00 p.m.: Online Reporting System (ORS) – *Score Reporting Project Manager, AIR*

a) Online Reporting System demonstration including test management center
b) Reports design
c) ORS specifications and recommended configurations

12:00 p.m. – 1:00 p.m.: Lunch

1:00 p.m. – 2:00 p.m.: Next Steps and Closing

Appendix K:
Sample Agenda: Status Meeting

New Hampshire Statewide Assessments
Weekly Status Meeting
Agenda
Date TBD

Call-In Information: [AIR will provide dedicated call-in information for weekly status meetings upon award of the contract.]

GoToMeeting Link: [AIR will provide a dedicated GoToMeeting link for weekly status meetings upon award of the contract.]

Attendees: New Hampshire Department of Education (NH DOE)
American Institutes for Research (AIR)

Welcome and Roll Call

System	UAT Period	Go-Live Date
TIDE	Round 1: Dates TBD Round 2: Dates TBD	Date TBD

Demo Topic 1 (e.g., Practice Tests)

Demo Topic 2 (e.g., Student Registration)

Demo Topic 3 (e.g., TIDE/Technology Requirements Webinar Registrations)

Topic	Training Dates
Technology Requirements for Online Testing	Date TBD
TIDE	Date TBD

Issues Log

Open Forum

Appendix L:
Sample Agenda: Exit Meeting

New Hampshire Statewide Assessments
Exit Meeting: Transition Planning Meeting
Agenda
Dates TBD

Incoming Vendor

Location: TBD

Call-In Information: [AIR will provide call-in information upon award of the contract.]

Day 1 GoToMeeting Link: [AIR will provide GoToMeeting link upon award of the contract.]

Day 2 GoToMeeting Link: [AIR will provide GoToMeeting link upon award of the contract.]

Attendees: New Hampshire Department of Education (NH DOE)
American Institutes for Research (AIR)
Incoming Vendor (TBD)

Day 1: Date TBD

9:00 a.m. – 9:15 a.m.: Introductions and roles on the project

9:15 a.m. – 9:45 a.m.: Overview of the New Hampshire Statewide Assessments, Lessons Learned - *New Hampshire and AIR*

9:45 a.m. – 10:45 a.m.: Draft transition specification document - *New Hampshire Technical Staff, AIR, and Incoming Vendor*

a) Relevant data
b) Test items
c) Testing information
d) Identify other materials necessary for successful transition

10:45 a.m. – 11:00 a.m.: Break

11:00 a.m. – 12:00 p.m.: Development of transition plan – *AIR and Incoming Vendor*

a) Identify all requisite transition information (critical elements)
b) Identify data and format of data delivery
   I. Identify mode of data delivery
   II. Identify timeline of data delivery

12:00 p.m. – 12:00 a.m.: Next Steps and Closing

Appendix M:
Sample Status Report

2017–2018 Status Report – Current Deliverables

9/1/2017

Program/System	Form/Document	Responsibility	Stage	Date Due to Next Stage	NH DOE Priority	Status	Post to Portal Date	Comments/Notes
Training Modules								
NH DOE Statewide Assessments	Insert Training Module Title Here	AIR	Insert description here (e.g., post to portal)	TBD	High	Early	TBD	Insert comments here
Online Webinars								
System Name	Insert Webinar Title Here	AIR	Insert description here (e.g., create presentation)	TBD	Low	Late	TBD	Insert comments here
Manuals								
NH DOE Statewide Assessments – Science	Test Administration Manual	AIR	Insert description here (e.g., ready for shipment)	TBD	Medium	On Sched	TBD	Insert comments here
NH DOE Statewide Assessments	Test Coordinator Manual	AIR	Insert description here (e.g., accessibility review)	TBD	High	On Sched	TBD	Insert comments here
Other Subjects as Required								
NH DOE Statewide Assessments	TBD	AIR	Insert description here	TBD	Low	On Sched	TBD	Insert comments here
Additional Subjects as Required								
NH DOE Statewide Assessments	TBD	AIR	Insert description here	TBD	Low	On Sched	TBD	Insert comments here

Appendix N: Sample Change Control Log

Change Control Log

OAKS Online Assessment				*Project Lead: Meredith Durgin*	*Date of Initial Issues Capture: DATE*											
#	**Status**	**Priority**	**Contract Deliverable**	**Description of Need or Issue**	**Owner**	**Evaluato r**	**Detailed Request**	**AIR Functional**	**Proposed Activity**	**Date16 Identified**	**Date of Decision**	**Estimated Budget**	**Planned Resolution**	**Actual Resolution**	**Closure Action**	**Comments**
1	Complete	Medium	Online Reporting	Oregon Department of Education has requested an additional report within the ORD system	**Austin Fossey**	**Oregon Dept. of Ed.**	**Yes**	**Online Reports**	Online Reporting will work with PM and ODE to determine the specifications of the additional deliverable.	3-Jun-16	7-Jun-16	none	27-Jun-16	25-Jun-16	UAT will be completed by AIR and Oregon Dept. of Ed. Oregon Dept. of Ed. will sign-off	
2	Pending	High	Item Development	Oregon Department of Education has requested additional item development for the ELPA test	**Oregon Dept. of Ed.**	**Oregon Dept. of Ed.**	**Yes**	**Content**	AIR content staff will develop 20 additional items per grade span.	3-Jul-16	8-Jul-16	XXX	TBD	TBD		Schedule needs to be adjusted, and additional resources added to the project.

Appendix P:
Sample Score Reports

Family of John W. Smith

Birth Date: 03/17/2005

School: ABC School (1234567)

District: ABC District (987654)

Ohio | Department of Education

Ohio's State Tests

GRADE 5

SCIENCE

SPRING 2017

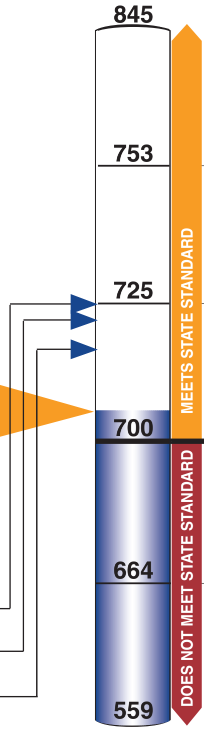
This report provides the score for the state test in Science that John took in spring 2017, explains what the score means, and includes ideas for how your family can help John improve, if needed.

Visit **reportcard.education.ohio.gov** to view your school and district report card.

For information on how you can help your child do better in school, subscribe to parent text alerts. Visit **education.ohio.gov/text** and sign up.

987654-123456-000001

Science assessment



John's score is 706.

He has performed at the proficient level and meets standards for Science.

School Average Score: 725

District Average Score: 721

State Average Score: 717

Advanced - A student with a score of Advanced compares changes in seasons around the world, makes conclusions using real-world facts and plans investigations to solve engineering problems (like choosing materials for a doghouse).

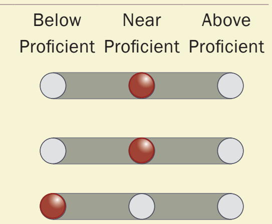
Accelerated - A student with a score of Accelerated can compare tools needed to study objects in our solar system, describe how changes in an ecosystem affect the flow of energy and plan an investigation to study how light is absorbed.

Proficient - A student with a score of Proficient explains differences between objects in our solar system, creates a food web that explain how energy flows through an ecosystem and uses data to compare changes in how an object moves.

Basic - A student with a score of Basic finds the inner and outer planets on a diagram and knows that gravity affects objects, that the sun is the source of energy in an ecosystem (things in an environment) and that light makes shadows.

Limited - A student with a score of Limited has little understanding of our solar system, how energy flows through the organisms in an ecosystem (everything that lives in an environment) and that forces change how an object moves.

Has John reached proficient in the areas of Science?



This chart shows you how well John performed in each area. He is near proficient in Earth Science, is near proficient in Life Science, and is below proficient in Physical Science.

What are your child's strengths and weaknesses in Science?

Earth Science

John Scored Near Proficient

Students understand that scientists use technology to study outer space. They describe general facts about objects in our solar system (like size, composition and movement). They compare the sun with other stars in the universe. They explain the cycles and patterns of motion between the sun and Earth like seasons and day and night.

WHAT THESE RESULTS MEAN
Your child knows that telescopes and satellites are used to study our solar system. He uses data to compare planets and asteroids or comets. He explains the difference between Earth's revolution and rotation.

NEXT STEPS
With your child, use the Internet or carry out an investigation to test the relationship between the angle of sunlight, the amount of sunlight and temperature. Then, use the results to describe differences in seasonal patterns around the world.

Life Science

John Scored Near Proficient

Students understand the relationships within an ecosystem. They recognize the roles of organisms (producers, consumers, decomposers) in an ecosystem are based on how they get energy. They know that sunlight is the major source of energy in ecosystems. They can describe the effects of a species invading an ecosystem.

WHAT THESE RESULTS MEAN
Your child identifies plants as producers in an ecosystem. He knows populations of organisms have relationships where one group may benefit or be harmed. He creates a food web to show how energy flows from herbivores (eat plants) to carnivores (eat meat).

NEXT STEPS
Use the Internet to research organisms in an ecosystem. Ask your child to make cards with one organism and its role (producer, consumer, decomposer) on each card. Have him create a food web using the cards to show how energy flows through the ecosystem.

Physical Science

John Scored Below Proficient

Students understand that forces affect the motion of an object. They know Earth pulls down on all objects (gravity). They measure how an object moves by using distance and time to solve for speed (how fast or slow). They know that light and sound are forms of energy. They compare behaviors that describe how light and sound travels.

WHAT THESE RESULTS MEAN
Your child may be able to identify changes in the motion of an object and explain how forces and the mass/weight of an object affect its motion. He may have little understanding of how light travels through some materials or how light behaves as it hits objects.

NEXT STEPS
Ask your child to gently push two different objects across the floor. Ask him to compare the mass/weight of the objects and describe the motion of the objects. Ask him to compare the distance and relative speed (how fast or slow) the objects traveled.

Student Name: Jane Doe
School: Aloha Elementary School
Complex Area: Ewa
Test Date: 2015–2016

The student's name may have been truncated due to space limitations.

FAMILY Report

Science
Assessment Results

Dear Doe Family:

The Hawai'i Department of Education is pleased to send you this report about Jane's performance on the Online Hawai'i State Science Assessment. The Science Assessment is designed to test students on the Hawai'i Content and Performance Standards, Third Edition (HCPS III) learned in the previous school year. The standards describe what students should know and be able to do in science.

Students take each assessment up to three times during the school year. This report shows Jane's best performance on the Science Assessment, which counts as her official score.

In addition to showing how well Jane did on the assessment, this report compares her score with those of other students in her school, her complex area, and the state. On the bottom of pages 2 and 3, the report also shows whether or not Jane reached proficiency in the different areas of science and suggests how you may help her to further her knowledge and skills.

This report is a starting point for a discussion with Jane's teacher. You may use it to talk about how you can support your child's learning at home. Informed students, parents, and schools working together provide the best education for our students.

Very truly yours,

Kathryn S. Matayoshi
Superintendent of Education

What is in this report?

- Jane's score on the Science Assessment
- How Jane's score compares
- The areas that make up the Science Assessment
- Whether Jane reached proficiency in the different areas of Science
- How you can help Jane improve her science knowledge and skills

For more information about this assessment, go to www.alohahsap.org

Level 1
Administered in Grade 4

2015-2016

Hawai'i Department of Education

Photograph: Lignum Vitae
Selvin Chin-Chance

FAMILY Report

Jane's Science Score

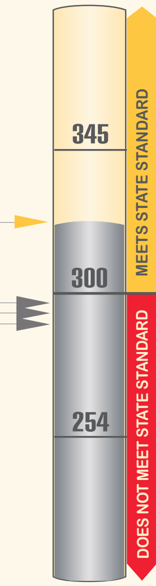
325
Meets Proficiency

How does Jane's score compare?

Jane's Science score is 325. This score is higher than the average score of fourth graders in her school, higher than that of fourth graders in her complex area, and higher than that of fourth graders statewide.

Jane's Score: 325

State Average: 297
Your Complex Area: 295
Your School: 290



Exceeds Proficiency - Students can create hypotheses; predict the effect of a decrease in the plant population on an ecosystem; give examples of differences between plant and animal cells; predict new substances formed when materials are combined; describe the direction of Earth's gravity; and use evidence to describe how fast and slow processes have shaped Earth's surface.

Meets Proficiency - Students can describe hypotheses; explain how simple food chains and webs can be traced to plants; identify basic differences between plant and animal cells; describe how materials may be combined to form new substances; describe that Earth's mass exerts a gravitational force; and describe how the shaping of Earth's surface is due to fast and slow processes.

Approaches Proficiency - Students can identify hypotheses; describe how simple food chains or webs can be traced to plants; recognize one difference between animal and plant cells; identify how materials may be combined to form new substances; recognize that Earth exerts a gravitational force; and identify examples of the shaping of Earth's surface by fast and slow processes.

Well Below Proficiency - Students at this level demonstrate skills and understanding of Science below the performance required to reach the Approaches Proficiency level.

Go to www.alohahsap.org to see a complete listing of knowledge and skills for each level.

A student's test score can vary if the test is taken several times. If your child were tested again, it is likely that Jane would receive a score between 300 and 350.

Has Jane Reached Proficiency in the Three Different Areas of Science?

Next Steps

Scientific Process

Discover, invent, and investigate using the skills necessary to engage in the scientific process; understand that science, technology, and society are interrelated.

Near **The test does not always provide enough information to tell if a student has reached the Meets Proficiency mark for this area of Science.**

Students may be able to describe a hypothesis, distinguish between observations and inferences, and describe how the use of technology has influenced Hawai'i's economy, demography, and environment.

For example, show your child an ice cube and a glass of water. Ask your child to form a hypothesis about what will happen if the ice cube is placed into the glass of water (e.g., "If I put the ice into the glass of water, then it will melt.") Ask her to test her hypothesis by putting the ice into the water and recording what she observes.

Life Science

Understand the interrelationships of organisms; understand the structures and functions of living organisms; understand the impact of genetics and biological evolution on the unity and diversity of organisms.

No **The score is below the Meets Proficiency range for this area of Science.**

Students may have difficulty explaining the role of plants in a food chain (diagram), identifying some differences between plant cells and animal cells, and describing how different organisms need specific environmental conditions to survive.

For example, help your child draw a food web using one plant and four animals that live in the ocean. Talk about how the plant is an important part of the food web (e.g., almost all animals' food can be traced back to plants). In addition, ask your child how one of the ocean animal's body parts helps it survive in the ocean (e.g., sea turtles have paddle-like front arms for swimming).

Similar Schools

This chart shows how fourth-grade students in Jane's school did compared with fourth-grade students in other schools when tested on concepts and skills that were taught in the previous years. School similarity is determined using three criteria: percentage of (1) disadvantaged students, (2) English language learners, and (3) students with disabilities.

The schools that were compared with your school were chosen because their fourth-grade students had backgrounds most similar to fourth-grade students in your school. Aloha Elementary School teaches many disadvantaged students and many students with disabilities.

Students come from many different environments. These differences do not necessarily affect student performance. Many issues contribute to student performance, such as administration and oversight, curriculum and content, teaching and testing, professional development, instructional materials, and parent and community support. You may want to contact your parent community networking coordinator to inquire about parent workshops that support the school's ongoing effort to improve student performance.

How your child's school compares

5%

About **5%** of similar schools did better than your school.

60%

About **60%** of similar schools had about the same average score as your school.

35%

About **35%** of similar schools did not do as well as your school.

Next Steps

Physical, Earth, and Space Sciences

Understand Earth and its processes, the solar system, and the universe and its contents.

Yes The score is at or above the Meets Proficiency range for this area of Science.

Students predict the new substances that are formed when some materials are combined, use materials to set up a circuit to create light and sound, describe that Earth exerts a gravitational force toward its center on all objects, use evidence to describe how fast and slow processes have shaped and reshaped Earth's surface, and use evidence to describe the relationship between the sun and Earth's daily rotation and annual revolution.

For example, in a darkened room, use a small lamp to represent the sun and a ball to represent Earth. Ask your child to move (e.g., rotate) the ball to represent alternating day and night. Ask your child to move (e.g., in a circle around the lamp) the ball to represent one year. Finally, ask your child to express the relationship between the motions associated with days and years (e.g., 365 rotations in every revolution around the lamp) by performing both motions at the same time.

Additional Resources

Smithsonian Education for Students
http://www.smithsonianeducation.org/students/

Exploratorium Snacks
http://www.exploratorium.edu/snacks/index.html

NASA's The Space Place for Kids
http://spaceplace.nasa.gov/en/kids/muses2.shtml

Energy for Kids
http://www.eia.gov/kids/

Kids.gov
http://kids.usa.gov/

Lawrence Hall of Science
http://www.lawrencehallofscience.org/kidsite
Click "Play Game & Activities" for activities and games in science.

National Science Foundation
http://www.nsf.gov/news/classroom
Click science topics such as "Astronomy & Space."

To see sample questions from the Science Assessment and how they were scored, go to
www.alohahsap.org

Frequently Asked Questions

Q: What is the Hawai‘i State Science Assessment?

A: The Hawai‘i State Science Assessment is a yearly testing program that measures student achievement in meeting Hawai‘i's science standards or expectations.

- During the 2015–2016 school year, Hawai‘i's public school and public charter school students in grades 4 and 8 were tested in science. The assessment measures whether students have learned the science knowledge and skills expected of them during the current school year.

Q: What are the Hawai‘i Content and Performance Standards (HCPS)?

A: In 1999, the Hawai‘i State Department of Education established learning expectations, known as the Hawai‘i Content and Performance Standards (HCPS), for Hawai‘i's students. These high academic standards give students clear achievement goals and help guide instruction in the schools.

- These standards identify important ideas, concepts, and skills students should know, care about, and be able to demonstrate.
- State or federal laws require yearly testing of students in certain grades in reading, mathematics, and science. The laws require that the Hawai‘i State Assessments provide clear information on how well your child is meeting these standards.

Q: How does the Hawai‘i State Science Assessment benefit my child?

A: The assessment can help identify whether a student needs extra support and practice in science. Teachers and families can then work together to ensure that a student gets the help he or she needs.

Prepared for

The Family of Jane Doe

Test Date: Spring 2014 **Birth Date:** 06/11/2003

School: ABC ES (012345)

District: ABC City (056789)

Ohio | Department of Education

Grade 5 Reading, Math and Science Achievement Assessments

Dear Family,

In May, your child Jane took the Ohio Achievement Assessments in Grade 5 reading, math and science. These tests measured how much she had learned at the end of the school year, based on state standards for what a fifth-grader should know and be able to do.

This report lists Jane's test scores and explains what they mean. I urge you to talk to her teacher about the report so you have a good understanding of what she needs to continue growing in knowledge.

Next spring we will begin moving into a new generation of assessments based completely on Ohio's New Learning Standards. Under these standards, teachers will focus on fewer topics within each classroom subject but go deeper into those topics. Over the course of her school years, this will give Jane the more in-depth education she will need to succeed in 21st-century work and life.

Thank you for all you do to support Jane in school. I wish her much success in her learning.

Sincerely,

Richard A. Ross

Dr. Richard A. Ross
Superintendent of Public Instruction

Quick Facts About Your Child and State Achievement Assessments

Q. Why does my child take achievement assessments?

The Reading, Math, and Science Achievement Assessments help ensure that all students have learned the level of reading, math, and science expected of Ohio students at the end of the fifth grade.

Q. What happens if my child does not pass the Grade 5 Achievement Assessments?

These test results are one of the factors, but not the sole factor, that teachers use to evaluate a child's readiness to move to the next grade. Classroom teachers determine whether your child should move to Grade 6 by considering her class work and school attendance.

If your child does not score in the proficient or above levels, this is a sign that she is having trouble learning the skills and knowledge expected of fifth-grade students and may need additional help. This report includes suggested activities to help your child practice specific skills. Your child's teacher may be able to recommend additional ways your child can improve her achievement.

Reading Achievement Results

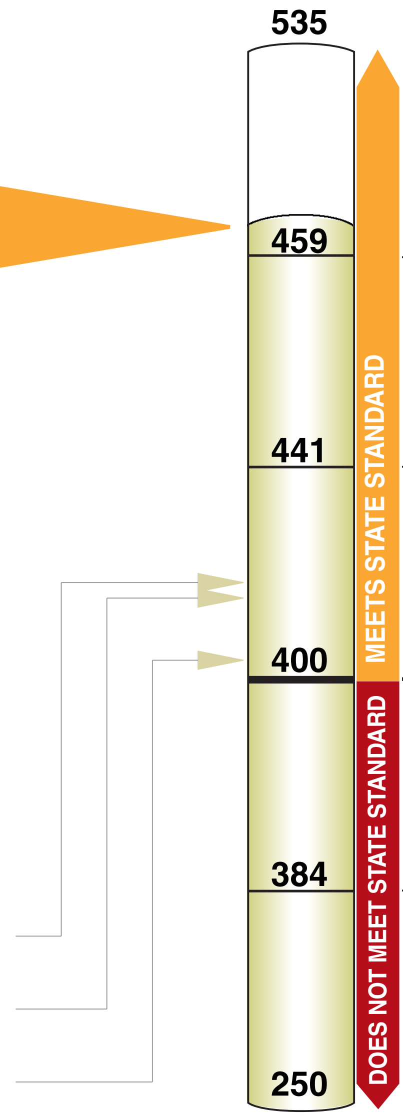
Jane's score is 470.

She has performed at the advanced level and meets standards for Grade 5 Reading.

School Average Score: 419

State Average Score: 416

District Average Score: 404



Advanced - Students go beyond simply understanding what they read and can make good judgments about the author's use of fact versus opinion.

Accelerated - Students understand what they read (e.g., plays, biographies, newspapers) and can summarize stated and unstated themes.

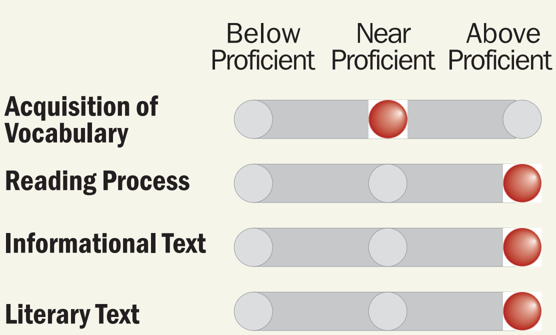
Proficient - Students understand what they read. They try to explain how an author's choice of words creates the story's mood and describes the story's setting.

Basic - Students understand some of what they read. They use strategies (e.g., authors' definitions and examples) to learn the meaning of new words.

Limited - Students may struggle with simple reading tasks (e.g., plays, biographies, newspapers).

Different versions of the test are used each time the test is given. The scores in this report are used to compare performance from year to year. Scores in different subjects cannot be compared.

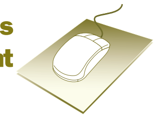
Has Jane reached proficient in the different areas of Reading?



This chart shows you how well Jane performed in each standard. She is near proficient in Acquisition of Vocabulary, above proficient in Reading Process, above proficient in Informational Text, and above proficient in Literary Text.

Note: There is some uncertainty in all test scores. Please see the interpretive guide for more information.

See the released test questions from earlier years' achievement assessments at www.success.ode.state.oh.us



What Are Your Child's Strengths and Weaknesses in Reading?

Acquisition of Vocabulary

Students use clues in the text, knowledge of word parts (e.g., *underneath*) and knowledge of where words come from (e.g., *autograph; graph* means *to write* in Greek) to learn new words. They understand figurative words (e.g., *the bus crawled through the traffic*).

Jane Scored Near Proficient

WHAT THESE RESULTS MEAN

Your child is learning to use word parts (e.g., *eraser*) and clues in the texts to learn the meaning of new words. She understands figurative language (e.g., *the saw bit into the wood*). Your child can also abbreviate words (e.g., *St.* for *Street*; *Ave.* for *Avenue*).

NEXT STEPS

Have your child hunt for words with a specific part, such as prefixes (e.g., *un-*, *unhappy*; *pre-*, *preview*) and suffixes (e.g., *-ing*, *writing*; *-less*, *careless*). Choose other word parts and do the searches together.

Reading Process

Students make and support predictions about what they read, and they summarize information. They can check their understanding and read in different ways (e.g., skimming, re-reading). They can use outlines to understand what they read.

Jane Scored Above Proficient

WHAT THESE RESULTS MEAN

Your child can skim materials (e.g., dramas, biographies, newspapers) to find facts that support her own judgments about a text. Your child can make judgments about whether certain reading materials provide the information she needs.

NEXT STEPS

Have your child read two articles on the same topic and make a chart of how the articles are similar or different. Have your child explain why one article is better than the other.

Informational Text

Students use text features (e.g., subheadings) to locate important information. Students compare information from different sources, summarize main ideas, and distinguish between important and unimportant details.

Jane Scored Above Proficient

WHAT THESE RESULTS MEAN

Your child is able to pull out complex new information from what she reads. She understands how the author organizes ideas (e.g., as they happen, in question-answer format). This helps her understand the author's message.

NEXT STEPS

Have your child preview difficult informational texts, such as articles in magazines or newspapers. Have your child read a text that includes a chart and explain the relationship between the chart and the article.

Literary Text

Students identify how the main events of a story affect what happens later in the story. They analyze setting, character and theme and how they affect each other. They explain how authors use figurative words (e.g., metaphors, similes) to create mood.

Jane Scored Above Proficient

WHAT THESE RESULTS MEAN

Your child can evaluate setting, character and theme and how these elements affect each other (e.g., setting can influence plot). She can explain how the story's point of view (first or third person) gives a reader a different understanding of character.

NEXT STEPS

Have your child read complex novels and discuss how the literary elements influence each other (e.g., how the setting affects the plot, how the characters affect one another). Ask your child to compare the theme of this novel with the themes of other books.

Math Achievement Results

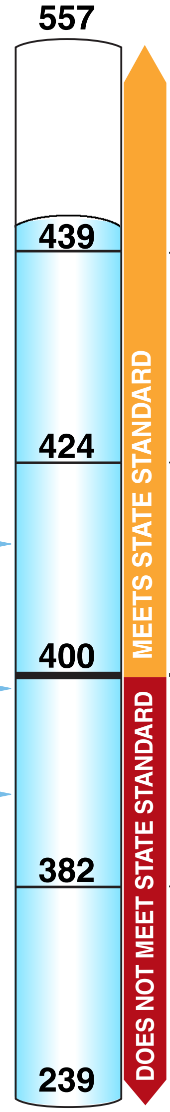
Jane's score is 453.

She has performed at the advanced level and meets standards for Grade 5 Math.

State Average Score: 415

School Average Score: 399

District Average Score: 390



Advanced - Students can solve new, complex problems (e.g., find the area of an irregular shape). They explain their reasoning using math terms (e.g., *probability*) and notation.

Accelerated - Students combine math ideas to solve problems (e.g., use multiplication to find area). They explain their reasoning (e.g., use terms and graphs to describe data).

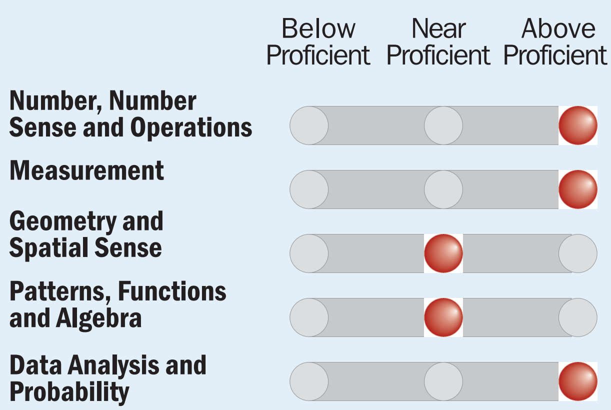
Proficient - Students can solve common problems with fractions, percents and tools (e.g., protractor). They estimate solutions and use informal reasoning skills.

Basic - Students can remember math terms (e.g., *mean*) and concepts (e.g., using variables). They carry out routine procedures, such as finding areas and reading graphs.

Limited - Students struggle with math terms (e.g., *ray*) or concepts. They have trouble adding decimals and negative numbers and using words to describe number patterns.

Different versions of the test are used each time the test is given. The scores in this report are used to compare performance from year to year. Scores in different subjects cannot be compared.

Has Jane reached proficient in the different areas of Math?



This chart shows you how well Jane performed in each standard. She is above proficient in Number, Number Sense, and Operations, above proficient in Measurement, near proficient in Geometry and Spatial Sense, near proficient in Patterns, Functions and Algebra, and above proficient in Data Analysis & Probability.

Note: There is some uncertainty in all test scores. Please see the interpretive guide for more information.

See the released test questions from earlier years' achievement assessments at www.success.ode.state.oh.us

What Are Your Child's Strengths and Weaknesses in Math?

Number, Number Sense and Operations

Students add and subtract fractions and decimals; show numbers in different ways (e.g., $1 = {}^{2}/_{2} = 1.0 = 100\%$); use rules of operations (+, -, ×, /); and understand ratios (1 to $2 = {}^{1}/_{2}$), negative numbers and perfect squares (e.g., $81 = 9 \times 9$). They solve problems in many ways.

Jane Scored Above Proficient

WHAT THESE RESULTS MEAN

Your child can add, subtract and round fractions, decimals and percents. She can create and use equivalent forms of fractions (e.g., ${}^{2}/_{6} = {}^{1}/_{3}$), decimals (e.g., ${}^{1}/_{2} = 0.5$) and percents (e.g., ${}^{3}/_{4} = 0.75 = 75\%$). She can make estimates and determine their reasonableness.

NEXT STEPS

Have your child find examples of ratios around the house (e.g., 5 blue plates to 9 total plates). Help her use ratios to solve related problems such as, "The ratio of apples to bananas is 3 to 2. There are 15 apples. How many bananas are there?"

Measurement

Students draw and measure angles. They develop formulas for finding perimeter, area and volume; describe the difference between surface area and volume; and make conversions among measurement units (e.g., inches to feet).

Jane Scored Above Proficient

WHAT THESE RESULTS MEAN

Your child can make conversions between units and solve multistep problems involving measurements. She can develop formulas for finding perimeter, area and volume.

NEXT STEPS

Have your child find the area and perimeter of more complex shapes, such as parallelograms. Encourage your child to use a variety of methods, such as estimating, measuring or using formulas.

Geometry and Spatial Sense

Students identify relationships among parts of circles (radius, diameter, circumference); understand congruence; plot negative values on a grid; and recognize two-dimensional shapes that can be folded to make a three-dimensional object.

Jane Scored Near Proficient

WHAT THESE RESULTS MEAN

Your child can use physical models to find the sum of interior angles of shapes. She can find points with negative values on the coordinate grid. She can usually use properties of congruent figures to solve problems.

NEXT STEPS

Cut, unfold and flatten a box, a hollow tube (paper towel tube) or an oatmeal container along the edges. Ask your child to imagine what the original container looked like and to name the object.

Patterns, Functions and Algebra

Students can write rules to represent patterns; use variables as unknown quantities; create and interpret equations and inequalities; and use graphs and tables to draw conclusions and make predictions.

Jane Scored Near Proficient

WHAT THESE RESULTS MEAN

Your child can use rules, models, tables, and graphs to represent and describe patterns. She can create equations to represent problems, use variables to represent unknowns and describe the rate of growth over time as constant or varying.

NEXT STEPS

Tell your child a rule (e.g., multiply by 5) and a number, and have her use the rule to tell you what number the rule would generate. Have her determine her own rule and use words, a table or a graph to show how the rule works.

Data Analysis and Probability

Students read, construct and interpret frequency tables, circle graphs, line graphs, and more complex displays of data. They understand the meaning of range, mean, median, and mode. They use ratios to represent the probability of simple events.

Jane Scored Above Proficient

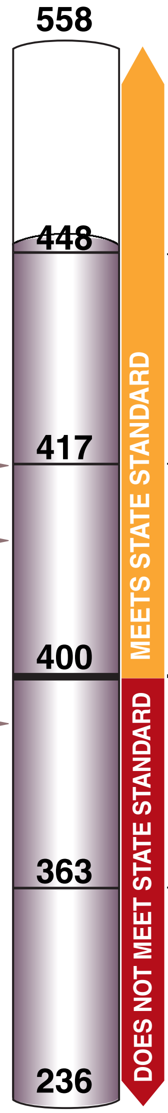
WHAT THESE RESULTS MEAN

Your child can interpret and construct circle graphs and double bar graphs. She can select the appropriate type of graph for a given situation. She can relate concepts of ratio and probability.

NEXT STEPS

Have your child record data, such as the amount of time spent one day in various activities, and then create a circle graph to display the data. Challenge her to relate the size of a section to the number of degrees in a circle.

Science Achievement Results



Jane's score is 452.

She has performed at the advanced level and meets standards for Grade 5 Science.

State Average Score: 417

School Average Score: 411

District Average Score: 392

Advanced - Students apply science facts and recognize relationships. They design investigations, draw conclusions, and apply knowledge to new situations.

Accelerated - Students use science information to discuss patterns; make reasonable predictions; make plans to collect data; and evaluate technological solutions.

Proficient - Students know science ideas and use them to make connections. They plan and carry out investigations and use the results to draw logical conclusions.

Basic - Students know some science facts and may know how they are connected. They may rely on simple ideas and facts to make obvious predictions.

Limited - Students may know few science facts and might have difficulty describing models or relationships. They may struggle to plan science investigations.

Different versions of the test are used each time the test is given. The scores in this report are used to compare performance from year to year. Scores in different subjects cannot be compared.

Has Jane reached proficient in the different areas of Science?

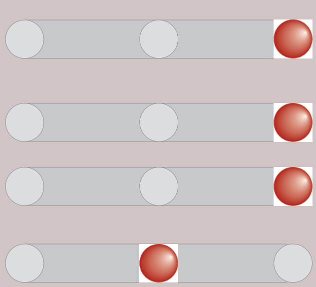
Below Proficient | Near Proficient | Above Proficient

Scientific Processes: Inquiry, Technology & Ways of Knowing

Physical Sciences

Earth & Space Sciences

Life Sciences

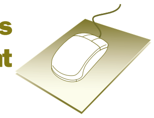


This chart shows you how well Jane performed in each standard. She is above proficient in Scientific Processes, above proficient in Physical Sciences, above proficient in Earth & Space Sciences, and near proficient in Life Sciences.

Note: There is some uncertainty in all test scores. Please see the interpretive guide for more information.

What Are Your Child's Strengths and Weaknesses in Science?

See the released test questions from earlier years' achievement assessments at



www.success.ode.state.oh.us

Scientific Processes: Inquiry, Technology and Ways of Knowing

Jane Scored Above Proficient

Students design and safely conduct experiments. They know how to share their results with others and they use observations to make their own conclusions. Students learn about the effects of technology. They use a process (method) to solve problems.

WHAT THESE RESULTS MEAN

Your child can design experiments and draw conclusions from the results. She can write notes that others can follow and can use a process to solve problems. She understands the effects of technology on human life and the environment.

NEXT STEPS

Discuss with your child how your daily lives are changed by technology. Ask her how human activity changes the environment. Encourage her to think of ways to manage the negative effects of technology.

Physical Sciences

Jane Scored Above Proficient

Students compare properties and movement of objects. They explain physical changes (like melting ice) and chemical changes (like iron rusting). Students trace the flow of electricity through a circuit (like a light bulb attached to a battery by wires).

WHAT THESE RESULTS MEAN

Your child can describe physical changes (like ice melting), chemical changes (like iron rusting), and the properties of light and sound. She can describe how heat is transferred from one object to another. She recognizes how forces cause movement.

NEXT STEPS

Talk with your child about different energy sources (like coal). Help her find out how these sources are used to make electricity or heat. Then, talk about how using different energy sources affects the environment. Have her think of ways to manage these effects.

Earth and Space Sciences

Jane Scored Above Proficient

Students explain Earth's cycles (like seasons) and its movement around the sun. They have knowledge of weather concepts and patterns, and can explain how weather and other processes shape Earth's surface. They describe ways to save natural resources (like trees).

WHAT THESE RESULTS MEAN

Your child understands Earth's cycles (like seasons), and its movement around the sun. She can explain weather patterns and how specific processes (like weathering) shape the Earth's surface. She can describe ways to save natural resources (like trees).

NEXT STEPS

Have your child explain how parts of the landscape (like hills or ponds) were formed. Ask her how the movements of shadows created by the sun are linked to how Earth turns. Ask how things (like changes in the amount of daylight) she sees each month are linked to Earth's movement around the sun.

Life Sciences

Jane Scored Near Proficient

Students compare the life cycles of plants and animals. They know how the structures of plants (e.g., flowers) and body parts of animals (e.g., teeth) help them survive. They explain how energy flows through an ecosystem (like a forest).

WHAT THESE RESULTS MEAN

Your child may be able to describe life cycles of plants or animals. She may have difficulty comparing life cycles. She may be able to tell how survival depends on changes in the environment, and the purpose of plant and animal parts (like leaves and wings).

NEXT STEPS

Have your child describe or draw two life cycles. Ask her to tell you what is the same and what is different about them. Discuss with her how plants and animals obtain food. Have her draw a food chain and ask what eats what and why.

To the Family of Jane Doe

Austin C. Smith

Grade: 10
Date of Birth: 3/25/2000
Student ID: 123456789
School District: AIR District (15)
School: AIR Academy (12345)
Test Date: Spring 2017

Delaware Department of Education

DCAS Science Assessment

Dear Parent/Guardian,

I am pleased to provide you with this report on the progress Austin made in science this past school year. The computer-based Science test is an important component of the Delaware System of Student Assessments (DeSSA), which provides you, your child and your child's teacher with valuable information about his/her learning.

This report also describes what the test covered, ideas for activities that will support continued learning and answers to frequently asked questions. I invite you to review this report with Austin and reach out to his teacher to discuss how you can be involved.

Sincerely,

Susan Bunting
Susan Bunting, Secretary of Education

DCAS Assessment Information

What is the purpose of the DCAS?

The Delaware Comprehensive Assessment System (DCAS) is a statewide standardized testing program tied to the Delaware content standards, which define the knowledge and skills required for our children to succeed beyond high school. DCAS is designed to

- help schools and districts determine whether children are making progress on meeting standards; and
- help the state learn how schools and districts are ensuring that children are meeting the standards.

What do the results of the DCAS mean, and how are they used?

DCAS results summarize a student's abilities as they relate to Delaware content standards. DCAS is one of the many tools used by teachers to help identify each child's strengths and weaknesses so that they can focus their instruction to meet the specific needs of their students.

For help in understanding Austin's scores and this report, contact Austin's teacher or school principal.

What is in this report?

- Austin's Science score
- Suggestions for Next Steps you can take with Austin
- FAQs and additional resources

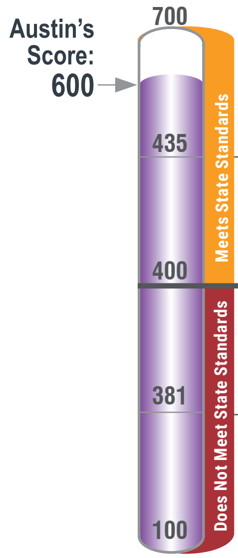
How did Austin do on the **Science Assessment?**

600

How does this compare?

Austin's Science score is 600. This score is **higher** than the average score of tenth graders in his school, **higher** than tenth graders in his district, and **higher** than tenth graders statewide.

	Average Score
State Average	409
District Average	390
School Average	415



Advanced – Students consistently demonstrate understanding of tenth grade Science concepts including properties of materials, energy and its effects, Earth's systems, and life processes.

Meets Standard – Students frequently demonstrate understanding of tenth grade Science concepts including properties of materials, energy and its effects, Earth's systems, and life processes.

Below Standard – Students partially demonstrate understanding of tenth grade Science concepts including properties of materials, energy and its effects, Earth's systems, and life processes.

Well Below Standard – Students minimally demonstrate understanding of tenth grade Science concepts including properties of materials, energy and its effects, Earth's systems, and life processes.

To see the complete grade-specific performance level descriptions and expectations, please view Science PLDs at **http://de.portal.airast.org/resources**.

Science Score Information

Students took DCAS Science once during the spring testing window; it is not tested in the fall testing window. DCAS Science is not tested in the previous grade, so no comparison is available with the previous year.

Science Categories Being Assessed in Grade Ten

Earth Science

Earth Science focuses on the solar system, components of Earth, and interactions throughout Earth's systems that cause earthquakes and volcanoes.

Life Science

Life Science focuses on structure and function relationships, cell structures and processes, matter and energy transformations that occur in photosynthesis and cellular respiration, heredity and genetics, and evolution.

Physical Science

Physical Science focuses on properties and structure of materials including atoms, elements, and compounds; chemical reactions; conservation of matter; forms of energy; energy transfer; and forces.

Next Steps for Austin Based on His Performance This Year

The following activities may help improve your child's knowledge and skills:

- Have your child predict the chemical properties of an element given its location on the Periodic Table. Ask him to predict how this element might interact with another element (e.g., high reactivity, low reactivity, electronegativity) under varying conditions.
- Have your child demonstrate and explain how changes in temperature, pressure, and concentration affect the rates of chemical reactions.
- Have your child illustrate the starting and ending structures of enzyme and substrate molecules in a reaction. Have him explain the "lock and key" enzyme/substrate model.
- Have your child predict the effect of DNA mutations on protein composition and function.
- Have your child explore the present rationales behind the formation of our Solar System by comparing theories of the evolution of the Solar System and Earth's composition with the Solar Nebular Theory.
- Have your child explain how tectonic plate movement affects volcanic and earthquake potential. Ask him to identify locations in the world that have high or low potential for such activity.

This space is reserved for notes.

Resources & FAQs

Are my child's scores good enough?

While it is the goal of the school to have your child achieve at the highest level possible, at a minimum, children are expected to perform at the Meets Standard level in each subject at each grade level.

How are students prepared for online testing?

A practice test is available for families to try out the online testing system at **http://de.portal.airast.org**. To use the practice test, you will need to install a free Mozilla Firefox browser on your home computer. The latest version of Adobe Flash must also be installed on the computer. A link to download the browser can be found on the website with the practice test.

How are accommodations provided for students with disabilities and English language learners?

Most accommodations that are provided to students with disabilities and English language learners in classes are also available for the Delaware System of Student Assessments (DeSSA). Therefore, decisions for your child concerning accommodations that were chosen by ELL, 504, and IEP teams during the school year will continue to be used for the administration of the DeSSA. For more information, please contact your child's school.

Visit http://de.portal.airast.org/
for more information on DE reports.

AIR District (15)
1000 Thomas Jefferson St NW
Washington, DC 20007

PRSRT First-Class
U.S. Postage Paid
Austin, TX
Permit 1143

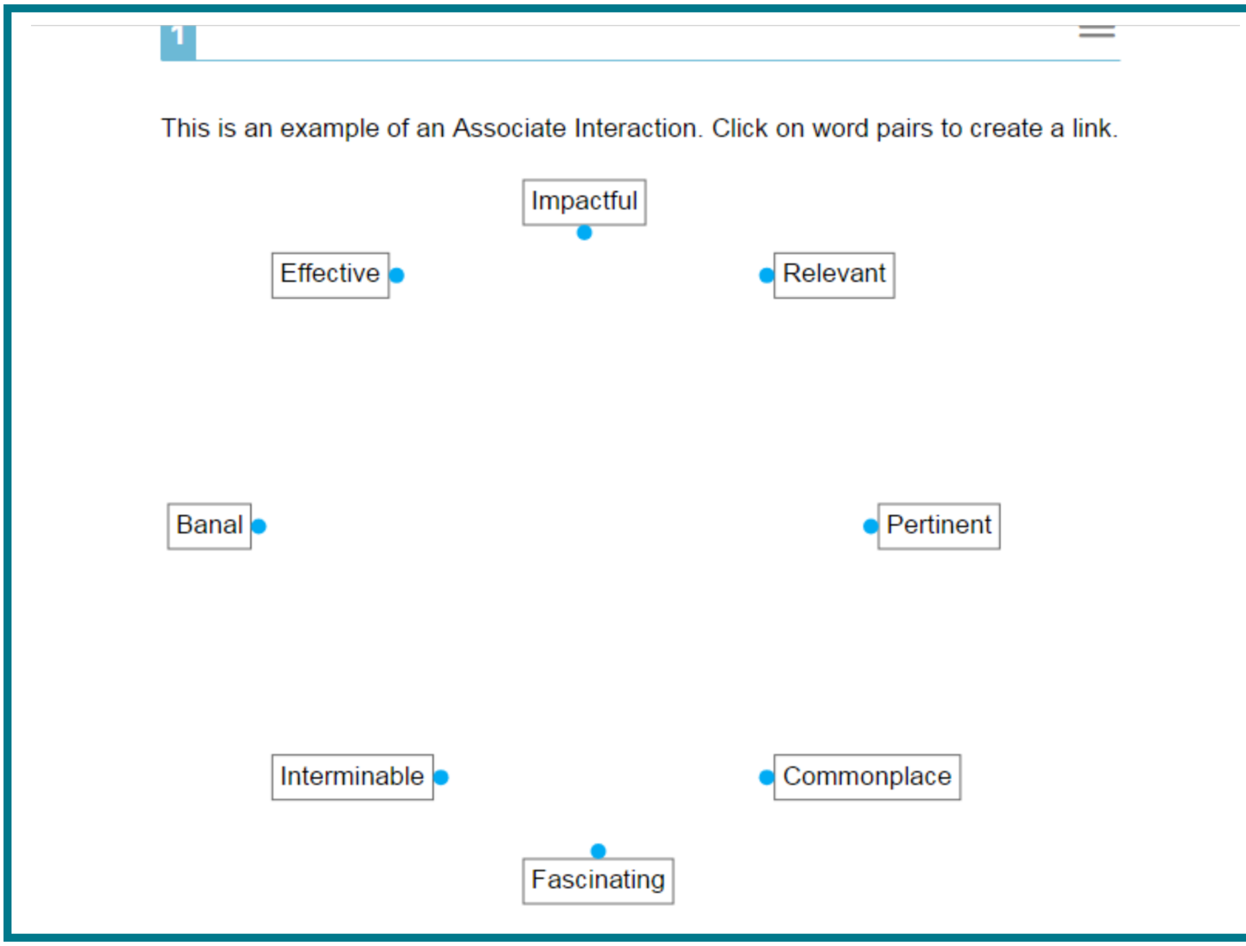
To the Parent/Guardian of Austin C. Smith
1000 Neil Armstrong Highway
PO Box 1120
Dover, DE 19904

15.10
48939390

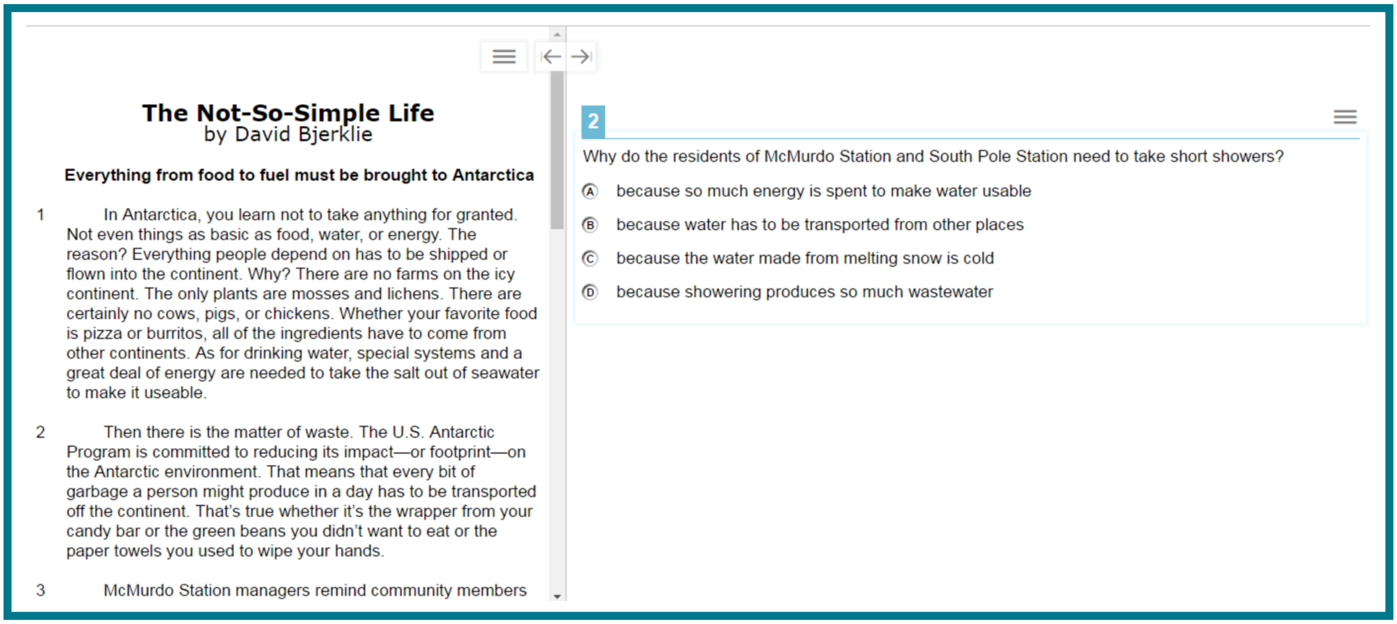
15.10

Appendix Q:
Screen Captures of Interaction Types

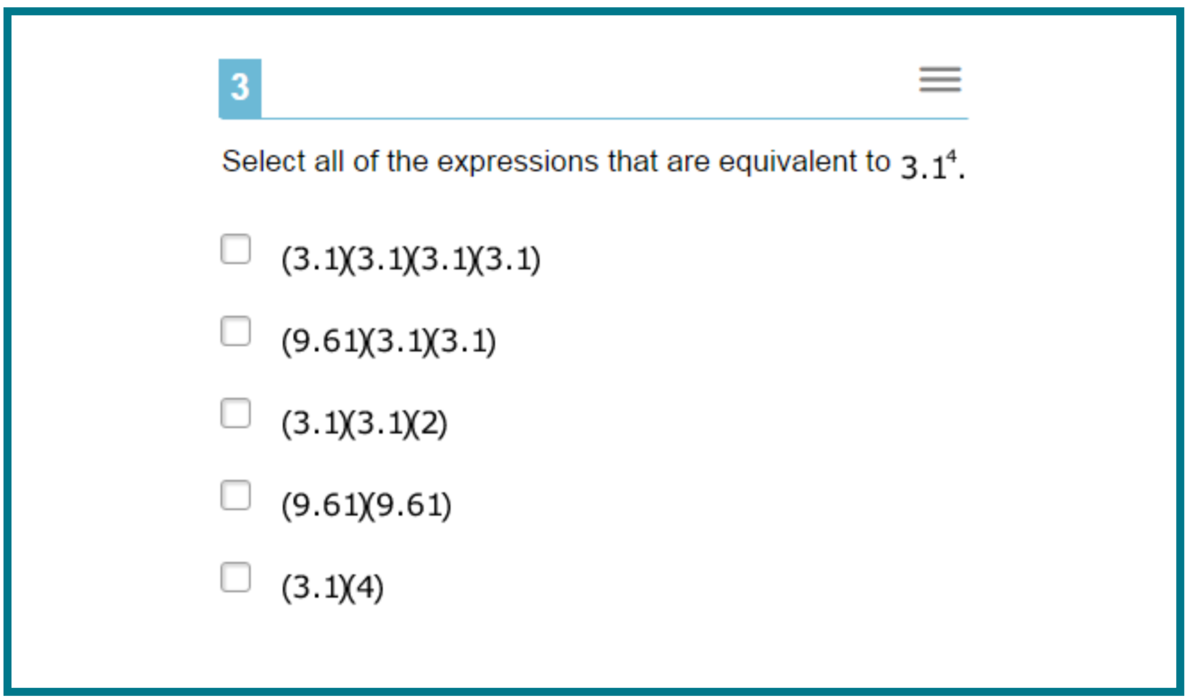
Interaction Type	Description
associateInteraction	A student creates a one-to-one or many-to-one relationship between two pieces of data. Machine-scored.



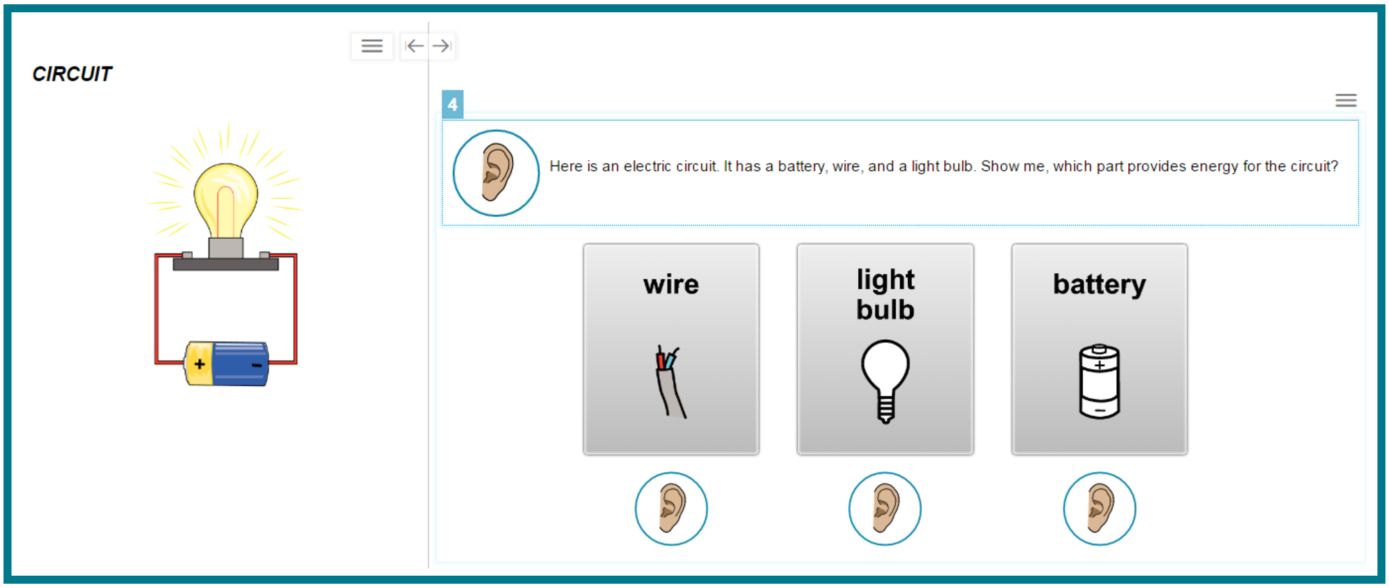
Interaction Type	Description
choiceInteraction:MultipleChoice	A student selects from traditional multiple-choice options using radio buttons. Machine-scored.



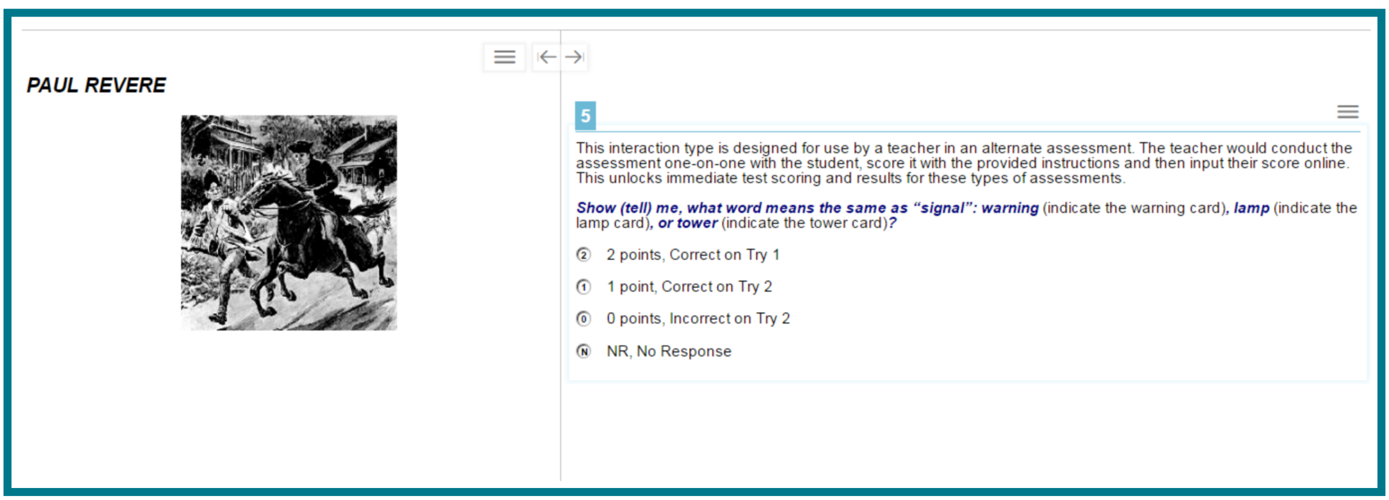
Interaction Type	Description
choiceInteraction:MultipleSelect	This is a traditional multi-select (checkboxes) with flexibility on minimum and maximum number of selections. Machine-scored.



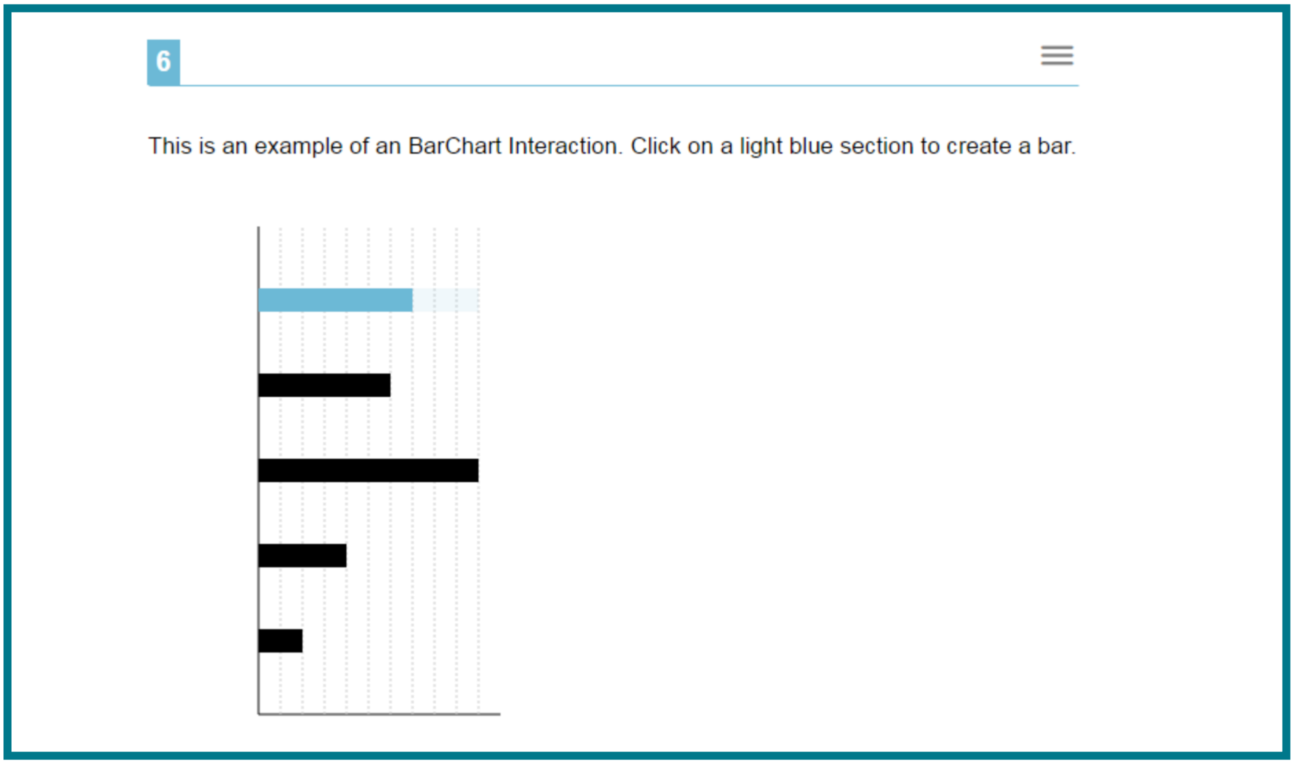
Interaction Type	Description
choiceInteraction:Scaffolding	This is a graphic-rich choiceInteraction variant often used with alternate assessment student populations, which removes an incorrect response and gives the student a second try. Machine-scored.



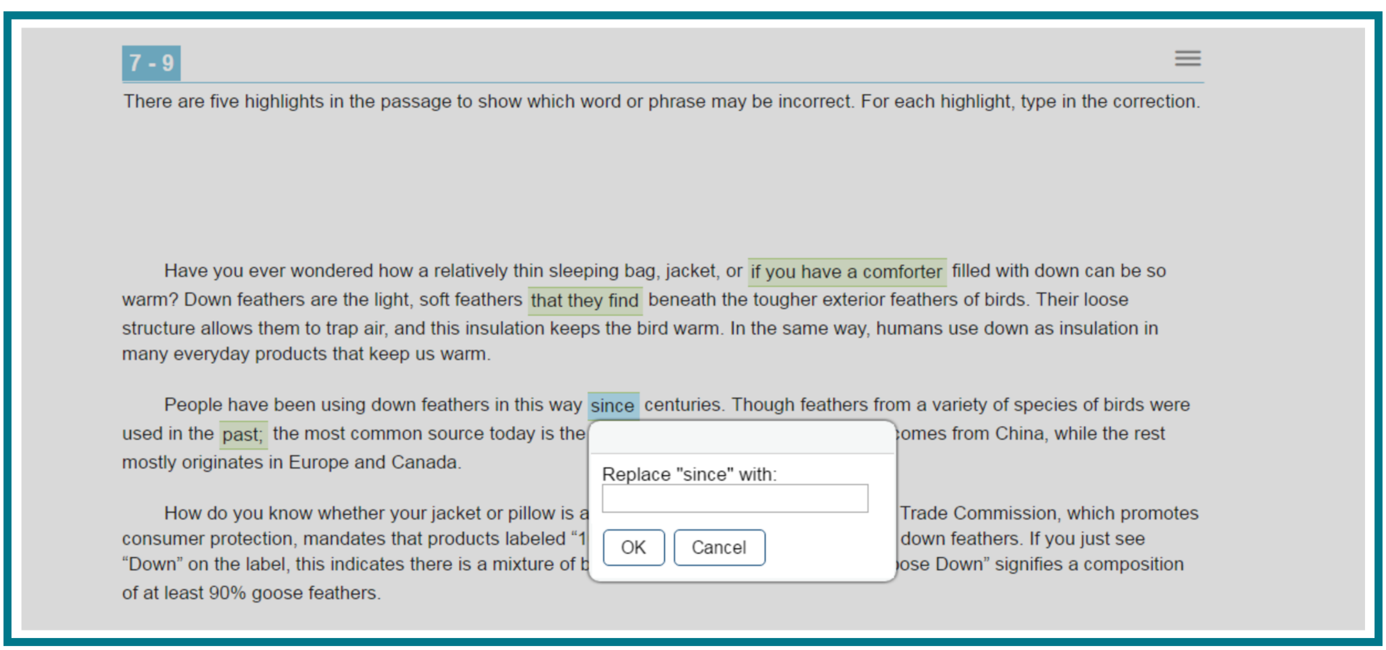
Interaction Type	Description
choiceInteraction:ScoreInput	This is a variation of choiceInteraction that allows a data entry clerk to directly enter the score a student achieved instead of the response.



Interaction Type	Description
customInteraction:BarChart	The student uses simple single-click actions to draw bar charts. Machine-scored.



Interaction Type	Description
customInteraction:EditTask	A student clicks a word and replaces it by typing in another word to revise a sentence. Machine-scored.

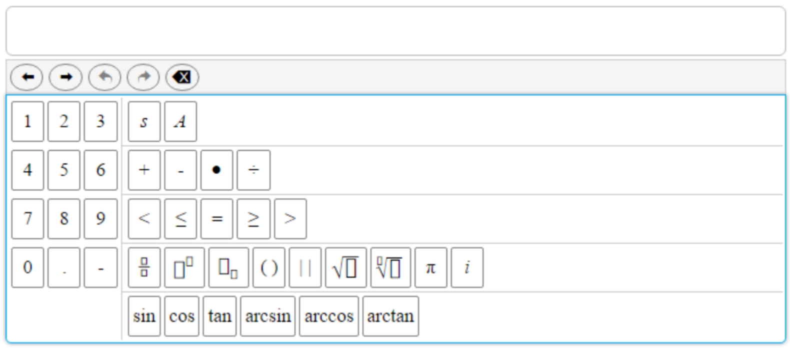


Interaction Type	Description
customInteraction:EquationEditor	A student uses flexible entry of one or more mathematical expressions. Machine-scored.

10

Joshua draws two squares. The area of square M is $\frac{3}{4}$ the area of square N.

Create an equation that relates the side length of square M, s, and the area of square N, A.

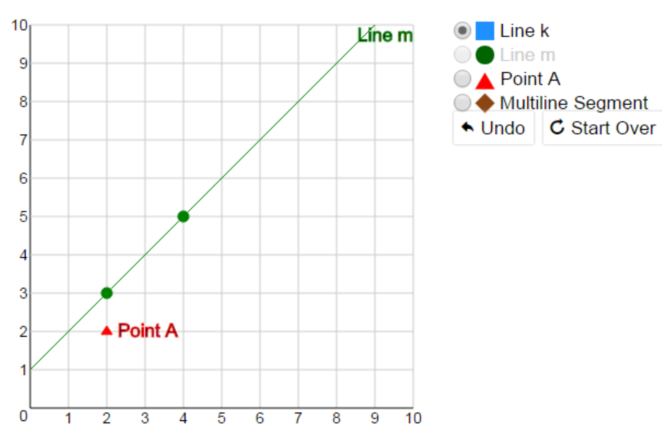


Interaction Type	Description
customInteraction:Graphing	A student graphs simple concepts like rays, line segments, and closed shapes. Machine-scored.

11

This is an example of an Graphing Interaction. Select an object to create, then click on the graph to add and/or move the object.

In the *xy*-plane below, point *A* has the coordinates $(2, 1)$. Draw line *k* through point *A* with slope -4. Then draw line *m* so that it is perpendicular to line *k* and contains the point $(5, 6)$.

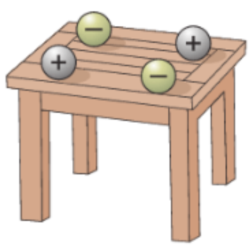


Interaction Type	Description
customInteraction:GraphicResponse	This is a flexible platform to create a wide variety of items from drag-and-drop, to drawing on a coordinate plane, to hot spots. Machine-scored.

Graphic Response (dragging and drawing):

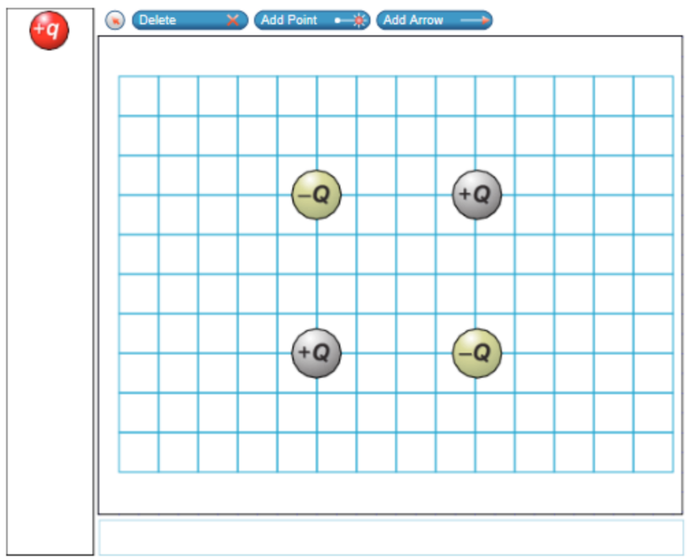
12

A scientist has fastened four charged metal spheres to a laboratory bench. The spheres all have the same amount of charge, Q, but the silver spheres are positively charged while the green spheres are negatively charged.



A. On each sphere, draw an arrow representing the net electric force due to the other three spheres. Ensure that the length of each arrow represents the relative strength of the force.

B. Move the red charged sphere to the point on the table where there will be no net force from the other four spheres.



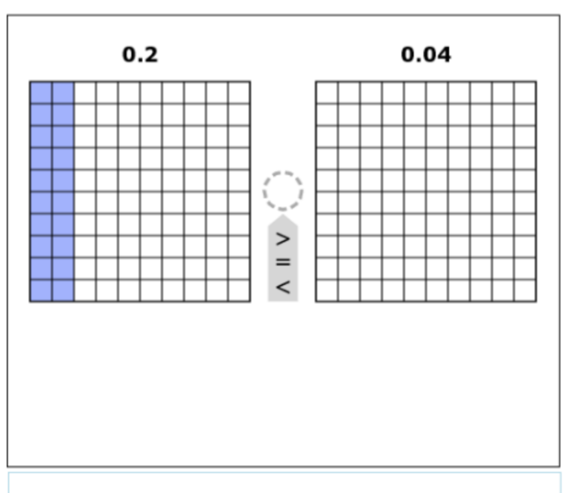
Graphic Response (hot spot):

13

A decimal model is shown.

A. Click on the grid on the right to create a decimal model that shows 0.04.

B. Click on the correct symbol to compare the two decimals.



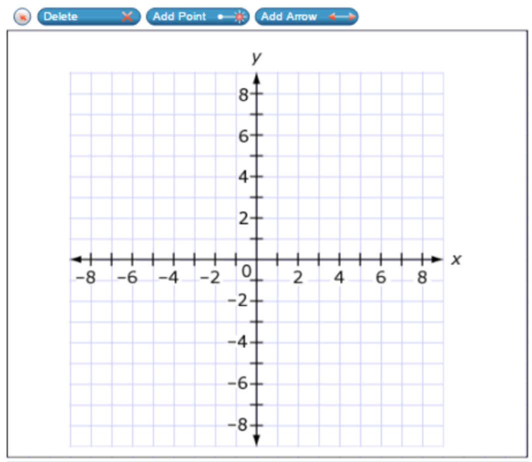
Graphic Response (graphing):

14

A linear function is represented in the table shown.

x	*y*
–1	–6
1	–2
3	2

Use the Add Arrow tool to draw a line on the coordinate grid that has a greater *y*-intercept than the function represented by the table and is perpendicular to the function $y+\frac{1}{4}x=2$.



Hot Spot Interaction:

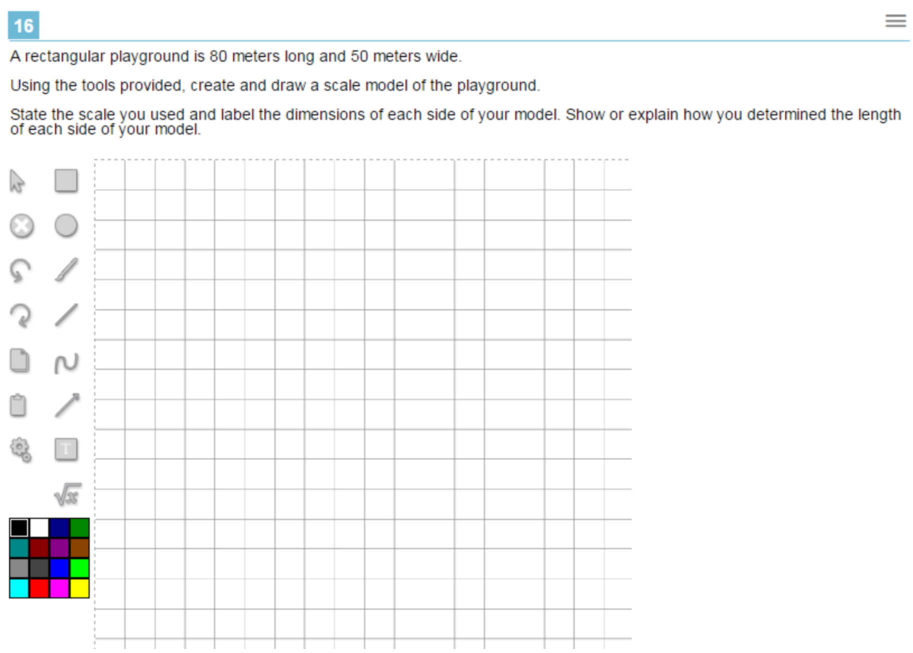
15

15

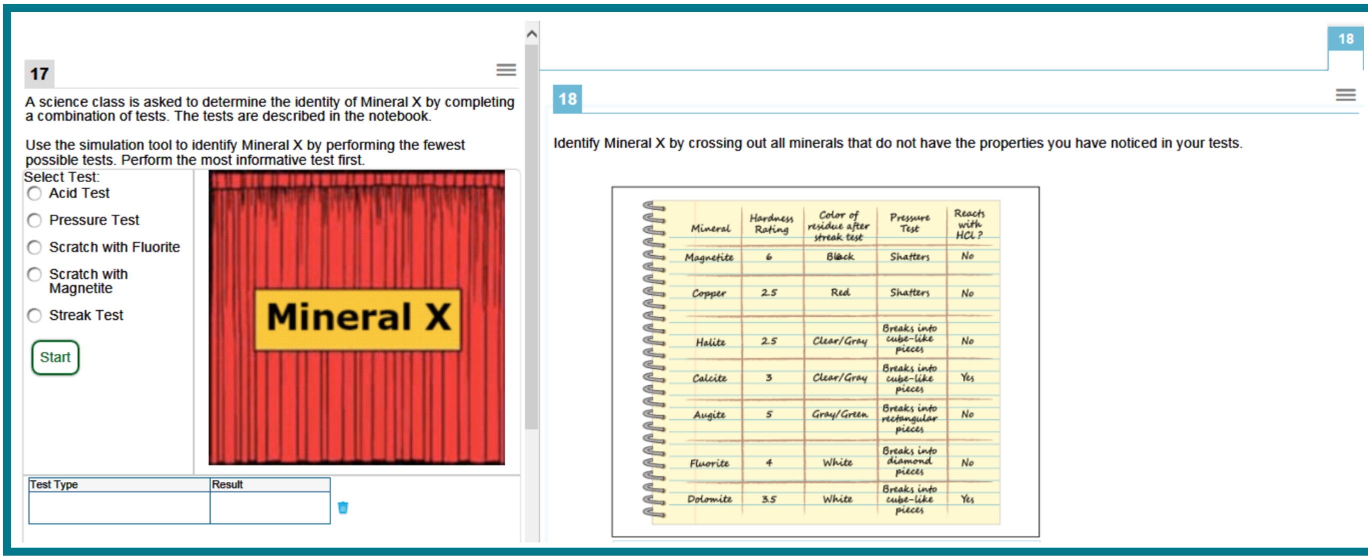
15

15

Interaction Type	Description
customInteraction:Scratchpad	This allows freeform drawing, text, or mathematics entry onto a canvas. Hand-scored.



Interaction Type	Description
customInteraction:Simulation	A student conducts an experiment by manipulating inputs until the desired output is reached. Machine-scored.



Interaction Type	Description
customInteraction:TableInput	This solicits a student to complete tabular data. Machine-scored.

19

The total number of hours, *t*, that Trent has practiced his guitar after *d* days is modeled by the equation shown.

$t = 3d$

Complete the table to describe this relationship.

d	*t*
1	
3	
	15

Interaction Type	Description
customInteraction:VerbalResponse	This records the student's voice using a microphone. Hand-scored.

20

Look at the picture. Answer the question.

Interaction Type	Description
customInteraction:WordBuilder (Numeric and Qwerty)	This is a fill-in-the-blank style response with an onscreen keyboard that shows the allowable keys. Machine-scored.

21

The table shows three days of weather data collected by students.

Recorded Weather Data

	Tuesday	Wednesday	Thursday
Temperature (°F)	34	30	36
Precipitation (inches)	1	3	0
Cloud cover	Partly cloudy	Cloudy	Clear

Identify the form of precipitation that occurred on Wednesday.

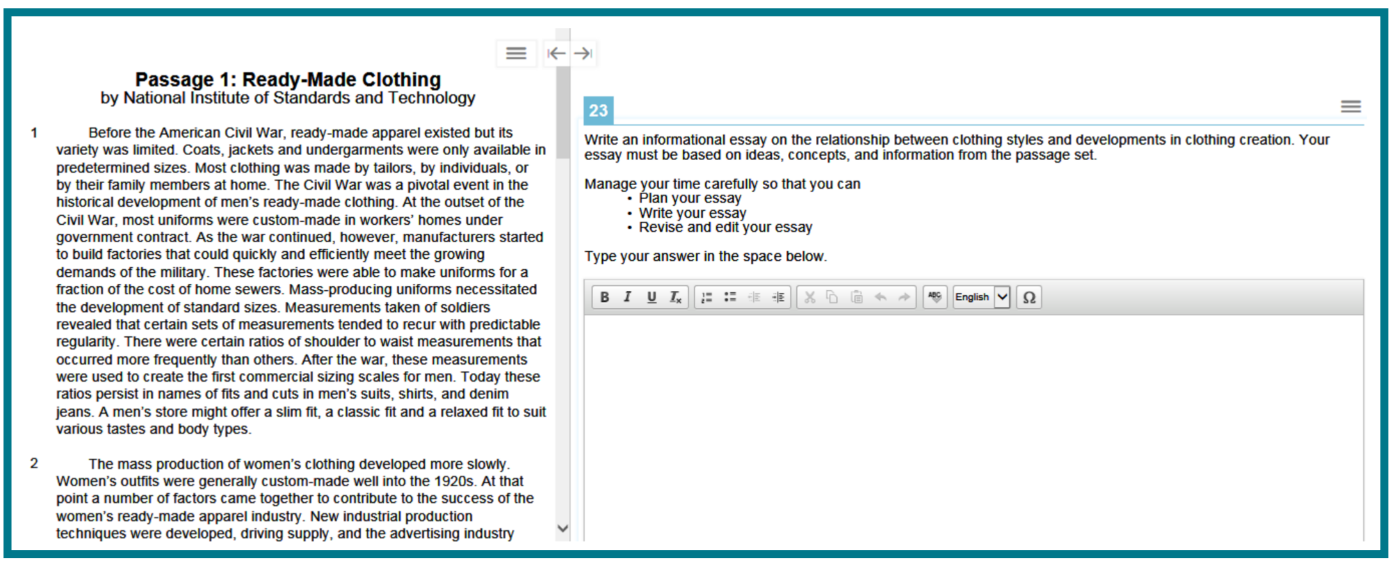
- Type only **one** word in the box.

A	B	C	D	E	F	G	H	I
J	K	L	M	N	O	P	Q	R
S	T	U	V	W	X	Y	Z	

Space Delete

Interaction Type	Description
textEntryInteraction	This is a short answer text response. Hand-scored or machine-scored using natural language scoring engines.

Interaction Type	Description
extendedTextInteraction	This is an extended response for essays. Hand-scored or machine-scored using essay scoring engines.



Interaction Type	Description
gapMatchInteraction	A student drags pre-written text responses into target boxes to respond. Machine-scored.

24

Move the words to make the sentence.

My teacher [ ] [ ] [ ] in my [ ] [ ] day.

write every journal lets me

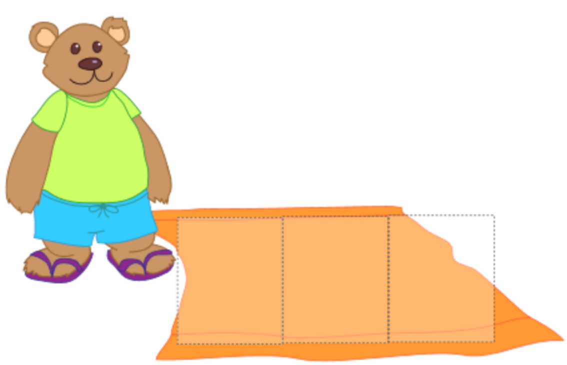
Interaction Type	Description
graphicGapMatchInteraction	A student drags prepared graphic responses into target boxes to respond. Machine-scored.

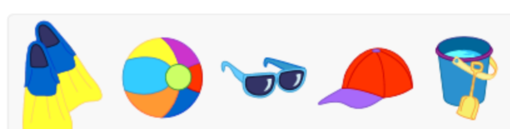
25

Read about Mr. Bear. Move the things Mr. Bear needs to the picture. Follow the directions.

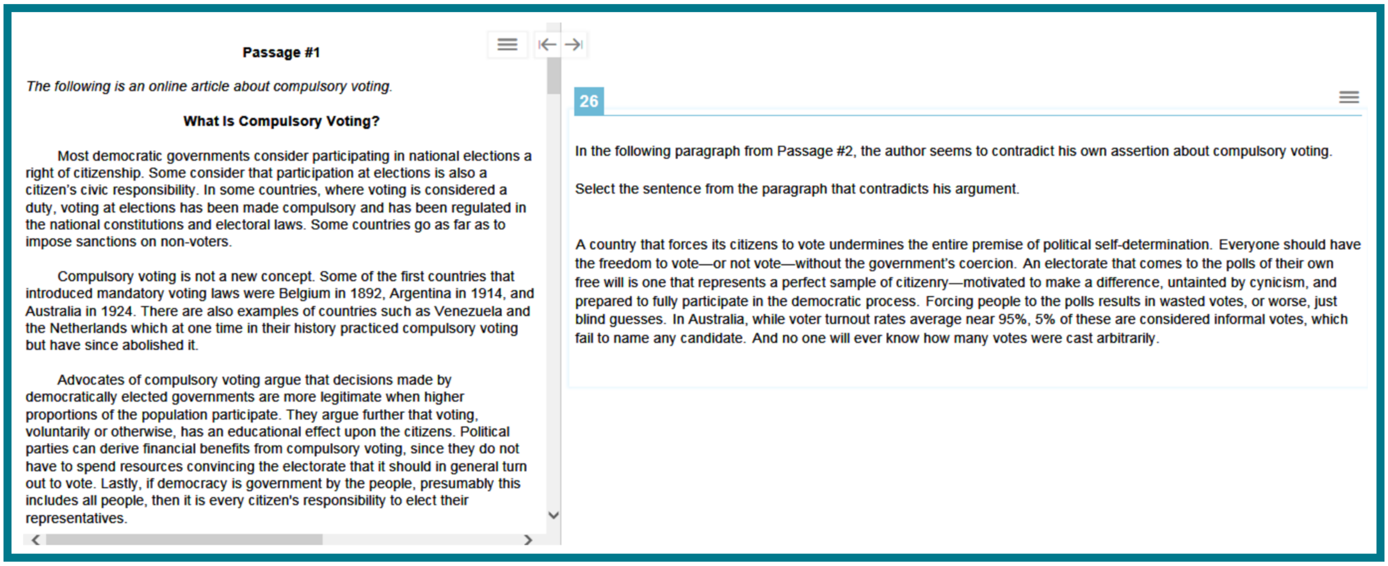
It is hot and sunny outside today. Mr. Bear is going to swim in the pool. He needs his sunglasses, flippers, and a hat.

Move the things Mr. Bear needs onto his towel.





Interaction Type	Description
hotSpotInteraction	A student clicks on clickable regions of a graphic to respond. Machine-scored.



Passage #1

The following is an online article about compulsory voting.

What Is Compulsory Voting?

Most democratic governments consider participating in national elections a right of citizenship. Some consider that participation at elections is also a citizen's civic responsibility. In some countries, where voting is considered a duty, voting at elections has been made compulsory and has been regulated in the national constitutions and electoral laws. Some countries go as far as to impose sanctions on non-voters.

Compulsory voting is not a new concept. Some of the first countries that introduced mandatory voting laws were Belgium in 1892, Argentina in 1914, and Australia in 1924. There are also examples of countries such as Venezuela and the Netherlands which at one time in their history practiced compulsory voting but have since abolished it.

Advocates of compulsory voting argue that decisions made by democratically elected governments are more legitimate when higher proportions of the population participate. They argue further that voting, voluntarily or otherwise, has an educational effect upon the citizens. Political parties can derive financial benefits from compulsory voting, since they do not have to spend resources convincing the electorate that it should in general turn out to vote. Lastly, if democracy is government by the people, presumably this includes all people, then it is every citizen's responsibility to elect their representatives.

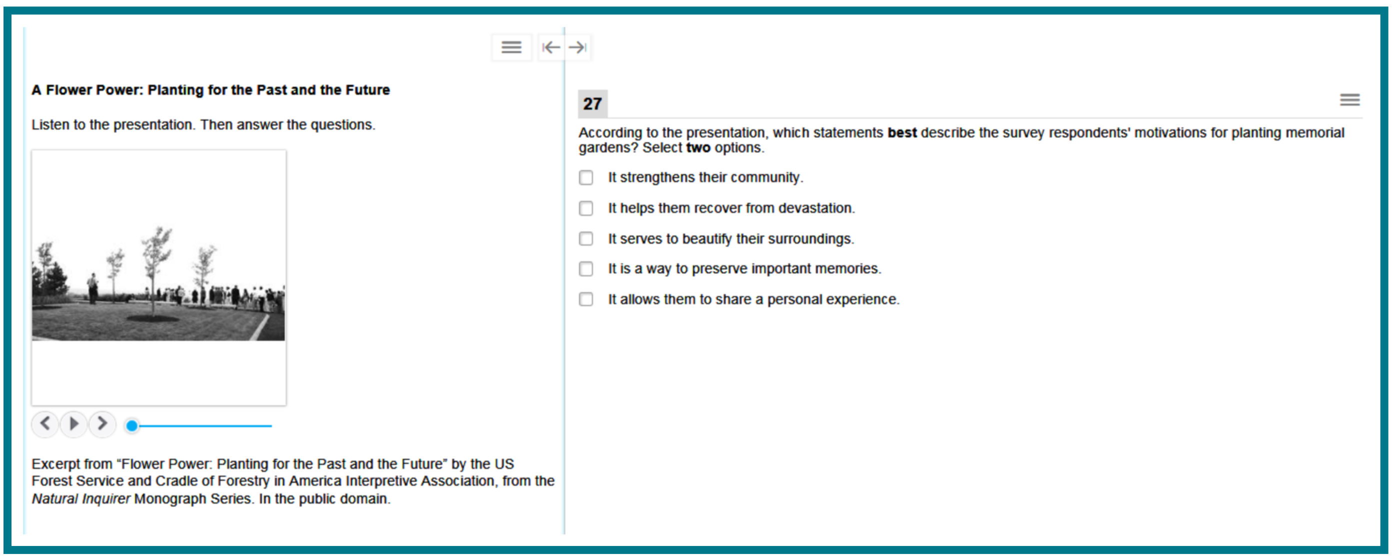
26

In the following paragraph from Passage #2, the author seems to contradict his own assertion about compulsory voting.

Select the sentence from the paragraph that contradicts his argument.

A country that forces its citizens to vote undermines the entire premise of political self-determination. Everyone should have the freedom to vote—or not vote—without the government's coercion. An electorate that comes to the polls of their own free will is one that represents a perfect sample of citizenry—motivated to make a difference, untainted by cynicism, and prepared to fully participate in the democratic process. Forcing people to the polls results in wasted votes, or worse, just blind guesses. In Australia, while voter turnout rates average near 95%, 5% of these are considered informal votes, which fail to name any candidate. And no one will ever know how many votes were cast arbitrarily.

Interaction Type	Description
hottextInteraction	A student clicks a word or sentence to respond. Machine-scored.



A Flower Power: Planting for the Past and the Future

Listen to the presentation. Then answer the questions.

Excerpt from "Flower Power: Planting for the Past and the Future" by the US Forest Service and Cradle of Forestry in America Interpretive Association, from the *Natural Inquirer* Monograph Series. In the public domain.

27

According to the presentation, which statements **best** describe the survey respondents' motivations for planting memorial gardens? Select **two** options.

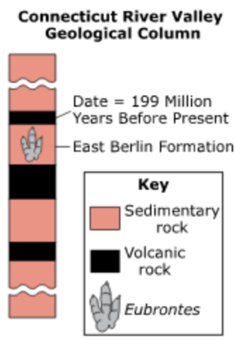
- It strengthens their community.
- It helps them recover from devastation.
- It serves to beautify their surroundings.
- It is a way to preserve important memories.
- It allows them to share a personal experience.

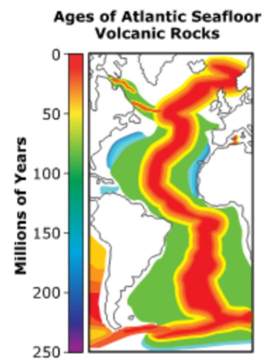
Interaction Type	Description
inlineChoiceInteraction	A student selects responses from drop-down menus throughout the text. Machine-scored.

The following question has two parts. First, answer part A. Then, answer part B.

Sedimentary rocks in the Connecticut River Valley of Connecticut and Massachusetts preserve numerous dinosaur footprints. These footprints have been given the name *Eubrontes*. *Eubrontes* footprints are particularly common in a rock unit called the East Berlin Formation. Volcanic rocks (ancient lava flows) that sit on top of the East Berlin Formation have been dated to about 199 million years ago. In the nearby Atlantic Ocean, the oldest volcanic rocks in the seafloor have been dated to about 180 million years ago.

The diagram shows a geological column for the East Berlin Formation in the Connecticut River Valley. The map shows the geologic ages of volcanic rocks on the Atlantic Ocean seafloor.





Part A

Click on each blank box to select the word that correctly describes the age relationships between the East Berlin Formation and the volcanic rocks of the Connecticut River Valley and the volcanic rocks of the Atlantic Ocean seafloor.

Eubrontes footprints are preserved in the sedimentary rocks of the East Berlin Formation. These sedimentary rocks sit [between ▼] volcanic rocks (ancient lava flows). These volcanic rocks are [younger ▼] than the oldest volcanic rocks of the Atlantic Ocean seafloor.

Part B

Click on each blank box to select the word that correctly describes the age relationship between the dinosaur that made the East Berlin Formation *Eubrontes* footprints and the formation of the Atlantic Ocean.

Based on the evidence provided, the dinosaur that made the East Berlin *Eubrontes* footprints lived [▼] the Atlantic Ocean began to form.

Interaction Type	Description
matchInteraction	A student creates a one-to-one relationship between two elements. Can be represented graphically or as a table. Machine-scored.

29

Select whether each equation has no solution, one solution, or infinitely many solutions.

	No solution	One solution	Infinitely many solutions
$7x+10=7x+10$	☐	☐	☐
$4x=4x+3$	☐	☐	☐
$5x+3=2x-3$	☐	☐	☐
$2x+8=3+2x$	☐	☐	☐

Interaction Type	Description
orderInteraction	A student reorders a list of elements to create the correct sequence. Machine-scored.

30

This is an example of an Order Interaction. Drag the words to put them in the correct order.

Fourth

First

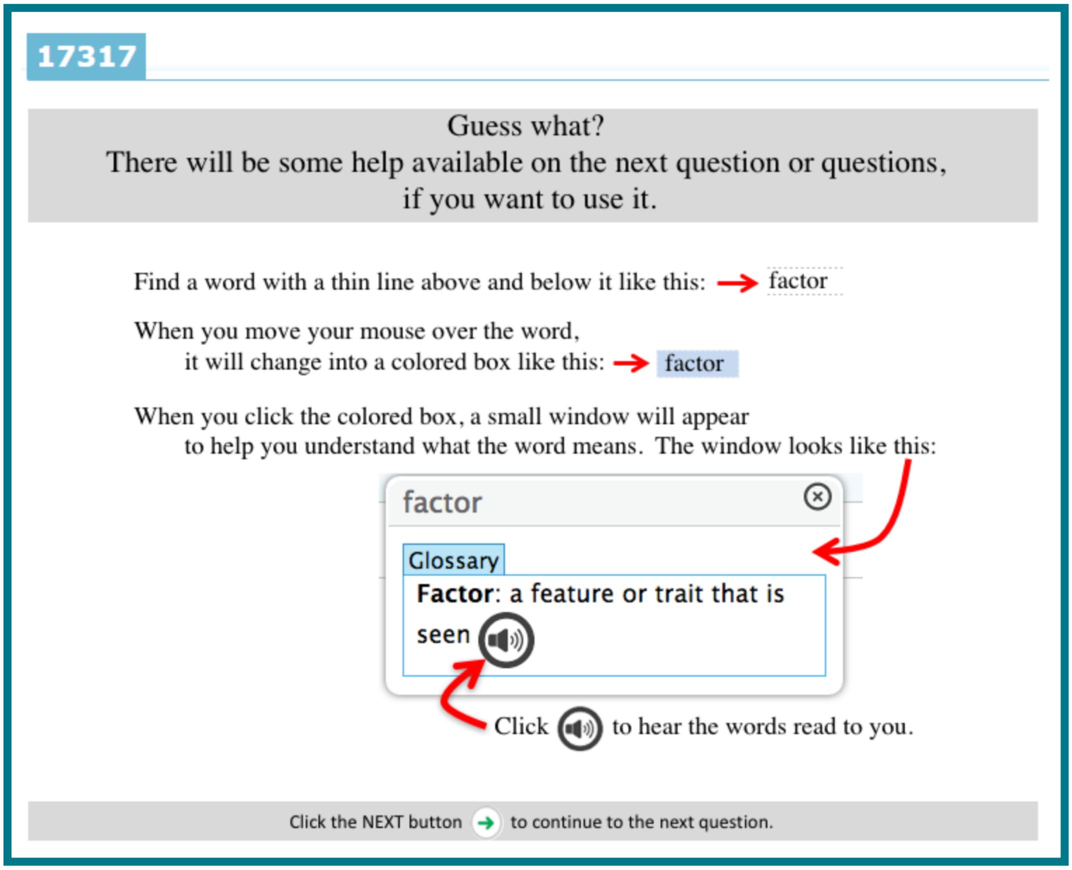
Fifth

Third

Second

Interaction Type	Description
html:GenericInstruction	This is a simple text display usually used for instructions. Not scored.

Student Help Generic Instruction:



Student Help Checklist (writing guide):

Informative/Explanatory Essay Writing Guide (Grades 6-8)

Purpose, Focus, and Organization	Evidence and Elaboration	Conventions
The response is fully sustained and consistently focused within the purpose, audience, and task; and it has a clear controlling idea and effective organizational structure creating coherence and completeness. The response includes most of the following: • Strongly maintained controlling idea with little or no loosely related material • Skillful use of a variety of transitional strategies to clarify the relationships between and among ideas • Logical progression of ideas from beginning to end with a satisfying introduction and conclusion • Appropriate style and objective tone established and maintained	The response provides thorough and convincing support, citing evidence for the controlling idea or main idea that includes the effective use of sources, facts, and details. The response includes most of the following: • Smoothly integrated, thorough, and relevant evidence, including precise references to sources • Effective use of a variety of elaborative techniques (including but not limited to definitions, quotations, and examples), demonstrating an understanding of the topic and text • Clear and effective expression of ideas, using precise language • Academic and domain-specific vocabulary clearly appropriate for the audience and purpose • Varied sentence structure, demonstrating language facility	The response demonstrates an adequate command of basic conventions. The response may include the following: • Some minor errors in usage but no patterns of errors • Adequate use of punctuation, capitalization, sentence formation, and spelling

References and Citations
When referring to evidence and information from passages, students should use paraphrasing and short quotations. To credit sources, students should use informal, in-text citations (e.g., MLA author or title tags).

Writing Test Scoring Guide:

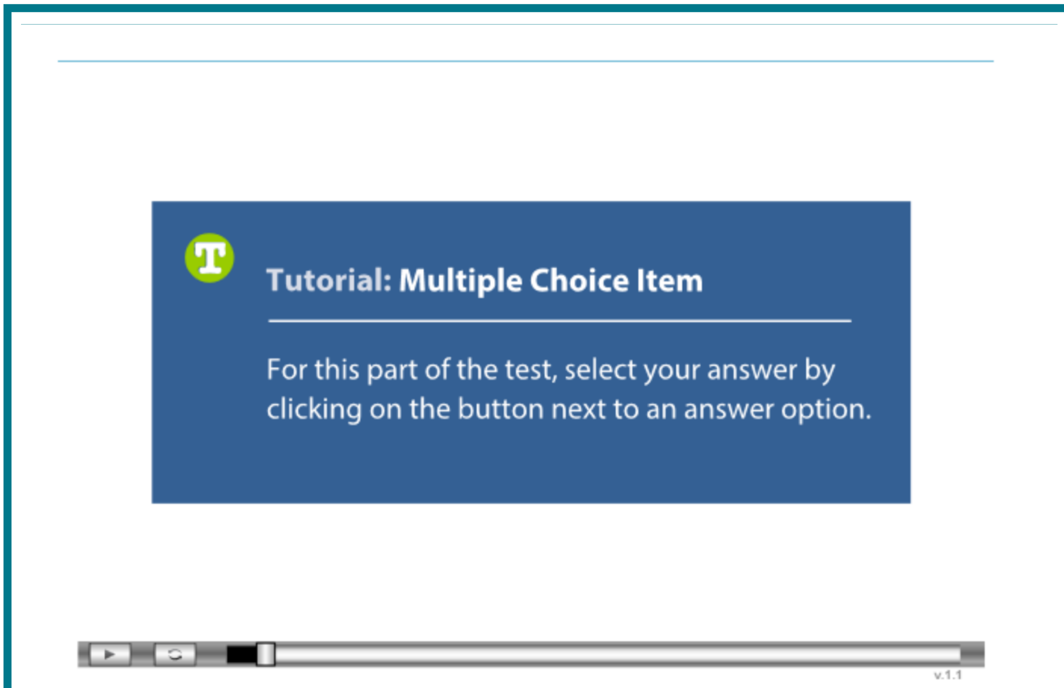
Click this icon - Scoring Guide - to get the guide below:

Scoring Guide

Argumentative Essay Writing Rubric Criteria (Grades 6–8)

Statement of Purpose/Focus and Organization Weight: 40%	Evidence/Elaboration Weight: 40%	Editing/Conventions Weight: 20%
The essay is fully developed and focused: • Claim is clearly stated, focused, and strongly maintained. • Claim is introduced and communicated appropriately for the purpose, audience, and task. • Opposing	The essay provides thorough and convincing support/evidence for the writer's claim and includes the effective use of sources, facts, and details: • Claims are supported with clear reasoning and relevant evidence from credible sources. • Use of evidence	The essay displays adequate command of all grade-level conventions of writing: • Some errors in usage and sentence formation may be present, but no systematic pattern of errors is displayed. • Use of punctuation, capitalization

Interaction Type	Description
html:Tutorial	This is animation that describes how to use that type of interaction. Not scored.



Appendix R:
Microsoft and Linux Security Patching in AIR's System Environments

Microsoft and Linux Security Patching in AIR's System Environments

This document outlines the process of installing the monthly security patches released by Microsoft and the office of Cybersecurity and Communications at the U.S. Department of Homeland Security (for Linux patches) in AIR's dedicated and cloud environments. Both organizations regularly release their respective security bulletins indicating security vulnerabilities they have found in their systems and the severity the risk poses if not patched.

In this document, we describe the process we follow to ensure that critical patches are regularly applied to our systems. But at the same time, AIR fully recognizes that we need our applications to be stable and reliable; thus, this policy also aims to minimize server downtime and risk to the functionality of our applications.

How does AIR determine what patches to apply?

AIR's hosting provider, Rackspace, has the following **Windows Update** policy for their dedicated hosting environment.

"**What patches are approved**:

- Updates announced on each month's security bulletin: <http://www.microsoft.com/technet/security/current.aspx>.
- We approve updates that apply to the Windows OS itself. This includes anything that is installed by default or is an available add-on role/service (i.e. Internet Explorer or IIS security updates).
- We approve security updates for the .NET Framework.
- We can approve non-security updates if support determines a specific need that applies to our customer base as a whole, (i.e. cumulative time zone updates).
- We approve the latest definitions for Windows Defender[1] each week during monthly patching.

What patches are not approved:
The following updates will not be approved because they have the potential of harming customized environments when not thoroughly tested:

- OS service packs
- New Internet Explorer versions
- Non-security related .NET Framework updates -- support only installs .NET Framework releases and service packs on customer request
- Microsoft Office updates -- Customers use Microsoft Office both as a client application on servers and also use its libraries in some web applications. We cannot update this software without the potential for breaking a custom application.

[1] This is Microsoft's anti-malware and anti-spyware software.

- Any add-on Microsoft products other than the core OS, and any third-party software

The customer can install any software and updates they wish by visiting the Windows Update website or downloading the software directly from Microsoft."

Regarding Windows Updates in the Rackspace cloud hosting environment:

- Rackspace does not have a centralized patching system on their cloud hosting environment. AIR uses the local Windows Update feature on each of the servers to download and install security patches.

Regarding Linux Updates in the Rackspace dedicated and cloud hosting environment:

- For both the dedicated and cloud hosting environments at Rackspace, AIR employs scheduled Cron[2] jobs on each of our Linux-based servers to download and install security patches from https://cve.mitre.org. This website is sponsored by the office of Cybersecurity and Communications at the U.S. Department of Homeland Security and provides a comprehensive list of all publicly known security vulnerabilities.
- The patch schedule for our Linux based systems will run concurrently with the Windows-based servers unless a particularly severe vulnerability, such as Heartbleed, is identified. We will then work to patch the affected systems as soon as agreed upon schedule is reached with our clients.

What is the schedule for making these patches?

On the second Tuesday of each month, regularly referred to as "Patch Tuesday", Microsoft releases a series of patches/bug fixes to plug security holes and ensure the stability of its operating systems and application suites. The patch schedule for our Linux based systems will run concurrently with the Windows-based servers.

We have distinct environments across our development, test and production systems, with a different release schedule for each. Refer to **Figure 1** for an illustrated schedule.

<Month> <Year>						
Sun	Mo	Tu	We	Th	Fr	Sa
		Patch Tuesday *(c/o Microsoft)*	**Apply patch to our dev/test** *(c/o AIR)*			
			Apply patch to TDS Reference System *(c/o AIR)*			
Apply patch to Production *(c/o AIR and Rackspace)*	**All systems ready with new patches**					

Figure 1: Monthly Patching Schedule

[2] This is a time-based job scheduler.

1. In our development and test environments, security patches are applied to our development and test servers the second Wednesday of each month. In other words, the day after Patch Tuesday.
2. Verification takes place on our development and test environments.
3. If there are no issues found, patching on our production systems at Rackspace occurs the second Saturday following Patch Tuesday.

How do we verify these patches do not break anything?

Before we patch our production systems, there is a series of checks to ensure that we do not push something to our production machines that will affect system functionality or stability:

1. At Rackspace patches are downloaded on Patch Tuesday of each month and run through a limited testing review by the WSUS Rackspace engineers. After review, they are made available to their clients for install on a pre-arranged schedule.
2. Patches are installed in our development and test environments which mirror our production environment on the second Wednesday of the month. If vulnerabilities arise, we will see them there first and resolve them before deploying to production.
3. Specifically for our Test Delivery System, patches are also installed on what we term a "Reference Cloud System" at Rackspace that is a small scale test version of our Production system also on the second Wednesday of each month. Again, if problems are seen in this environment, we will be able to address them before releasing the patches across the entirety our production environment.
4. Assuming we do not see any issues during Steps 1-3, patches are pushed to the rest of our production systems on the Saturday after patching occurs on our Reference System.
5. Linux patches are installed first on our development servers in our cloud environment before being pushed out to Production.
6. If we encounter issues during Steps 1-5, depending on the severity, we may suspend the patching process until we resolve the issue. We can either wait until the next regular round of patches, unless the security vulnerability that will be patched is considered critical. We will then work with the clients to patch this as soon as possible.

How long does applying these patches take?

The patches are installed beginning at 12 midnight EST on the second Saturday following Patch Tuesday. Once patching begins it will take anywhere from 5-15 minutes to complete the installation.

Servers are then rebooted and should be back up and functioning approximately 30 minutes after the patching process begins.

What post-installation testing and system verification does AIR do?

The verification of the system/application integrity will begin the following morning after the patches have been installed.

Two (2) ASCM technicians will be available to respond to any system outages or application failures that may arise from the patching. This is done by reviewing any tickets that are generated from the Rackspace-provided SCOM monitoring system along with any alerts generated by the AIR-deployed Idera SQL Diagnostic tool, as well as PRTG.

Application level functionality will also be verified using internal monitoring tools.

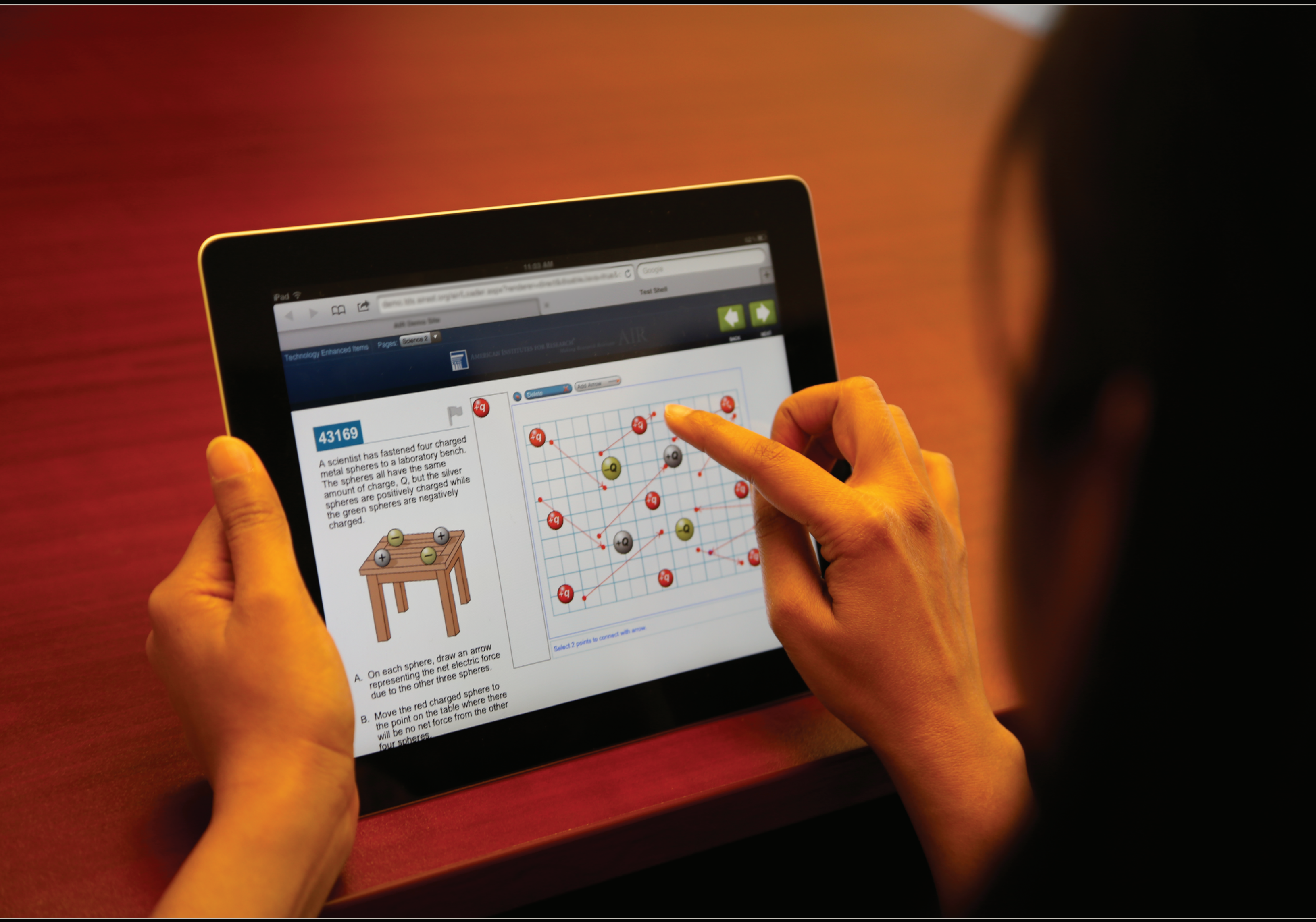
If it is determined that a security patch has broken application functionality, we will conduct an immediate rollback and un-install the update on all affected machines.

Appendix S:
State Staff Resource Worksheet

Team	Name	Assigned Responsibilities	Year 1 Initiation	Year 1 Configuration	Year 1 Implementation	Year 1 Close Out	Year 1 Total	Year 1 Other Projects	Years 2-E Initiation	Years 2-End Configuration	Years 2-End Implementation	Years 2-End Close Out Final Year Only	Years 2-End Total	Years 2-End Other Projects
Project Management	Heather Hayes	Vice President					As needed							
	Tom Glorfield	Program Manager	200	40	960		1,200	600			1,080	120	1,200	600
		Program Coordinator and Assistants	240	120	840		1,200				856	120	976	
	Cindy Benis	Call Center Manager		8	112		120	1,500			120		120	1,500
		Customer Service Staff			5,250		5,250				5,250		5,250	
Psychometrics	Gary Phillips	Vice President and Chief Psychometrician					As needed						As needed	
	Ahmet Turan	Lead Psychometrician			400		400	800			120	80	200	1,000
		Psychometricians			800		800				520	80	600	
		Research Associates/Assistants			800		800				620	80	700	
Test Development	June Zack	Vice President					As needed						As needed	
	Jacob Wilkes	Content Development Lead and ELA Lead		40	200		240	1,360			160		160	1,440
	Alyssa Kartee	Content Specific Area Lead, Math		40	200		240	1,360			160		160	1,440
	Rachel Aazzerah	Content Specific Area Lead, Science		40	120		160	1,440			120		120	1,480
	Matthew Greathouse	Special Populations Consultant			80		80	1,520			60		60	1,540
		Test Development Specialists			600		600				600		600	
	Kara Prezocki	Production Manager			80		80				80		80	
		Publication Staff, including Production, Graphics, and Editorial			200		200				120		120	
Information Technology	Selina Tolosa	Vice President					As needed						As needed	
	Sonja Hubbard	Technology Consultant	40	120	740		900	500			720	180	900	500
Test Delivery	Scott Christian Redmond, M.S., PMP	Software Project Manager	16	24	20		60	1,340			60		60	1,340
	Alan Reeve, M.S.	Senior Application Developer	16	24	20		60	1,340			60		60	1,340
Student Registration	Nikhil Polepally, B.S., PMP	Software Project Manager	16	24	20		60	1,340			60		60	1,340
	Eyal Moses, Ph.D.	Senior Application Developer	16	24	20		60	1,340			60		60	1,340
Item Banking	Bruce Malone, M.B.A.	Software Project Manager	16	24	20		60	1,340			60		60	1,340
	Teddy Xiong, M.S.	Senior Application Developer	16	24	20		60	1,340			60		60	1,340
Analysis and DoR	Muhammad Salman Aslam, Ph.D., PMP	Software Project Manager	16	24	20		60	1,340			60		60	1,340
	Adam McLaughlin, B.A.	Senior Application Developer	16	24	20		60	1,340			60		60	1,340
Online Reporting	Kushal Budhwar, M.S.	Software Project Manager	16	24	20		60	1,340			60		60	1,340
	Deepak Sharma, M.S.	Senior Application Developer	16	24	20		60	1,340			60		60	1,340
		Lead Developers and Data Analysts, All Systems	80	800	1,520		2,400				1,900		1,900	
Software QA	Dorian Sofiaj, M.S.	Lead, Software Quality Assurance			80		80	1,320			80		80	1,320
		QA Leads and Staff			240		240				190		190	
Network Engineering	Fara Tapscott, B.S.	Lead, Network Engineering			80		80	1,320			80		80	1,320
		Network Engineers			420		420				420		420	
Operations and Scoring	Damon Hartzler	Performance Scoring Lead	40	120	240		400	1,000			400		400	1,000
		Scoring Leads and Scorers			1,260		1,260				1,260		1,260	
Score Reporting	Dina Booher	Score Reporting Lead		80	120		200	1,400			80		80	1,520
		Score Reporting Specialists		80	200		280				160		160	

AIR
AMERICAN INSTITUTES FOR RESEARCH

Making Research Relevant



Cost Proposal Copy

State of New Hampshire
Department of Education
c/o Sandie MacDonald
101 Pleasant Street
Concord, New Hampshire 03301

RFP 2017-073 DOE
New Hampshire Statewide Assessments
April 26, 2017

New Hampshire Statewide Assessments: ELA, Mathematics, Science

1000 Thomas Jefferson Street NW, Washington, DC 20007

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

STATE OF NEW HAMPSHIRE
Department of Education

RESPONSE TO PROPOSAL: RFP 2017-073
DOE New Hampshire Statewide Assessments

COST PROPOSAL

American Institutes for Research®
1000 Thomas Jefferson Street, NW
Washington, DC 20007-3835
Contact: Jon Cohen, President, AIR Assessment
Phone: 202-403-5420
Fax: 202-403-5303
Email: JCohen@air.org

April 26, 2017

COPY

F-1 Activities/Deliverables/Milestones Pricing Worksheet – Deliverables List

The vendor must include, within the not-to-exceed for IT service activities, tasks and preparation of required deliverables, pricing for the deliverables required based on the proposed approach, and methodology and tools. The following format must be used to provide this information. Please note: this information is for Year 1. **Please add row**s for activities, deliverables, or milestones not included in the table below. **Please add columns** if specific on-going costs, not covered by costs in Tables F4 and F5, should be included.

Table F-1: Activities/Deliverables/Milestones Pricing Worksheet

		Required						
	Activity, Deliverable, or Milestone	Deliverable Type	Projected Delivery Date	Projected Payment Date	Projected Year 1 Price	Projected Year 2 Price	Projected Year 3 Price	Projected Year 4 Price
PLANNING AND PROJECT MANAGEMENT								
1	Conduct Project Kickoff Meeting	Non-Software	July	September	26,153	26,153	26,153	26,153
2	Project Status Reports	Written	Monthly	Quarterly	212,000	212,000	212,000	212,000
3	Work Plan	Written	July	September	7,500	7,500	7,500	7,500
4	Infrastructure Plan, including Desktop and Network Configuration Requirements	Written	July	September	1,000	-	-	-
5	Security Plan	Written	July	September	1,000	-	-	-
6	Communications and Change Management Plan	Written	July	September	1,000	-	-	-
7	Requirements Trace Ability Matrix	Written	July	September	1,000	-	-	-
8	Software Configuration Plan	Written	July	September	1,000	-	-	-
9	Systems Interface Plan and Design/Capability	Written	July	September	7,500	7,500	7,500	7,500
10	Testing Plan	Written	July	September	1,000	-	-	-
11	Data Conversion Plan and Design	Written	July	September	1,000	-	-	-
12	Deployment and Roll-out Plan	Written	July	September	1,000	-	-	-
13	Comprehensive Training Plan and Curriculum	Written	July	September	12,139	12,139	12,139	12,139
14	End User Support Plan	Written	July	September	1,000	-	-	-
15	Business Continuity Plan	Written	July	September	1,000	-	-	-

		Required						
	Activity, Deliverable, or Milestone	**Deliverable Type**	**Projected Delivery Date**	**Projected Payment Date**	**Projected Year 1 Price**	**Projected Year 2 Price**	**Projected Year 3 Price**	**Projected Year 4 Price**
16	Documentation of Operational Procedures	Written	July	September	1,000	-	-	-
INSTALLATION								
17	Provide Software Licenses, if needed	Written	included in fees below					
18	Provide Fully Tested Data Conversion Software	Software	n/a	n/a	n/a	n/a	n/a	n/a
19	Provide Software Installed, Configured, and Operational to Satisfy State Requirements	Software	n/a	n/a	n/a	n/a	n/a	n/a
TESTING								
20	Conduct Integration Testing	Non-Software	included in fees below					
21	Conduct User Acceptance Testing	Non-Software	included in fees below					
22	Perform Production Tests	Non-Software	included in fees below					
23	Test In-Bound and Out-Bound Interfaces	Software	n/a	n/a	n/a	n/a	n/a	n/a
24	Conduct System Performance (Load/Stress) Testing	Non-Software	included in fees below					
25	Certification of 3rd Party Pen Testing and Application Vulnerability Scanning	Non-Software	included in fees below					
SYSTEM DEPLOYMENT								
26	Converted Data Loaded into Production Environment	*Software*	n/a	n/a	n/a	n/a	n/a	n/a
27	Provide Tools for Backup and Recovery of all Applications and Data	*Software*	n/a	n/a	n/a	n/a	n/a	n/a
28	Conduct Training	*Non-Software*	October	December	86,530	86,530	86,530	86,530

		Required						
	Activity, Deliverable, or Milestone	Deliverable Type	Projected Delivery Date	Projected Payment Date	Projected Year 1 Price	Projected Year 2 Price	Projected Year 3 Price	Projected Year 4 Price
29	Cutover to New Software	*Non-Software*	n/a	n/a	n/a	n/a	n/a	n/a
30	NA	n/a	n/a	n/a	n/a	n/a	n/a	n/a
31	Provide Documentation	*Written*	included in fees below					
32	Execute Security Plan	*Non-Software*	included in fees below					
OPERATIONS								
33	Ongoing Hosting Support	Non-Software	included in fees below					
34	Ongoing Support & Maintenance	Software	included in fees below					
35	Item Banks and Configuration for Testing (UAT)	Non-Software	July	September	207,902	207,902	207,902	207,902
36	Test Administration	Non-Software	May	June	808,500	808,500	808,500	808,500
37	Test Scoring	Non-Software	June	June	150,000	150,000	150,000	150,000
38	Student Data File	Non-Software	June	July	157,500	157,500	157,500	157,500
39	Online Reporting System	Non-Software	November	December	95,332	95,332	95,332	95,332
40	Deliver Individual Student Reports (ISR)	Non-Software	June	July	134,878	134,878	134,878	134,878
41	TAC Meeting (2)	Non-Software	January / June	March / June	113,101	113,101	113,101	113,101
42	Technical Report		June	June	50,000	50,000	50,000	50,000
43	Conduct Project Exit Meeting	Non-Software			n/a	n/a	n/a	12,000
ITEM DEVELOPMENT and TEST CONSTRUCTION								
44	Test Design	Non-Software	October	December	44,958	n/a	n/a	n/a
45	Summative Test Construction	Non-Software	January	March	75,019	30,000	30,000	30,000
46	Interim Form Construction	Non-Software	January	March	31,115	5,000	5,000	5,000
47	Practice Tests	Non-Software	October	December	42,039	n/a	n/a	n/a
#REF!	Bias/Sensitivity Committee Review	Non-Software	August	September	70,054	42,830	42,830	42,830
#REF!	Content Review	Non-Software	August	September	207,218	99,513	99,513	99,513
#REF!	Standard Setting	Non-Software	August	September	n/a	168,437	n/a	n/a
#REF!	Standards Revision/Performance Level Descriptors incl. Meeting	Non-Software	August	September	n/a	80,021	n/a	n/a
	Grand Total				**2,550,438**	**2,494,836**	**2,246,378**	**2,258,378**

F-2 Proposed vendor Staff, Resource Hours and Rates Worksheet
Use the proposed vendor staff position, resource hours and rates worksheet to indicate the individuals that will be assigned to the project, hours and applicable rates. Names must be provided for individuals designated for key roles, but titles are sufficient for others. Information is required by phase.

Table F-2: Proposed Vendor Staff, Resource Hours and Rates Worksheet SCHOOL YEAR 2017-18

Title	Name	Initiation	Implementation	Project Close out	Category Hourly Rate	Hours X Rate
Project Manager	*Tom Glorfield*	200	1,000		$ 174.14	$ 208,968
Technology Consultant	*Sonja Hubbard*	40	860		$ 164.98	$ 148,482
Program Coordinator and Assistants		240	960		$ 79.38	$ 95,256
Call Center Manager	*Cindy Benis*	8	112		$ 122.21	$ 14,665
Lead Psychometrician	*Ahmet Turan*		400		$ 219.98	$ 87,992
Psychometricians			800		$ 141.49	$ 113,192
Research Associates/Assistants			800		$ 103.93	$ 83,144
Content Development Lead	*Jacob Wilkes*		240		$ 112.82	$ 27,077
Test Development Specialists			1,080		$ 90.55	$ 97,794
Production Manager			80		$ 149.71	$ 11,977
Production, Graphics, and			200		$ 85.55	$ 17,110
Software Project Manager	*Scott Christian Redm*	16	44		$ 180.26	$ 10,816
Senior Application Developer	*Alan Reeve*	16	44		$ 164.98	$ 9,899
Lead Developers and Data Analysts, All Systems		240	2,760		$ 127.21	$ 381,630
Lead, Software Quality Assurance	*Dorian Sofiaj*		80		$ 205.42	$ 16,434
QA Leads and Staff			240		$ 106.93	$ 25,663
Lead, Network Engineering	*Fara Tapscott*		80		$ 192.48	$ 15,398
Network Engineers			420		$ 113.04	$ 47,477
Performance Scoring Lead	*Damon Hartzler*	40	360		$ 132.32	$ 52,928
Scoring Leads and Scorers			1,260		$ 106.93	$ 134,732
Score Reporting Lead	*Dina Booher*		200		$ 103.85	$ 20,770
Score Reporting Specialists			280		$ 81.38	$ 22,786
TOTALS		**800**	**12,300**			**$ 1,644,189**

Add rows as appropriate for proposal.

*This information is important and required so that the NH DOE can break down the costs for services and personnel when presenting a contract for approval by the Governor and Executive Council

F-2 Proposed vendor Staff, Resource Hours and Rates Worksheet

Use the proposed vendor staff position, resource hours and rates worksheet to indicate the individuals that will be assigned to the project, hours and applicable rates. Names must be provided for individuals designated for key roles, but titles are sufficient for others. Information is required by phase.

Table F-2: Proposed Vendor Staff, Resource Hours and Rates Worksheet SCHOOL YEAR 2018-19 (and beyond)

Title	Name	Initiation	Implementation	Project Close out (final year only)	Category Hourly Rate	Hours X Rate
Project Manager	*Tom Glorfield*		1,080	120	$ 177.62	$ 213,147
Technology Consultant	*Sonja Hubbard*		900		$ 168.28	$ 151,452
Program Coordinator and Assistants			976		$ 80.97	$ 79,024
Call Center Manager	*Cindy Benis*		120		$ 124.65	$ 14,959
Lead Psychometrician	*Ahmet Turan*		200		$ 224.38	$ 44,876
Psychometricians			600		$ 144.32	$ 86,592
Research Associates/Assistants			700		$ 106.01	$ 74,206
Content Development Lead	*Jacob Wilkes*		160		$ 115.08	$ 18,412
Test Development Specialists			940		$ 92.36	$ 86,819
Production Manager			80		$ 152.70	$ 12,216
Production, Graphics, and			120		$ 87.26	$ 10,471
Software Project Manager	*Scott Christian Redmond*		60		$ 183.87	$ 11,032
Senior Application Developer	*Alan Reeve*		60		$ 168.28	$ 10,097
Lead Developers and Data Analysts, All Systems			2,500		$ 129.75	$ 324,386
Lead, Software Quality Assurance	*Dorian Sofiaj*		80		$ 209.53	$ 16,762
QA Leads and Staff			190		$ 109.07	$ 20,723
Lead, Network Engineering	*Fara Tapscott*		80		$ 196.33	$ 15,706
Network Engineers			420		$ 115.30	$ 48,426
Performance Scoring Lead	*Damon Hartzler*		400		$ 134.97	$ 53,987
Scoring Leads and Scorers			1,260		$ 109.07	$ 137,426
Score Reporting Lead	*Dina Booher*		80		$ 105.93	$ 8,474
Score Reporting Specialists			160		$ 83.01	$ 13,281
TOTALS		**-**	**11,166**	**120**		**$ 1,452,475**

Add rows as appropriate for proposal.

*This information is important and required so that the NH DOE can break down the costs for services and personnel when presenting a contract for approval by the Governor and Executive Council

F-3 Future Vendor Rates Worksheet

The State may request additional services from the selected vendor and requires rates in the event that additional service is required. The following format must be used to provide this information. "SFY" refers to State Fiscal Year. The New Hampshire fiscal year runs from July 1 through June 30 of the following calendar year. Positions not identified in the Proposed Position Worksheet may be included in the future vendor rates worksheet.

Table F-3: Future Vendor Rates Worksheet

Required					Optional		
Position Title	**SFY 2018**	**SFY 2019**	**SFY 2020**	**SFY 2021**	**SFY 2022**	**SFY 2023**	**SFY 20214**
Project Manager	$ 174.14	$ 177.62	$ 181.18	$ 184.80	$ 188.49	$ 192.26	$ 196.11
Technology Consultant	$ 164.98	$ 168.28	$ 171.65	$ 175.08	$ 178.58	$ 182.15	$ 185.79
Program Coordinator and Assistants	$ 79.38	$ 80.97	$ 82.59	$ 84.24	$ 85.92	$ 87.64	$ 89.39
Call Center Manager	$ 122.21	$ 124.65	$ 127.15	$ 129.69	$ 132.28	$ 134.93	$ 137.63
Lead Psychometrician	$ 219.98	$ 224.38	$ 228.87	$ 233.44	$ 238.11	$ 242.88	$ 247.73
Psychometricians	$ 141.49	$ 144.32	$ 147.21	$ 150.15	$ 153.15	$ 156.22	$ 159.34
Research Associates/Assistants	$ 103.93	$ 106.01	$ 108.13	$ 110.29	$ 112.50	$ 114.75	$ 117.04
Content Development Lead	$ 112.82	$ 115.08	$ 117.38	$ 119.73	$ 122.12	$ 124.56	$ 127.05
Test Development Specialists	$ 90.55	$ 92.36	$ 94.21	$ 96.09	$ 98.01	$ 99.97	$ 101.97
Production Manager	$ 149.71	$ 152.70	$ 155.76	$ 158.87	$ 162.05	$ 165.29	$ 168.60
Publication staff, including Production, Graphics, and Editorial	$ 85.55	$ 87.26	$ 89.01	$ 90.79	$ 92.60	$ 94.45	$ 96.34
Software Project Manager	$ 180.26	$ 183.87	$ 187.54	$ 191.29	$ 195.12	$ 199.02	$ 203.00
Senior Application Developer	$ 164.98	$ 168.28	$ 171.65	$ 175.08	$ 178.58	$ 182.15	$ 185.79
Lead Developers and Data Analysts, All Systems	$ 127.21	$ 129.75	$ 132.35	$ 135.00	$ 137.70	$ 140.45	$ 143.26

Table F-3: Future Vendor Rates Worksheet

Required					Optional		
Position Title	**SFY 2018**	**SFY 2019**	**SFY 2020**	**SFY 2021**	**SFY 2022**	**SFY 2023**	**SFY 20214**
Lead, Software Quality Assurance	$ 205.42	$ 209.53	$ 213.72	$ 217.99	$ 222.35	$ 226.80	$ 231.34
QA Leads and Staff	$ 106.93	$ 109.07	$ 111.25	$ 113.47	$ 115.74	$ 118.06	$ 120.42
Lead, Network Engineering	$ 192.48	$ 196.33	$ 200.26	$ 204.26	$ 208.35	$ 212.51	$ 216.76
Network Engineers	$ 113.04	$ 115.30	$ 117.61	$ 119.96	$ 122.36	$ 124.81	$ 127.30
Performance Scoring Lead	$ 132.32	$ 134.97	$ 137.67	$ 140.42	$ 143.23	$ 146.09	$ 149.01
Scoring Leads and Scorers	$ 106.93	$ 109.07	$ 111.25	$ 113.47	$ 115.74	$ 118.06	$ 120.42
Score Reporting Lead	$ 103.85	$ 105.93	$ 108.05	$ 110.21	$ 112.41	$ 114.66	$ 116.95
Score Reporting Specialists	$ 81.38	$ 83.01	$ 84.67	$ 86.36	$ 88.09	$ 89.85	$ 91.65

Add rows as appropriate for proposal.

F-4 Software Licensing, Maintenance, and Support Pricing Worksheet

Table F-4: Software Licensing, Maintenance, and Support Pricing Worksheet

Required						Optional		
Software Name	Initial Cost	Maintenance Support and Upgrades				Maintenance Support and Upgrades		
		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
License Fees included in per-student test administration costs	Included in Table F-1					Included in Table F-1		

Add rows as appropriate for proposal.

End User Support Plan

F-5 Web Site Hosting, Maintenance, and Support Pricing Worksheet

Table F-5: Web Site Hosting, Maintenance, and Support Pricing Worksheet

Required						Optional		
HOSTED SERVICES	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**TOTAL**	**Year 5**	**Year 6**	**Year 7**
Web Site Hosting Fee	Included in Table F-1					Included in Table F-1		
Technical Support and updates								
Maintenance and Updates								
GRAND TOTAL								

Add rows as appropriate for proposal.

Cost Proposal

Cost Proposal

Below, we present our separate cost proposal as referenced in Topic 42 Pricing Model and Section VII: Pricing Model.

American Institutes for Research (AIR) has based our proposal and fixed prices on the terms, conditions, and schedule outlined in the State of New Hampshire's NH Statewide Assessments Request for Proposal (RFP) #2017-073 as amended by:

- Addendum 1 issued on April 10, 2017
- Addendum 2 issued on/about April 13, 2017

We have documented our assumptions throughout the proposal. We have relied on the RFP and Q&A documents in every case where they clarify or further define the requirements of the RFP.

Assumptions Underlying our Fixed Prices

AIR is bidding on all five components of the RFP and the fixed price.

- Summative ELA and Mathematics assessments in grades 3–8 will be online, adaptive tests built –using AIR's Independent College and Career Readiness (ICCR) item banks. The online adaptive algorithm will be finalized in Year 1.
- Summative Science in grades 5, 8, and 11 will be online, fixed-form tests built using ICCR item bank Science items. These fixed-form assessments will be created each year of the contract.
- Interim ELA and Mathematics assessments in grades 3–8 will be online, fixed-form tests built by licensing an existing item bank developed to measure college and career ready skills in English language arts and mathematics and linked to the ICCR summative item bank. These fixed-form assessments will be created in Year 1.
- Interim Science in grades 5, 8, and 11 will be online, fixed-form tests built by licensing an existing item bank developed to measure college and career ready skills and three-dimensional science skills described in the K–12 Framework and linked to the ICCR summative item bank. These fixed-form assessments will be created in Year 1.
- The Reporting Portal will be available through our Online Reporting System (ORS). The Interim reports will be available through our AIRWays reporting system.

In this section, we wanted to highlight several important assumptions discussed elsewhere, all of which we believe are consistent with the RFP.

In the event of a reduction or change in the scope of work or partial termination for convenience, AIR assumed that any adjustment in contract price will account for potential increases in the price of the remaining tasks occasioned by such reduction or change in the scope of work.

We assume that AIR and New Hampshire will agree on terms whereby AIR will retain ownership of all software and other AIR intellectual property (IP) developed before entering into the Contract as well as any derivations of this IP, and New Hampshire will own all of its data and deliverables.

Any conflicts between the provided documents (e.g., the RFP, Terms and Conditions, and Proposal) will be resolved during negotiations. We assume that the final contract will include an order of precedence in which AIR's proposal is read together with the RFP and Q&A documents to determine the contract requirements.

Schedule and Contract Requirements

We assume that we will build mutually agreeable schedules to allow both AIR and New Hampshire to meet all quality standards. The format and general content for the online score reporting, data file formats, and other reports and documents will be finalized with the first operational use.

Item Development and Release

Our base costing assumed that 100% of the items will be sourced from the ICCR item bank's ELA, Mathematics, and Science item pools and a licensed bank for interim assessments. All license fees are included in our fixed prices. This is the most cost-effective option and meets all of the RFP requirements. All items will be brought to New Hampshire Content and Bias meetings for approval or rejection before being placed in the operational item pool. Rubric Validation and Data Review will be done (and paid for) by AIR.

If New Hampshire would like to develop their own items, we are happy to do so at the following rates (plus committee time, any additional copyright permissions, and optional accommodations):

- New Hampshire-owned Science clusters at $9,500 per cluster
- New Hampshire-owned Science, ELA, and Mathematics stand-alone items at $1,500 per item on field-test form
- New Hampshire-owned Writing prompts at $2,544 per prompt
- Text-to-Speech (TTS) of each item at $20 per item
- American Sign Language (ASL) of each item at $1,600 per video
- Text-to-Braille of each item at $100 per item

Item Release

A bank of released ICCR items currently exists, and AIR is committed to growing this released item bank each year. AIR will work with New Hampshire (and other states using ICCR items) to identify which additional items will be released each year.

License Fee: All license and necessary copyright fees are included in the Base Price.

Handscoring

We have assumed a fully machine-scored assessment, with validation handscoring. As outlined in the proposal, we have assumed that the ELA test will include one writing prompt per student and that it will be scored using our AI scoring system. We have included a 15% human read-behind.

Reporting

Our pricing includes one copy of the Individual Student Report (ISR) shipped to school districts. The report will be a two-page, double-sided print ISR with a Parent Brochure. All other educator reports and interpretive guides will be housed either on the Online Reporting System or the AIRWays reporting system.

Accommodations

All ICCR items will be tagged for Text-to-Speech accommodation. For Braille students, Text-to-Braille tagging will be performed for one form per grade. If New Hampshire would like additional online accommodations, we are happy to do so at the prices listed above in the item development section.

For New Hampshire students who require a paper test accommodation, we offer AIR's print-on-demand feature. This accommodation allows the test administrator's work station to print out items as they are administered to a student. The student can then be given the paper, and the student or a scribe enters the response. Using this system, students receive exactly the same adaptive test as students who do not require this accommodation (although some items may be excluded from the pool). This feature is available for both visual print and Braille.

19.2 Equating Verification

We have included costs to support an independent real-time review of the equating process, analyses, and results by an independent vendor(s) but have not included costs to pay the independent vendor(s).

Committee Meetings/Workshops

AIR assumed that the project management meetings, the committee meetings, and workshops will be held in Concord, New Hampshire. Based on the RFP, we have assumed that all participants will require travel and lodging reimbursement in order to attend meetings. These participants would require mileage, dinner, and hotel accommodations. For meeting participants (teachers or educators), either stipends at $150 per day or substitute teacher reimbursements have been budgeted.

For the regional training workshops, we have included four workshops per year. Costs include the facility, refreshments, and up to two AIR staff members and two New Hampshire staff members. Consistent with the RFP and our experience, AIR has assumed we are not responsible for any reimbursement (e.g., travel, lodging, substitute teacher payments, etc.) for participants in training.

During the item review meetings, New Hampshire educators will have the ability to accept or reject items.

For the TAC meeting, we assumed that there will be seven TAC members. We assumed that each of them will require flight, lodging, and per diems. There was no honorarium amount listed, so we included $1,500 per day.

A summary of meetings is listed in Exhibit I below.

Exhibit I: Summary of Meetings

		Assumptions of Logistical Costs within Meeting Price					
Year	**Meeting Name**	**Meetings Per Year**	**Days per Meeting**	**NH Staff**	**Meeting Participants**	**AIR Staff**	**Total Participants**
All	Kickoff / Annual Meeting	1	2	7	0	7	14
Year 1	Test Blueprint Review Meeting	1	3	6	0	6	12
Year 1	Performance Level Descriptors (PLD) Review	1	3	6	18	6	30
Year 1	ELA Bias/Sensitivity Review	1	2	1	5	2	8
Year 1	Mathematics Bias/Sensitivity Review	1	2	1	5	2	8
Year 1	Science Bias/Sensitivity Review	1	2	1	5	2	8
Year 1	ELA Content Review	1	3	3	18	6	27
Year 1	Mathematics Content Review	1	3	3	18	6	27
Year 1	Science Content Review	1	3	3	9	6	18
All	Regional Training Workshops	4	1	2	100	2	104
All	TAC Meeting	2	2	4	7	3	14
Year 1	ELA Standard-Setting Workshop	1	3	6	15	6	27

Exhibit I: Summary of Meetings (continued)

		Assumptions of Logistical Costs within Meeting Price					
Year	**Meeting Name**	**Meetings Per Year**	**Days per Meeting**	**NH Staff**	**Meeting Participants**	**AIR Staff**	**Total Participants**
Year 1	Mathematics Standard-Setting Workshop	1	3	6	15	6	27
Year 1	Science Standard-Setting Workshop	1	3	6	15	6	27
Years 2–4	ELA Bias/Sensitivity Review	1	1	1	5	1	7
Years 2–4	Mathematics Bias/Sensitivity Review	1	1	1	5	1	7
Years 2–4	Science Bias/Sensitivity Review	1	1	1	5	1	7
Years 2–4	ELA Content Review	1	1	3	18	6	27
Years 2–4	Mathematics Content Review	1	1	3	18	6	27
Years 2–4	Science Content Review	1	1	3	9	3	15

Notes on Completing the Cost Tables

We did our best to allocate costs on the Appendix F cost forms to the sequential parts of the program; however, spreading costs across all of the specific deliverables is not an exact science. Many deliverables are interrelated and/or completed in combination with other deliverables. In addition, we have found that clients prefer to make larger payments when specific milestone tasks are accomplished, such as the testing window has closed, and delivery of test materials, score reports, and technical reports have occurred. We would be happy to provide further detail.

- We used the Appendix F cost format, but did add text and rows within Table F-1 to help clarify our costs. Any changes to the template were made with red text.
- In Table F-1, there are no Software deliverables.
- Table F-2 was created for two years. One table for Year 1 and one table for Year 2.
- In Tables F-4 and F-5, all license fees and costs are included in our deliverables.

Within our total proposed annual prices, we will be happy move costs between line items if any of our assumptions are incorrect, and/or to negotiate other changes to specific payments with the Department.

F-1 Activities/Deliverables/Milestones Pricing Worksheet – Deliverables List

The vendor must include, within the not-to-exceed for IT service activities, tasks and preparation of required deliverables, pricing for the deliverables required based on the proposed approach, and methodology and tools. The following format must be used to provide this information. Please note: this information is for Year 1. **Please add row**s for activities, deliverables, or milestones not included in the table below. **Please add columns** if specific on-going costs, not covered by costs in Tables F4 and F5, should be included.

Table F-1: Activities/Deliverables/Milestones Pricing Worksheet

		Required						
	Activity, Deliverable, or Milestone	**Deliverable Type**	**Projected Delivery Date**	**Projected Payment Date**	**Projected Year 1 Price**	**Projected Year 2 Price**	**Projected Year 3 Price**	**Projected Year 4 Price**
PLANNING AND PROJECT MANAGEMENT								
1	Conduct Project Kickoff Meeting	Non-Software	July	September	26,153	26,153	26,153	26,153
2	Project Status Reports	Written	Monthly	Quarterly	212,000	212,000	212,000	212,000
3	Work Plan	Written	July	September	7,500	7,500	7,500	7,500
4	Infrastructure Plan, including Desktop and Network Configuration Requirements	Written	July	September	1,000	-	-	-
5	Security Plan	Written	July	September	1,000	-	-	-
6	Communications and Change Management Plan	Written	July	September	1,000	-	-	-
7	Requirements Trace Ability Matrix	Written	July	September	1,000	-	-	-
8	Software Configuration Plan	Written	July	September	1,000	-	-	-
9	Systems Interface Plan and Design/Capability	Written	July	September	7,500	7,500	7,500	7,500
10	Testing Plan	Written	July	September	1,000	-	-	-
11	Data Conversion Plan and Design	Written	July	September	1,000	-	-	-
12	Deployment and Roll-out Plan	Written	July	September	1,000	-	-	-
13	Comprehensive Training Plan and Curriculum	Written	July	September	12,139	12,139	12,139	12,139
14	End User Support Plan	Written	July	September	1,000	-	-	-
15	Business Continuity Plan	Written	July	September	1,000	-	-	-

		Required						
	Activity, Deliverable, or Milestone	**Deliverable Type**	**Projected Delivery Date**	**Projected Payment Date**	**Projected Year 1 Price**	**Projected Year 2 Price**	**Projected Year 3 Price**	**Projected Year 4 Price**
16	Documentation of Operational Procedures	Written	July	September	1,000	-	-	-
INSTALLATION								
17	Provide Software Licenses, if needed	Written	included in fees below					
18	Provide Fully Tested Data Conversion Software	Software	n/a	n/a	n/a	n/a	n/a	n/a
19	Provide Software Installed, Configured, and Operational to Satisfy State Requirements	Software	n/a	n/a	n/a	n/a	n/a	n/a
TESTING								
20	Conduct Integration Testing	Non-Software	included in fees below					
21	Conduct User Acceptance Testing	Non-Software	included in fees below					
22	Perform Production Tests	Non-Software	included in fees below					
23	Test In-Bound and Out-Bound Interfaces	Software	n/a	n/a	n/a	n/a	n/a	n/a
24	Conduct System Performance (Load/Stress) Testing	Non-Software	included in fees below					
25	Certification of 3rd Party Pen Testing and Application Vulnerability Scanning	Non-Software	included in fees below					
SYSTEM DEPLOYMENT								
26	Converted Data Loaded into Production Environment	*Software*	n/a	n/a	n/a	n/a	n/a	n/a
27	Provide Tools for Backup and Recovery of all Applications and Data	*Software*	n/a	n/a	n/a	n/a	n/a	n/a
28	Conduct Training	*Non-Software*	October	December	86,530	86,530	86,530	86,530
29	Cutover to New Software	*Non-Software*	n/a	n/a	n/a	n/a	n/a	n/a
30	NA	n/a	n/a	n/a	n/a	n/a	n/a	n/a

		Required						
	Activity, Deliverable, or Milestone	**Deliverable Type**	**Projected Delivery Date**	**Projected Payment Date**	**Projected Year 1 Price**	**Projected Year 2 Price**	**Projected Year 3 Price**	**Projected Year 4 Price**
31	Provide Documentation	*Written*	included in fees below					
32	Execute Security Plan	*Non-Software*	included in fees below					
OPERATIONS								
33	Ongoing Hosting Support	Non-Software	included in fees below					
34	Ongoing Support & Maintenance	Software	included in fees below					
35	Item Banks and Configuration for Testing (UAT)	Non-Software	July	September	207,902	207,902	207,902	207,902
36	Test Administration	Non-Software	May	June	808,500	808,500	808,500	808,500
37	Test Scoring	Non-Software	June	June	150,000	150,000	150,000	150,000
38	Student Data File	Non-Software	June	July	157,500	157,500	157,500	157,500
39	Online Reporting System	Non-Software	November	December	95,332	95,332	95,332	95,332
40	Deliver Individual Student Reports (ISR)	Non-Software	June	July	134,878	134,878	134,878	134,878
41	TAC Meeting (2)	Non-Software	January / June	March / June	113,101	113,101	113,101	113,101
42	Technical Report		June	June	50,000	50,000	50,000	50,000
43	Conduct Project Exit Meeting	Non-Software			n/a	n/a	n/a	12,000
ITEM DEVELOPMENT and TEST CONSTRUCTION								
44	Test Design	Non-Software	October	December	44,958	n/a	n/a	n/a
45	Summative Test Construction	Non-Software	January	March	75,019	30,000	30,000	30,000
46	Interim Form Construction	Non-Software	January	March	31,115	5,000	5,000	5,000
47	Practice Tests	Non-Software	October	December	42,039	n/a	n/a	n/a
#REF!	Bias/Sensitivity Committee Review	Non-Software	August	September	70,054	42,830	42,830	42,830
#REF!	Content Review	Non-Software	August	September	207,218	99,513	99,513	99,513
#REF!	Standard Setting	Non-Software	August	September	n/a	168,437	n/a	n/a
#REF!	Standards Revision/Performance Level Descriptors incl. Meeting	Non-Software	August	September	n/a	80,021	n/a	n/a
	Grand Total				**2,550,438**	**2,494,836**	**2,246,378**	**2,258,378**

F-2 Proposed vendor Staff, Resource Hours and Rates Worksheet

Use the proposed vendor staff position, resource hours and rates worksheet to indicate the individuals that will be assigned to the project, hours and applicable rates. Names must be provided for individuals designated for key roles, but titles are sufficient for others. Information is required by phase.

Table F-2: Proposed Vendor Staff, Resource Hours and Rates Worksheet SCHOOL YEAR 2017-18

Title	Name	Initiation	Implementation	Project Close out	Category Hourly Rate	Hours X Rate
Project Manager	*Tom Glorfield*	200	1,000		$ 174.14	$ 208,968
Technology Consultant	*Sonja Hubbard*	40	860		$ 164.98	$ 148,482
Program Coordinator and Assistants		240	960		$ 79.38	$ 95,256
Call Center Manager	*Cindy Benis*	8	112		$ 122.21	$ 14,665
Lead Psychometrician	*Ahmet Turan*		400		$ 219.98	$ 87,992
Psychometricians			800		$ 141.49	$ 113,192
Research Associates/Assistants			800		$ 103.93	$ 83,144
Content Development Lead	*Jacob Wilkes*		240		$ 112.82	$ 27,077
Test Development Specialists			1,080		$ 90.55	$ 97,794
Production Manager			80		$ 149.71	$ 11,977
Production, Graphics, and			200		$ 85.55	$ 17,110
Software Project Manager	*Scott Christian Redmo*	16	44		$ 180.26	$ 10,816
Senior Application Developer	*Alan Reeve*	16	44		$ 164.98	$ 9,899
Lead Developers and Data Analysts, All Systems		240	2,760		$ 127.21	$ 381,630
Lead, Software Quality Assurance	*Dorian Sofiaj*		80		$ 205.42	$ 16,434
QA Leads and Staff			240		$ 106.93	$ 25,663
Lead, Network Engineering	*Fara Tapscott*		80		$ 192.48	$ 15,398
Network Engineers			420		$ 113.04	$ 47,477
Performance Scoring Lead	*Damon Hartzler*	40	360		$ 132.32	$ 52,928
Scoring Leads and Scorers			1,260		$ 106.93	$ 134,732
Score Reporting Lead	*Dina Booher*		200		$ 103.85	$ 20,770
Score Reporting Specialists			280		$ 81.38	$ 22,786
TOTALS		**800**	**12,300**			**$ 1,644,189**

Add rows as appropriate for proposal.

*This information is important and required so that the NH DOE can break down the costs for services and personnel when presenting a contract for approval by the Governor and Executive Council

F-2 Proposed vendor Staff, Resource Hours and Rates Worksheet

Use the proposed vendor staff position, resource hours and rates worksheet to indicate the individuals that will be assigned to the project, hours and applicable rates. Names must be provided for individuals designated for key roles, but titles are sufficient for others. Information is required by phase.

Table F-2: Proposed Vendor Staff, Resource Hours and Rates Worksheet SCHOOL YEAR 2018-19 (and beyond)

Title	Name	Initiation	Implementation	Project Close out (final year only)	Category Hourly Rate	Hours X Rate
Project Manager	*Tom Glorfield*		1,080	120	$ 177.62	$ 213,147
Technology Consultant	*Sonja Hubbard*		900		$ 168.28	$ 151,452
Program Coordinator and Assistants			976		$ 80.97	$ 79,024
Call Center Manager	*Cindy Benis*		120		$ 124.65	$ 14,959
Lead Psychometrician	*Ahmet Turan*		200		$ 224.38	$ 44,876
Psychometricians			600		$ 144.32	$ 86,592
Research Associates/Assistants			700		$ 106.01	$ 74,206
Content Development Lead	*Jacob Wilkes*		160		$ 115.08	$ 18,412
Test Development Specialists			940		$ 92.36	$ 86,819
Production Manager			80		$ 152.70	$ 12,216
Production, Graphics, and			120		$ 87.26	$ 10,471
Software Project Manager	*Scott Christian Redmond*		60		$ 183.87	$ 11,032
Senior Application Developer	*Alan Reeve*		60		$ 168.28	$ 10,097
Lead Developers and Data Analysts, All Systems			2,500		$ 129.75	$ 324,386
Lead, Software Quality Assurance	*Dorian Sofiaj*		80		$ 209.53	$ 16,762
QA Leads and Staff			190		$ 109.07	$ 20,723
Lead, Network Engineering	*Fara Tapscott*		80		$ 196.33	$ 15,706
Network Engineers			420		$ 115.30	$ 48,426
Performance Scoring Lead	*Damon Hartzler*		400		$ 134.97	$ 53,987
Scoring Leads and Scorers			1,260		$ 109.07	$ 137,426
Score Reporting Lead	*Dina Booher*		80		$ 105.93	$ 8,474
Score Reporting Specialists			160		$ 83.01	$ 13,281
TOTALS		**-**	**11,166**	**120**		**$ 1,452,475**

Add rows as appropriate for proposal.

*This information is important and required so that the NH DOE can break down the costs for services and personnel when presenting a contract for approval by the Governor and Executive Council

F-3 Future Vendor Rates Worksheet

The State may request additional services from the selected vendor and requires rates in the event that additional service is required. The following format must be used to provide this information. "SFY" refers to State Fiscal Year. The New Hampshire fiscal year runs from July 1 through June 30 of the following calendar year. Positions not identified in the Proposed Position Worksheet may be included in the future vendor rates worksheet.

Table F-3: Future Vendor Rates Worksheet

Required					Optional		
Position Title	**SFY 2018**	**SFY 2019**	**SFY 2020**	**SFY 2021**	**SFY 2022**	**SFY 2023**	**SFY 20214**
Project Manager	$ 174.14	$ 177.62	$ 181.18	$ 184.80	$ 188.49	$ 192.26	$ 196.11
Technology Consultant	$ 164.98	$ 168.28	$ 171.65	$ 175.08	$ 178.58	$ 182.15	$ 185.79
Program Coordinator and Assistants	$ 79.38	$ 80.97	$ 82.59	$ 84.24	$ 85.92	$ 87.64	$ 89.39
Call Center Manager	$ 122.21	$ 124.65	$ 127.15	$ 129.69	$ 132.28	$ 134.93	$ 137.63
Lead Psychometrician	$ 219.98	$ 224.38	$ 228.87	$ 233.44	$ 238.11	$ 242.88	$ 247.73
Psychometricians	$ 141.49	$ 144.32	$ 147.21	$ 150.15	$ 153.15	$ 156.22	$ 159.34
Research Associates/Assistants	$ 103.93	$ 106.01	$ 108.13	$ 110.29	$ 112.50	$ 114.75	$ 117.04
Content Development Lead	$ 112.82	$ 115.08	$ 117.38	$ 119.73	$ 122.12	$ 124.56	$ 127.05
Test Development Specialists	$ 90.55	$ 92.36	$ 94.21	$ 96.09	$ 98.01	$ 99.97	$ 101.97
Production Manager	$ 149.71	$ 152.70	$ 155.76	$ 158.87	$ 162.05	$ 165.29	$ 168.60
Publication staff, including Production, Graphics, and Editorial	$ 85.55	$ 87.26	$ 89.01	$ 90.79	$ 92.60	$ 94.45	$ 96.34
Software Project Manager	$ 180.26	$ 183.87	$ 187.54	$ 191.29	$ 195.12	$ 199.02	$ 203.00
Senior Application Developer	$ 164.98	$ 168.28	$ 171.65	$ 175.08	$ 178.58	$ 182.15	$ 185.79

Table F-3: Future Vendor Rates Worksheet

Required					Optional		
Position Title	**SFY 2018**	**SFY 2019**	**SFY 2020**	**SFY 2021**	**SFY 2022**	**SFY 2023**	**SFY 20214**
Lead Developers and Data Analysts, All Systems	$ 127.21	$ 129.75	$ 132.35	$ 135.00	$ 137.70	$ 140.45	$ 143.26
Lead, Software Quality Assurance	$ 205.42	$ 209.53	$ 213.72	$ 217.99	$ 222.35	$ 226.80	$ 231.34
QA Leads and Staff	$ 106.93	$ 109.07	$ 111.25	$ 113.47	$ 115.74	$ 118.06	$ 120.42
Lead, Network Engineering	$ 192.48	$ 196.33	$ 200.26	$ 204.26	$ 208.35	$ 212.51	$ 216.76
Network Engineers	$ 113.04	$ 115.30	$ 117.61	$ 119.96	$ 122.36	$ 124.81	$ 127.30
Performance Scoring Lead	$ 132.32	$ 134.97	$ 137.67	$ 140.42	$ 143.23	$ 146.09	$ 149.01
Scoring Leads and Scorers	$ 106.93	$ 109.07	$ 111.25	$ 113.47	$ 115.74	$ 118.06	$ 120.42
Score Reporting Lead	$ 103.85	$ 105.93	$ 108.05	$ 110.21	$ 112.41	$ 114.66	$ 116.95
Score Reporting Specialists	$ 81.38	$ 83.01	$ 84.67	$ 86.36	$ 88.09	$ 89.85	$ 91.65

Add rows as appropriate for proposal.

Table F-3: Future Vendor Rates Worksheet

Required					Optional		
Position Title	SFY 2018	SFY 2019	SFY 2020	SFY 2021	SFY 2022	SFY 2023	SFY 20214

End User Support Plan

F-4 Software Licensing, Maintenance, and Support Pricing Worksheet

Table F-4: Software Licensing, Maintenance, and Support Pricing Worksheet

Required						Optional		
Software Name	Initial Cost	Maintenance Support and Upgrades				Maintenance Support and Upgrades		
		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
License Fees included in per-student test administration costs	Included in Table F-1					Included in Table F-1		

Add rows as appropriate for proposal.

End User Support Plan

F-5 Web Site Hosting, Maintenance, and Support Pricing Worksheet

Table F-5: Web Site Hosting, Maintenance, and Support Pricing Worksheet

Required						Optional		
HOSTED SERVICES	Year 1	Year 2	Year 3	Year 4	TOTAL	Year 5	Year 6	Year 7
Web Site Hosting Fee	Included in Table F-1					Included in Table F-1		
Technical Support and updates								
Maintenance and Updates								
GRAND TOTAL								

Add rows as appropriate for proposal.