# Table of Contents

Cover Letter

Executive Summary

Detailed Response

I. Summative & Interim Technology
   I.C Technology Requirements - Overview ............................................................. I-1
   I.C.1 Item Bank .................................................................................................. I-7
   I.C.2 Test Administration .................................................................................. I-15
   I.C.3 Test Delivery System ............................................................................... I-41
   I.C.4 Technology Requirements ....................................................................... I-91
   I.D Psychometrics ............................................................................................ I-103
   I.D.2 Blueprints .............................................................................................. I-128
   I.D.3 Standard Setting ..................................................................................... I-131

II. Summative & Interim Test Content—ELA, Mathematics, Science
   II.B.1 Overall Approach and Sample Items and Item Types................................ II-1
   II.B.2 Item Development/Procurement............................................................. II-55
   II.B.3 Annual Analysis, Review, Revision....................................................... II-89
   II.B.4 Technology Requirements ..................................................................... II-91
III. Formative Assessments

III.B. Overall Approach ...............................................................................................................III-1

III.B.2 Educator and User Functionality .....................................................................................III-5

III.B.3 System Functionality ........................................................................................................III-13

III.B.4 Additional Requirements ...............................................................................................III-13

III.B.5 Technology Requirements ..............................................................................................III-17

IV. Other Contract Requirements

IV.A Return and Destruction of Items and Data .........................................................................IV-1

IV.B Timetables .............................................................................................................................IV-3

IV.C Quality Assurance ..............................................................................................................IV-7

IV.D Project Management ...........................................................................................................IV-15

IV.E Technical manuals ...............................................................................................................IV-41

IV.F Contract finalization plan ...................................................................................................IV-45

V Proposal Content Requirements

V.A Demonstration of Expertise in Developing Large-Scale Assessments ......................V-1

V.B Demonstration of Expertise in Computer Based and/or Computer Adaptive Testing .........................................................................................................................V-16

V.C Development of a Project Plan ............................................................................................V-31

V.D Cost Proposal .......................................................................................................................V-32
Appendices

A. Blueprints ..............................................................................................................A-1
B. Sample Communications Materials .................................................................B-1
C. Sample Panelist Evaluation Form .......................................................................C-1
D. ELA Sample Performance Task Reading .........................................................D-1
E. Resumes ..................................................................................................................E-1
F. Project Calendar ....................................................................................................F-1
September 30, 2012

Pam Rugg
Purchasing Agent
State of Utah

RE: Solicitation PR13015
Utah Statewide Computer Adaptive Assessment System

Dear Ms. Rugg and members of the evaluation committee:

In response to the referenced solicitation, the American Institutes for Research (AIR) is pleased to join Data Recognition Corporation in a proposal to support the Utah Statewide Computer Adaptive Assessment System.

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 over 1,600 people working in the areas of assessment, education research and technical assistance, health, human development, and international development. In student assessment, pursuing our mission means collaborating with our clients to deliver assessments that

- lead educators to teach all students all the content;
- model the types of activities and exercises that should occur in the classroom;
- report immediate, actionable feedback to educators, students, and families; and
- support all students with the tools needed for an equal opportunity to effectively access and use the online, adaptive test without distraction.

AIR is currently the only organization delivering statewide, online adaptive tests approved for ESEA accountability. We are doing so in four states, and we have been selected to deliver the Smarter Balanced Assessment Consortium’s pilot test and field test using our online testing system.

Our proposal is for the full scope of work.

- For Section I, Summative and Interim Technology, our online testing system is fully Internet-based and has the smallest technology footprint in the school. We are committed to 100% accessibility for all students and are proud that virtually all students are receiving fully accessible, adaptive tests in and . No existing system is more accessible than ours, and only our system has been proven in real statewide accountability tests.

- For Section II, Summative and Interim Test Content, AIR proposes to work with USOE and Utah educators to build a custom test, rather than shoehorning Utah’s standards into an existing assessment. We will work with USOE on every aspect of design, beginning with the test blueprint. We have proposed initial test blueprints that cover the Utah standards, with a wide variety of machine-scored constructed-response and other technology-enhanced items, and include performance tasks.

- For Section III, Formative Assessments, we will deploy AIR’s Learning Point Navigator instructional and formative assessment delivery system. The system integrates with our interim and summative reporting system, giving teachers the power to probe deeply into individual strengths and weaknesses. Navigator can deliver resources from multiple libraries, and we offer AIR’s open-source AIR Core Library along with a Utah-specific library imported from UTIPS and including items and other resources newly constructed by Utah’s educators.

Thank you in advance for your consideration.

Jon Cohen
Executive Vice President and Director, Assessment Program
Executive Summary
Executive Summary

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 over 1,600 people working in the areas of assessment, education research and technical assistance, health, human development, and international development.

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

- lead educators to teach all students all the content;
- model the types of activities and exercises that should occur in the classroom;
- report immediate, actionable feedback to educators, students, and families; and
- support all students with the tools needed for an equal opportunity to effectively access and use the online adaptive test without distraction.

AIR is currently the only organization delivering statewide online adaptive tests approved for ESEA accountability. We are doing so in four states, and we have been selected to deliver the Smarter Balanced Assessment Consortium’s pilot test and field test using our online testing system.

We are pleased to offer the full scope of work. Under our plan, Utah will be able to discontinue the current CRT after the spring 2013 administration and begin assessing the new Utah Core during the 2013–2014 school year.

The Tests

AIR proposes to work with USOE and Utah educators to build a custom test, rather than shoehorning Utah’s standards into an existing assessment. We will work with USOE on every aspect of design, beginning with the test blueprint. We have proposed initial test blueprints that cover the Utah standards, with a wide variety of machine-scored constructed-response and other technology-enhanced items, and include performance tasks. Each test will include about 40–50 items plus 4–6 activities associated with a performance task. We expect that each test should take approximately 70–120 minutes to administer. The performance task provides our best opportunity to probe the depth of student understanding and to model the sorts of challenging activities that should occur in the classroom.

The vast majority of the items will be machine-scored. Our section on scoring describes our scoring engines, many of which build on work pioneered by researchers in artificial intelligence. We propose that the new items developed be heavily weighted toward real constructed-response items that are machine-scored. These include items in which the student responds with a drawing, equation, written statements, sequence of interactions and choices, or longer written texts.

The test will require human scoring in two instances.

First, each mathematics performance task will include an open-ended question requiring the student to communicate mathematical reasoning and targeting very high cognitive complexity (Depth of Knowledge 4). Despite claims to the contrary by advocates or vendors, validating the logic and reasoning of open-ended, possibly multi-mode math communication is beyond the current capabilities of artificial intelligence scoring. Therefore, we propose to have these responses human-scored. We have partnered with Data Recognition Corporation (DRC) to conduct this scoring on a fast-turnaround basis.

Second, the ELA performance tasks will incorporate a human second read on targeted machine-scored responses. The system can detect scores that have an increased probability of deviating from the intended rubric, and these will be routed to human scorers for independent review.
We have designed science performance tasks that can be entirely machine-scored.

**The Delivery System**

Our online testing system is fully Internet-based and has the smallest technology footprint in the school. The system requires no technology expertise at the schools and works on very primitive hardware and networks, as well as the newest most advanced tablets. The system requires only about 5Kbs of bandwidth per student, requires no caching or local proxy servers, works on machines 10 years old or more, and requires only our secure browser (a special build of the Mozilla browser) installed on the student machines.

We are committed to 100% accessibility for all students and are proud that virtually all students are receiving fully accessible, adaptive tests in and . Our unparalleled suite of embedded supports and access-enhancing tools even enables us to deliver real-time adaptive tests to students who are blind, students who cannot look at a computer monitor (through secure print-on-demand), and students with very limited mobility or dexterity. No existing system is more accessible than ours, and only our system has been proven in real statewide accountability tests.

**The Plan**

Key operational features of our plan include

- importing, sharing, and developing items between contract award and January 2014 to build a robust item pool for the adaptive assessment;
- administering the spring 2014 summative administration as an “operational field test,” designed to provide both calibrations for the entire summative and interim item pool and valid summative scores for accountability; and
- deploying Learning Point Navigator, our online formative assessment and curricu-

lum support tool, in fall 2013 with a collection of open-source curricular material compiled or created by AIR, items imported from UTIPS, and newly created materials created by Utah teachers.

We have contacted both consortia, and neither is yet ready to commit to making its items available. We propose a mix of new development, including development by Utah teachers, use of some portion of existing items, and inviting Utah to join an item-sharing arrangement with our existing clients. The latter gives Utah immediate access to many technology-enhanced and machine-scored constructed-response items written to the Common Core State Standards.

The operational field test implements a model that we have used successfully many times. The approach yields valid summative scores and calibrations for the entire item bank and provides statistics based on motivated administration conditions for standard setting. During this assessment, students will receive tests constructed by the adaptive engine but with an algorithm designed to (a) match the blueprint for every student and (b) deliver each item to a representative sample of students.

The final part of this proposal describes Learning Point Navigator, an instructional and formative assessment delivery system. The system integrates with our interim and summative reporting system, giving teachers the power to probe deeply into individual strengths and weaknesses. Navigator can deliver resources from multiple libraries. We propose to offer AIR’s open-source AIR Core Library, along with a Utah-specific library imported from UTIPS and including items and other resources newly constructed by Utah’s educators.

**Proprietary Information**

AIR has completed the Confidentiality Claim Form to request protection of proprietary information, including names and resumes of our staff, clients lists, and non-public internal financial information.
Specialist
CLAIM OF BUSINESS CONFIDENTIALITY

Pursuant to Utah Code Annotated, Subsections 63G-2-305(1) and (2), and in accordance with Section 63G-2-309, American Institutes for Research (company name) asserts a claim of business confidentiality to protect the following information submitted as part of a Request for Proposals.

This claim is asserted because this information requires protection as it includes:

- Trade secrets as defined in Utah Code Annotated Section 13-24-2 ("Trade secret" means information, including a formula, pattern, compilation, program, device, method, technique, or process, that: (a) derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable by proper means by, other persons who can obtain economic value from its disclosure or use; and (b) is the subject of efforts that are reasonable under the circumstances to maintain its secrecy).

- Commercial information or non-individual financial information obtained from a person if: (a) disclosure of the information could reasonably be expected to result in unfair competitive injury to the person submitting the information or would impair the ability of the governmental entity to obtain necessary information in the future; [and] (b) the person submitting the information has a greater interest in prohibiting access than the public in obtaining access.

This statement of reasons supporting the claim of business confidentiality applies to the following information in this proposal:

<table>
<thead>
<tr>
<th>Page</th>
<th>Paragraph</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple</td>
<td>Names and resumes of AIR and DRC employees; competitors can use this information to steal AIR employees</td>
<td></td>
</tr>
<tr>
<td>Multiple</td>
<td>Names and contact information for AIR clients; competitors can use this information</td>
<td></td>
</tr>
<tr>
<td>Cost Proposal</td>
<td>Individual pricing elements (staff salaries, indirect rates, cost elements). AIR is not attempting to protect our proposed prices as shown on the required budget form. However, as this is a cost-reimbursement contract, AIR is required to submit detailed internal pricing information, which is confidential business information.</td>
<td></td>
</tr>
</tbody>
</table>

Please use additional sheets if needed.

You will be notified if a record claimed to be protected herein under Utah Code Annotated § 63G-2-305(1) or (2) is classified public or if the governmental entity determines that the record should be released after weighing interests under Utah Code Annotated § 63G-2-201(5)(b) or Utah Code Annotated § 63G-2-401(6). See Utah Code Annotated § 63G-2-309.

Signed: ___________________________ Date: 9/28/22
Detailed Response
I. Summative and Interim Technology
I.C. Technology Requirements

The American Institutes for Research is pleased to propose a proven, online adaptive testing solution to assess student proficiency with the Utah Core Standards. Our proposed solution offers

- a robust, online adaptive test delivery system that has proven effective, requiring little equipment, bandwidth, expertise, or support;
- a standards-based adaptive engine capable of supporting sophisticated blueprints, ensuring that every student has a fair chance to show what he or she knows and can do across the full range of the curriculum;
- 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 through simulations;
- the richest collection of embedded supports and accommodations in use in a statewide testing system to ensure access for all students; and
- a system that is guaranteed to be compatible with the Smarter Balanced Assessment Consortium items because it is the same system being used to deliver the pilot test and field test.

AIR’s test delivery system currently delivers statewide assessments in and and is scheduled to deliver the Smarter Balanced pilot test this February. In our adaptive assessments in math and reading have received full approval from federal peer reviewers (in, approval of the full assessment systems, including the alternate assessments, is still pending), and the peer review submission is currently under preparation in, the most recent state to adopt standards-based adaptive assessments.

Robust, Online Adaptive Test Delivery System

In schools and districts, the AIR test delivery system has succeeded even where other systems have failed. 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 purely Internet-based system, and we support operating systems and browsers for 10 years from their release. This covers almost all of the computers found in schools right now. Within the schools, the only software required is our secure browser, which is a secure build of the Mozilla (Firefox) browser. It can be installed with the click of a button, and installation options are available that do not even require administrative privileges on the computer.

Our system does not require any local caching servers, proctor caching servers, or other specialized software requiring expertise, support, or maintenance. Such devices increase the support load at the school, depend on vulnerable within-school servers, and historically have required support and periodic upgrades.

Instead, our system relies on highly optimized, secure Internet delivery. Install the secure browser on the machines that the students will use and the rest happens between the student workstations and our servers. A typical test requires only approximately 5 kilobits of bandwidth per second. This means that a typical school can easily test hundreds of students simultaneously. In fact, 5 to 10 students can be tested on the bandwidth typically available through a single smartphone.
Our approach means a smooth, robust testing experience for students. Actual results from last year show that the typical students waited less than one half-second between the time they finished one item and the time the next item was fully displayed on their screen. The system recovers from interruptions in Internet service without loss of data and often recovers without users even becoming aware that the service was interrupted.

Standards-based Adaptive Engine

While AIR’s test delivery system is able to deliver fixed-form, adaptive, and other types of tests, we are most proud of our ability to deliver standards-based adaptive tests. Our flexible adaptive algorithm, developed in partnership with our client states over the years, implements complex blueprints. The blueprints may contain specifications from the content hierarchy (strand, benchmark, standard, etc.) as well as other constraints such as depth of knowledge, item type, or any other attribute that may be stored about test items. The algorithm ensures that each student receives a test that (a) matches the blueprint and (b) does so with the items that best match student performance, given the blueprint constraints.

Our adaptive engine supports, and we advise all of our clients to use, blueprints that meet the following conditions (which have been advocated by the Consortium for Citizens with Disabilities, an umbrella group encompassing most national advocacy groups for students with disabilities and other exceptional students):

1. Every student is tested on the full range of grade-level content with no discernible differences in the content assessed.
2. Every student is tested on items measuring the same mix of cognitively complex skills, with no discernible difference—regardless of student proficiency.
3. Every student is tested on items reflecting the full range of other aspects of the grade-level curriculum as may be appropriate for the grade and subject.
4. Students are tested on items that provide the best measurement possible within these constraints.

These four principles serve the objective of making sure that every student can accurately demonstrate his or her academic skills and knowledge across the entire grade-level curriculum. AIR’s adaptive algorithm supports blueprints that meet these goals, and we propose to work with USOE to design Utah blueprints that adhere to these principles.

Our standards-based adaptive engine is discussed at length in Section I.D, Psychometrics.

Industry’s Richest Repertoire of Item Types

AIR’s test delivery system is able to deliver the item types defined in the QTI specification, but these item types barely access the measurement capabilities of our system. Typical “technology-enhanced” items, such as hotspot, drag and drop, matching, and hot text, comprise a variety of representations of selected-response items. While these items have their place, they rarely extend measurement beyond the reach of multiple-choice items. We support all of these item types and recognize that they may improve engagement for some students. However, we move far beyond that.

Our experience has proven that technology can extend the measurement range of tests in at least two ways:
Some of the rich, constructed-response item types that heretofore have been scored by humans at great cost in terms of time and money can be captured and scored automatically.

Computers can present challenges to students that were not possible on paper, requiring students to integrate skills to solve problems in ways that are integral to the Common Core State Standards.

Implementing constructed-response items like those on paper tests with automated scoring reduces costs and allows the system to provide educators, students, and families with immediate feedback. Our system offers a range of capabilities in these areas. For example, paper tests often present geometric items or tasks that require students to apply their knowledge to construct shapes with various properties, graph scatter-plots and best-fitting lines, and create concept maps. AIR’s Graphic Response mechanism allows all of these sorts of items to be captured on computer and automatically scored according to sophisticated, explicit rubrics. Our Proposition Scoring Engine can score some types of natural language constructed-response items, and our essay scoring engine can score other types of natural language responses. Our Equation Response mechanism enables students to enter equations that can be automatically scored according to sophisticated rubrics.

Through our Simulation Tool we can go well beyond what can be accomplished on paper. For example, students can conduct a realistic laboratory study using a simulated microscope and virtual samples. “Hard copy” versions of these sorts of performance tasks were prohibitively expensive in the past. Even more exciting possibilities exist to evaluate student choices. For example, in Section II.B.2, Item Development, you will see a sample math item in which students enter values for x in a quadratic equation and see y graphed and returned. They are asked to use the Equation Response mechanism to write the specific quadratic equation graphed. By capturing a student’s trials (the values of x that he or she enters), we can infer the strategy that the student uses to solve the problem and the depth of his or her understanding of the structure of the problem.

The possibilities for new measurement approaches are boundless. In , we piloted a prototype item in which students demonstrate knowledge by playing an interactive computer game. In this social studies prototype, students were asked to apply the lessons learned by the Jamestown colonists in a new but otherwise similar colony. Players (examinees) found themselves in situations comparable to salient situations studied in the Jamestown curriculum, and they were scored on the basis of their appropriate application of the Jamestown lessons. Tasks similar to this one can challenge students to apply ideas learned through a complex reading passage or passages. These are some of the ways that we use technology to probe more deeply into what students know and can do.

Readers will find challenging, innovative, and exciting examples of items in Section II.B.2, Item Development.

Richest Collection of Embedded Supports and Accommodations

AIR has a long history of educational research about and support for students with disabilities and English language learners. We have been and remain committed to making our online testing system accessible to every student. In Section II.B.2, you will read a catalogue of the accommodations and embedded supports available through our system.
Before discussing specific accommodations, we note that our standards-based adaptive approach is itself an accessibility feature. The engine ensures that each student sees the full range of content and the full range of depth of knowledge, regardless of his or her performance; however, these goals can be accomplished while still varying the difficulty of the items. Consider, for example, a math problem requiring students to multiply fractions. Common fractions, such as one-half, make for easier items than do some other values (say, one-seventh). The items may measure exactly the same skill at the same depth of knowledge, but the difficulty (as defined by psychometricians) would differ.

The standards-based adaptive engine ensures that every student (a) will be able to access items that he or she sees on the test (struggling students will see some easy items that they can access) and (b) will be challenged by the test that he or she receives. Each student will see a test that is both accessible and challenging.

Here, we highlight just a few of the more powerful supports provided by the system:

- The ability to deliver real-time, adaptive paper tests, including Braille and large print
- An Accessibility Interface that optimizes presentation for flawless interaction with the JAWS screen reader for use with a keyboard or refreshable Braille display
- One hundred percent keyboard navigability
- Configurable text-to-speech accommodating multiple, different levels of support
- Ability to deliver sign language

Our system offers a real-time, secure “print-on-demand” capability. This enables the system to deliver truly adaptive tests even to students who cannot look at a computer screen (in some cases a scribe may be needed to enter the responses).

Our system provides complete support for Braille. The print-on-demand feature can “print” to an embosser in the school. While our system also provides 100% support for refreshable Braille displays, our research and work with stakeholder groups have revealed significant advantages of the real-time embossing system. For example, refreshable Braille displays show very little text at a time (usually 40 or 80 characters), they cannot display tactile graphics, and they do not display Nemeth Braille (the Braille used in math and science textbooks). These limitations are overcome by our unique real-time embossing solution. Because the cost of brailing the full item bank and the cost of embossers were not been foreseen by the RFP, we include these costs as a separate option so that our price may be fairly compared with comparable solutions from other vendors.

To support refreshable Braille displays, the system can render the items in an alternative format, designed for optimal navigation using a screen-reader, which may drive a refreshable Braille display. This interface, which we call the Accessibility Interface, does not require any different software (except the screen reader itself) but automatically renders when the accommodation is set for a student.

Our system is 100% keyboard-navigable, so students with dexterity issues need not use a mouse. In addition, the system can be navigated with a very limited number of keychords, making it compatible with a range of third-party switch arrays and other assistive devices for mobility- or dexterity-impaired students.

AIR’s powerful text-to-speech system allows varying levels of support, according to rules that we can configure in consultation with USOE. Thus, while some students may have
(for example) only descriptions of graphics and tables read to them, other students may have entire items and response options read. These differentiated supports can be configured to meet Utah’s needs within broad boundaries.

Finally, our system has the capability to deliver sign language. We have met with experts on sign language and advocates for students who are deaf. In all cases, we have been advised against using signing avatars. Sign language stands as an independent language, and the avatars are not currently capable of translating appropriately from English to American Sign Language (ASL). Furthermore, subtler aspects of the communication, such as facial expressions, are not present in the avatars.

Therefore, our sign language capability is through recorded sign translation. Sign language translation was not included in the RFP, so we have not priced it here. The system has the capability to deliver sign language, and we will be happy to price the translation/recording of the sign language at USOE request.

Section II.B.2 presents a full catalog of our tools and embedded supports. Interested readers are directed there.

### Compatibility with Smarter Balanced Items

AIR is honored to have been selected by the Smarter Balanced Assessment Consortium to deliver its pilot and field test. One implication of this is that, since our system will be delivering the Smarter Balanced test, Utah can be certain of complete compatibility between Smarter Balanced items and AIR’s test delivery system. Therefore, if Utah chooses to use or license Smarter Balanced items, integration with our system will be quick and easy (that is, the systems will already be integrated).

Going forward, AIR is building Smarter Balanced an open-source testing system. As part of that effort, our client-side code (the code that runs in the student browser and test administrator browser) will become open source. This client-side code includes the renderers that display (or “render”) the items and manage the interactions with examinees. Therefore, if Utah selects AIR for this contract, the display code used by the Consortium and integrated into its new testing system will be the same as that used in Utah’s system. We believe that this will lower Utah’s long-term cost of ownership and reduce the risk associated with integrating Consortium items.
I.C.1 Item Bank

Item Bank: Item Tracking System

AIR’s item bank, the Item Tracking System (ITS), meets all of the requirements outlined in the RFP and offers some unique capabilities that will make Utah’s program more efficient, innovative, and sustainable. ITS includes full lifecycle support for items and item pools and full content development and management capabilities. In addition, ITS integrates powerful tools for test developers that enable them to develop sophisticated “technology-enhanced” and simulation items without having to engage software developers in the tasks.

The power and flexibility of AIR’s approach to using technology for educational measurement have enabled us to implement virtually all of the technology-enhanced item types identified in both consortia’s specifications (Smarter Balanced) or examples (PARCC). In fact, when Smarter Balanced decided to deploy sample items, some of its more interesting examples strayed from or extended their specifications in important ways; AIR’s technology was able to immediately implement these extensions without involving software developers. Smarter Balanced’s Director of Mathematics wrote to us:

I wanted to thank you and the AIR team for the incredible effort in putting together the innovative items for the upcoming Smarter Balanced sample item release. The speed with which your team was able to turn ideas into items is commendable. We displayed the items to a live crowd tonight of state and higher education leads and they were thrilled with the results.

We are excited to think about what is possible over the next several months given what was accomplished in two short weeks. On behalf of the consortium, please give my thanks to the AIR team. (email from Shelbi Cole, September 11, 2012)

AIR’s item bank is the item bank being used by the Smarter Balanced Assessment Consortium until its item bank is ready, and our item bank will serve all of the item banking and test construction functions for the Smarter Balanced pilot test.

Below, we describe our item bank, which is a pure Internet solution over secure connections.

ITS serves at least three roles:

- A content development and management tool
- An item bank
- A publication system supporting both paper and online publication

The item development workflow guides items from inception through a series of content, fairness, graphic, and other reviews to final publication. 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.

When an item is published, ITS tracks its use on a form or in an adaptive item pool. After an item is used, the ITS stores the resulting statistics, including exposure statistics, classical item statistics, and statistics based on item response theory.
Technology-Enhanced, Machine-Scored Items

First in the industry, AIR has introduced technology that enables test developers—content experts, editors, and graphic designers—to develop complex, machine-scored items, including their scoring rubrics, without the involvement of software developers (although flash-based or HTML-5 simulations or item stimuli require graphic artists with animation skills for display). This technology represents the items and their scoring rubrics in a declarative format (i.e., as data rather than programs), and generalized display and scoring engines read the data to deliver the items to students and to score student responses.

ITS contains tools that empower test developers. We recognize that the expertise in both assessment and the content rests with test developers, not software developers. Therefore, ITS enables them to point-and-click their way to finished items, scoring rubrics, and simulations. Exhibit I.C.1-1 presents a collage of screenshots from some of these tools. We are happy to demonstrate these tools for USOE.

The benefits of our approach are threefold. First, by eliminating software developers from the mix, we keep item and rubric development costs comparable to those of paper and pencil items and rubrics. Second, we expedite and simplify the development process because our content experts can work directly with USOE content experts to create items without having to translate the content to a format that can be programmed by software developers. Finally, AIR makes available our XML format and interpretation rules, so any test delivery system can implement the software necessary to deliver the items. From the perspective of Utah, this process ensures that you can change testing contractors, share items across states, and otherwise maintain your investment over time.

Integrated Accessibility and Accommodation Development

Particularly when developing items for an adaptive test, integrating transcription notes and other accessibility features with the development process smoothes the workflow. Unlike a fixed-form test in which only a small portion of the item bank is published and made accessible, an adaptive test will have hundreds of items per grade/subject, any one of which might be administered to any student. The development of transcription notes (a process that we call tagging for web-delivered items) proves most tractable when it is done upon approval of the base item. That way, tagging is...
accomplished (and reviewed by USOE) on a flow basis instead of requiring the review of thousands of items at the same time.

Currently, our tagging includes translation, text-to-speech, and text-to-Braille. The system is capable of delivering recorded audio or video, including sign language.

**Summary**

ITS helps simplify and enforce the workflow for items as they move from conception to publication. The system includes test-developer-friendly tools for the development of sophisticated technology-enhanced, machine-scored items, which reduces costs, increases flexibility, and facilitates collaboration on innovative items. Our system also supports the integrated development of item-specific accommodation information, including language, audio, and video accommodations.

**Item Inventory, Item Banking, and Content Management**

ITS is completely configurable, including templates for new item types, the particular metadata stored, the software-enforced review sequences, and item status identifiers.

We will work with USOE to determine the best configuration of ITS to meet the requirements of each assessment. AIR’s ITS integrates a full-featured item banking system. This fully configurable system implements specifications such as input templates, output (destination) templates, XML-based content storage and representation of graphics and other binary objects, item attributes, review histories, item statistics, and item use histories.

**Review of Current ITS Configuration With USOE**

ITS is an integral component of all our assessment programs, AIR proposes to conduct a meeting with USOE to review and refine the configuration of ITS. During this meeting, AIR will provide USOE with examples of the configurable modules so that collectively we can decide what configuration will best serve the needs of USOE and your testing programs.

The information that follows gives more specific examples of the types of modules that can be configured.

**Input Templates**

Input templates capture an item in “parts,” including the stem, the key, the distractors, and, for constructed-response items, the rubric, including examples of responses for each potential score point. We will work with USOE to design additional templates if necessary to support your assessments.

**Destination Templates**

Each item bank specification includes multiple destination templates. Destination templates describe the format in which items will be rendered. These templates are customized for each program and may include, among others, item review templates, which print the item along with selected item attributes; committee review templates, which present an item’s attributes on the left-hand page and the item in very much the same format that will appear on the test form on the right-hand page; data review templates (sometimes called item cards) that present the item statistics on the left-hand page and the item on the right-hand page; and form-layout templates, which enable ITS to directly generate items into InDesign format for final test formats and online lay-
outs. This automated process eliminates many opportunities for human error to be introduced into the test production process.

**XML-Based Content Storage**

AIR’s item bank stores all the item content in XML, with text (e.g., passages, stems, distractors) directly in the XML document and binary documents (such as graphics) represented as links to files within the XML. Integrated version control ensures that the most current version of every element of an item or a stimulus is accessed. All prior versions are maintained in an archive and can be recalled for comparison with the current version. The system enforces the documentation of changes to each element of an item.

**Item Attributes**

ITS is configured to store specific information (meta-data) about items, called attributes. Attributes often contain information about the academic standard alignment, language load, cognitive complexity, distractor and key rationales, or other characteristics of items. Every item is stored along with its set of attributes. Attributes are completely configurable and can be added, modified, or deleted at any time (costs may be incurred if new information must be gathered about a large number of items when adding attributes). We will work collaboratively with the USOE to identify the attributes for items and stimuli that will meet your current and future needs.

**Tagging Items for Accessibility**

Items in ITS can be tagged for accessibility, such as for text-to-speech, translations, and text-to-Braille.

**Test Construction Process**

ITS supports and automates the test construction process. As we described previously, each item moves through a customizable sequence of development and review steps.

ITS not only manages the content of the test items but also incorporates tools and processes for constructing the test for an online administration. Form Builder, a component of ITS, allows content experts and psychometricians to collaborate in the construction of a form or a pool of items to be published to an adaptive test. Form Builder facilitates the review of the form in many ways, including statistical review and review of item use histories. Another module, Blueprint Tool, represents the blueprint for adaptive tests.

**Blueprints (Test Specifications)**

The test development process starts with blueprints. ITS incorporates a module for defining blueprints, which may include both content constraints from the standards hierarchy, and other constraints defined by any of the meta-data available for the items. It is typical for blueprints to contain specifications of minimum and maximum numbers of items at each content level, targets for each item type, and targets for depth of knowledge.

In collaboration with USOE content experts, AIR will propose blueprints to be stored in ITS. We note that as with all of AIR’s systems, the content standards are accessed from a single source—our Standards Repository, a necessary element for seamless interoperability of the components of our system.
**Item Pool Construction (Form Building)**

ITS has a “form building” module for the constitution of fixed form tests or the assembly of item pools for adaptive delivery. The form building tool, Form Builder, allows item developers to associate items with common stimuli where necessary and assemble those items onto forms.

Our adaptive algorithm will support adaptation within groups of items grouped under a common stimulus (such as reading items grouped under a passage). Form Builder allows test developers to identify the number of items to be administered from each group. For example, a test developer might associate 15 reading items with a given passage, and use Form Builder to indicate that each student only sees (say) seven of these items. This characteristic of the item pool allows the adaptive algorithm to select the seven items that best match the blueprint and the student performance on the test. This parameter can be set separately for each item group.

ITS can package item pools for use in the adaptive algorithm simulator (see Section I.C.3.c) or the real test, as necessary.

**Field-Test Item Selection**

The field-testing under this contract will be conducted online.

Field-test item selection is a simpler process for the online test because editorial issues do not arise—by the time items are eligible for selection onto field-test forms, they have already received the online analog to blue line approval. In Section II.B.2, Item Development, readers will learn about web preview. This step in the item development process causes the final lockdown of content to display on the test. Approved items are eligible for selection into the field-test pool. Items will be selected to optimize the functioning of the algorithm—first building robust pools to ensure that many unique, conforming tests can be generated. As the pool matures, it becomes possible to extend the range of item difficulties within and across content categories.

**Preparation of Tests for Publication**

Below, we describe the processes for preparing a test for publication and publishing the test. We begin with the online test, focusing on adaptive tests.

**Preparing an Online Test for Publication**

Exhibit I.C.1-2 displays AIR’s internal process for preparing a test for publication and encompasses the entire test construction process. The graphic color-codes the team responsible for each step.

The Test Delivery System (TDS) requires two inputs for each test—content and configuration information. Content is represented as a set of XML files and associated media files. The configuration file is actually a set of tables in a relational database. We focus first on the development and deployment of the configuration file.

The exhibit shows the process after we have received USOE’s approval for the blueprint and the item pools. Our lead psychometricians and item development manager must each sign off on the pools and blueprints to confirm that they conform to client expectations.

The pink box at the top center represents our configuration generation process. This highly automated process

- transforms the blueprint into the form required by our TDS;
- gathers configuration information;
checks that every item on the form has been web-approved and that all media files are available;

cconducts a series of internal consistency checks;

provides a preliminary indication about whether the item pool will support the blueprint;

compares the configuration file with any prior version; and

produces a report including a differential comparison to previous version if applicable, a list of any internal inconsistencies, warnings about missing items or media files, and warnings about likely item pool adequacy.

The tables that make up the configuration are built on a test server, and the report is delivered to relevant staff members. The project director, item development manager, and lead psychometricians must approve the report before the configuration file is deployed.

Upon signoff, the configuration file is deployed to the simulation, staging, or production server, and a secondary process runs that compares the configuration file deployed with the one on the test server. These should be identical. In the unlikely event that they are not identical, the deployment is rolled back and the differences are resolved.

Before any deployments to production, we run an extensive and highly automated set of simulations. The simulator allows psychometricians to adjust parameters within 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,
items on each content standard or with each characteristic. 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 USOE’s decision.

Psychometricians interact with the simulator by modifying the configuration parameters in a spreadsheet. When the configuration parameters are finalized, this same spreadsheet 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 changed and that procedures were followed. USOE will receive a final report on all simulations containing the detailed statistical output. We encourage USOE to send your psychometric staff (at our expense) to participate in this process with us.

Deployment, the pink box on the bottom right, is similarly automated. When the configuration file is generated, the content is also gathered to a holding space on our servers. On deployment, which is always done during off-hours, the content is loaded on the production servers and the configuration information is loaded into the production database, which is then compared with the configuration data on the staging database. Any differences are resolved.

We typically deploy to a staging server first, and that server is the platform for user acceptance testing (UAT) and, should USOE desire, final approval. On the staging server, USOE can test the full functionality of the test. Because the test is adaptive, it is difficult to check the content of the test. Therefore, on the staging server we also deploy fixed-form versions that include the entire item bank. These, of course, are only for quality assurance purposes and do not appear on the operational site.

USOE may or may not choose to check the content at this stage. Recall that our development process includes a web preview review, in which USOE views the item exactly as it will appear to students. Note that our ITS and TDS both use exactly the same display engine, and web approval causes the display code to be stored for deployment. This is a direct analog to blueline review. Additional review after deployment to the staging server is more analogous to checking the first run from the printer for a paper test.

**Summary**

Above we described ITS, and how it supports the full lifecycle for items and tests. ITS

- facilitates and manages item reviews by both internal and external reviewers, ensuring that every item undergoes every review;
- enforces a formal approval is recorded at each level of review. In the section on item development, readers will learn about our Web Approval level, which “freezes” the item as it appears for delivery on tests;
- supports test construction, from the definition of blueprints and item pools through publication and deployment;
- incorporates both workflow management and real-time item inventories for managing the content pool;
- records every change to each item, along with a rational for the change, storing each version.
- supports, in conjunction with our administration simulation tool, item pool planning pools specifically designed for management and planning of pools for standards-based adaptive tests; and
supports all activities necessary for full item lifecycles for both summative and interim assessments.

ITS can import items from QTI format and will do so. We will import items from any source. We can support machine-readable formats other than QTI in addition to QTI as long as they separate the logical components of items in a readily format that can be parsed and reformatted into QTI, and provided that they contain sufficient information to display and score the item.

Where items are stored in formats that cannot be readily parsed (for example, non-templated MS Word files), AIR will be pleased to separately evaluate and price the effort required to transform the items into QTI format for import.

The Item Bank supports the mandated parental review, and this committee process is described in Section II.B.2.
I.C.2 Test Administration

Below, we outline the processes and procedures that will support the delivery of the new assessments. We specifically address:

- testing times and our support for flexible testing schedules, including pausing and resuming tests;
- training, including the early communication that will help Utah transition to this new testing system with buy-in and support from key stakeholders, and with each party understanding their responsibilities and benefits from the new system;
- administration practices and procedures (requirements);
- help desk support; and
- our proposed approach to paper-based administration.

I.C.2.a. Test Administration Times

The testing times derive from the blueprints, and Part II, Item Development, describes our proposed measurement approach and blueprints in detail. As we discuss there, the blueprints will be developed in partnership with USOE and do not reflect an off-the-shelf product. Appendix A offers draft blueprints for each of the assessments. These blueprints include items of all types, including a broad range of machine-scored items, covering the complete depth of knowledge and range of content in each subject area. We propose to use machine-scoring, including a variety of artificial intelligence approaches, and human scoring. This is described in Section I.C.3.d, Scoring.

In order to cover the full range of content and reach the required depth of knowledge, we propose to include one performance task in each content area. The adaptive test, including the performance task, in each subject area should take approximately 90–120 minutes to complete. Performance tasks, samples of which appear under Item Development, are designed to take approximately 20–45 minutes to complete.

Our proposed draft blueprints cover the entire breadth and depth of the content standards. These blueprints, however, are drafts. We intend to work with USOE content experts to refine these drafts to develop a customized blueprint that will exactly match Utah’s needs.

AIR generally recommends that our clients allow all students to pause and resume tests as necessary. Security features (discussed below) protect the integrity of the test while allowing this flexible administration. This feature enables each school to tailor testing sessions to its own bell schedules, allowing assessment to work more harmoniously with instruction. This universal design feature also naturally accommodates students who require more frequent pauses in their tests, as well as students who tend to require additional time.

To maintain the security of the assessment:

- Students can be prevented from pausing a test while unanswered items are on screen (we recommend allowing the test administrator to pause tests at all times).
- Security periods can be established so that a student whose test is paused for more than a typical bathroom break, for example, can be prevented from navigating back to questions answered before the test was paused.

These features help prevent abuse of the flexibility offered by our pause-anytime approach. We note that the system can also prevent pauses, if Utah prefers such a policy.
I.C.2.b. Training and Communication

Providing comprehensive, ongoing training throughout the year on the new online testing system for Utah educators and administrators is essential to success. Below we propose an annual training plan that includes a series of interactive presentations on each of our online systems to be delivered through a combination of webinars, self-guided and self-paced online tutorials, face-to-face trainings, and train-the-trainer sessions. Our experience transitioning large-scale assessment programs in , and has also taught us that clear, early communication about the new online testing system is an important training component. While training teaches stakeholders the hands-on skills needed to use each component of the new online testing system, communication lays the foundation for what to expect: How does the new online testing system differ from the old testing system? What kinds of training will be provided and when? Will my child have more than one opportunity to take the test and, if so, how will his or her final score be determined? Addressing these questions early on, in addition to having a comprehensive training plan, has proved effective in calming concerns and building critical stakeholder support. Below we discuss our proposed plan for early communication and our plan for ongoing, interactive training in more detail.

Proposed Plan for Early Communication

Transitioning a large-scale assessment program to online adaptive testing typically produces an enormous amount of information. The objective of an early communication strategy is to provide stakeholders with a clear structure for organizing and absorbing the information they will be receiving. We propose to accomplish this by collaborating with USOE to customize two key communication tools for Utah parents, educators, and administrators: first, we propose to work with USOE to develop a customized information portal for all Utah educators, administrators, families, and students that functions as a central repository for all communication and training materials. Second, we propose to collaborate with USOE to develop a series of informational brochures for each user role that answer common questions about the new online testing system.

1. A Customized Information Portal for All Stakeholders. The structure of the portal may be organized according to role (Test Coordinator, Teacher, Technology Coordinator, Student/Parent) or by 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 home page 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 USOE to configure the portal in the most optimal way for Utah educators and administrators.

2. A Series of Informational Brochures for All Stakeholders. AIR proposes to collaborate with USOE to develop a series of customized one- to two-page, color informational brochures for parents, educators, administrators, and technology coordinators. The brochures will answer questions about the new online system: How is the new online testing system different from the old testing system? Will students be
required to take a test multiple times? What guidelines can schools use to schedule tests using the new online adaptive testing system? We include sample brochures from in Appendix B, and in Exhibit I.C.2.b-1 propose a list of brochures for Utah stakeholders. We propose to deliver the brochures in PDF format, which is appropriate for delivery via email or websites.

**Proposed Annual Training Plan**

Below we propose an annual training plan that includes a series of interactive presentations on each of our online systems to be delivered through a combination of webinars; self-guided, self-paced online tutorials; face-to-face trainings; and train-the-trainer sessions. The content of the webinar presentations, as described below, would largely be the same as the face-to-face training presentations and will form the basis for the online tutorials. AIR’s online tutorials are designed to teach both sophisticated technology users as well as users new to the system 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 tutorial they find useful, new system users typically find it useful to have instruction at a very detailed level. The online tutorials will also provide follow-up training throughout the year to educators and administrators who require it. In addition, the online tutorials can be developed to include quiz questions to provide the users with a gauge of his or her knowledge after the training. USOE may choose to develop one 15–20 minute tutorial per online system or may choose to develop three to four separate tutorials of approximately three minutes each to more easily allow users to choose the specific system feature they would like to learn more about.

A detailed description of proposed training presentations and materials in Exhibit I.C.2.b-2 is provided on the next page:

1. **User Roles Chart.** This reference document indicates, at a glance, the access each user role is permitted in each online system. (Please see the sample user role chart in Appendix B.) System access for each user role will be reviewed in each webinar as appropriate. The user role chart will be posted on the portal; will be available in the manuals; and will form part of the online, self-guided tutorials as determined in collaboration with USOE.

---

### Exhibit I.C.2.b-1: Summary of Proposed PDF Brochures

<table>
<thead>
<tr>
<th>Primary Audience for Brochure</th>
<th>Informational Brochure Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educators/Test Administrators/LEA Assessment Staff</td>
<td>Overview of the New Online Testing System and What It Means for Educators and Test Administrators</td>
</tr>
<tr>
<td>Parents/Students</td>
<td>Overview of the New Online Testing System and What It Means for Students and Families</td>
</tr>
<tr>
<td>District/School Administrators/LEA Assessment Staff</td>
<td>Preparing and Training Staff for Online Adaptive Testing</td>
</tr>
<tr>
<td>Educators/Test Administrators</td>
<td>Scheduling Online Test Sessions Worksheet</td>
</tr>
<tr>
<td>Technology Coordinators</td>
<td>Minimum Technical Requirements for Online Testing</td>
</tr>
<tr>
<td>District Data Coordinators</td>
<td>Preparing to Upload Student Information for Online Testing</td>
</tr>
<tr>
<td>Principals, Teachers, Test Administrators, Data Administrators</td>
<td>Quick Guide to User Roles</td>
</tr>
</tbody>
</table>
## Exhibit I.C.2.b-2 Proposed Presentations and Formats for Annual Training

<table>
<thead>
<tr>
<th>Presentation Title and Primary Audience</th>
<th>Proposed Training Topics</th>
<th>Proposed Training Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Phase I: Preparing for Online Testing</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| User Roles: Principals, Teachers, Test Administrators, Data Administrators, Technology Coordinators | User roles: who does what and in which system | 1. User role chart  
2. Will be a component of each webinar and face-to-face training |
| Uploading the Pre-Load File: Data Administrators | Learn how to upload student information into TIDE and to add and delete students | 1. Webinar presentation  
2. Face-to-face training presentation  
3. Online tutorial on TIDE |
| Preparing Your School’s Computer Labs: Technology Coordinators | Steps for secure browser installation and minimum hardware requirements | 1. Webinar presentation  
2. Instructions on portal |
| **Training Phase II: Administering Online Tests** | | |
| Online Test Administrator Certification Course: Test Administrators, Proctors, Teachers | Learn how to use the new online testing system, set up a test session, manage and monitor testing, assist students with online tools. Certification earned only by passing a mandatory quiz | Online Test Administration Certification Course |
| Online Practice Tests: Test Administrators, Educators | Practice setting up, pausing, resuming, and ending live test sessions; setting accommodations; using the same functionality as the operational Test Administrator site | Online practice tests |
| Online Practice Tests: Students, Parents, Educators | 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 strikethrough and answering all item types | 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 (in addition, learn how to reset and invalidate tests) | 1. Webinar presentation  
2. Face-to-face training presentation  
3. Online tutorial |
| **Training Phase III: Post Testing** | | |
| How to Use the Online Reporting System: Teachers, Principals, District Administrators | Learn how to access student scores, create class rosters, and drill down from district and school results | 1. Webinar presentation  
2. Face-to-face training presentation  
3. Online tutorial |
| How to Use Learning Point Navigator: Teachers | Learn how to identify instructional resources for your class, create student assignments, and generate reports | 1. Webinar presentation  
2. Face-to-face training presentation  
3. Online tutorial |
2. **Webinar for Data Administrators:**

*Uploading PreLoad Files.* This training includes instructions on how to upload preload files with student and school information, file formats, adding students who were not in the initial preload file, deleting students, moving students to another school, and setting student test eligibility and student accommodations.

3. **Webinar for Technology Coordinators:**

*Preparing Your Schools’ Computer Lab.* This training includes information for school and district network administrators and provides guidelines for varied technical setup 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.

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 that course-takers have learned the skills necessary to proctor tests using the new online system, each user must pass a quiz at the end of the course 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.

5. **Webinar for Test Administrators: How to Start and Monitor Testing and Modify Test Settings.** This training includes the same content as the Online Test Administrator Certification Course but is presented in a webinar and face-to-face presentation format.

6. **Online Practice Tests.** Online practice tests for all 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.

7. **Webinar on How to Use the Online Reporting System.** This training teaches authorized users how to use two types of reports: test management reports and performance reports. Users are shown how to use test management reports to monitor student testing status for their schools or districts and to identify students who have and have not tested. Users also learn how to use these reports to help schedule test activities for their school. Turning to performance reports, 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; filter score reports by student subgroup (e.g., gender, ethnicity, ELL); and manage online rosters (groups).

8. **Webinar How to Use Learning Point Navigator.** This training includes instruc-
tions for identifying and using supplementary instructional resources to help educators address areas of weakness at the classroom level. Users are shown how to create individual and class assignments using targeted instructional resources and generate reports on the status of the assignments.

Each year, AIR will collaborate closely with USOE to update the webinar presentations, face-to-face training presentations, and online tutorials as needed. Follow-up training during each year will be achieved through the availability of online tutorials year round and by offering a second round of repeat webinars, if needed, later in the year. In the first year, we propose to focus all presentations on the basics of the program and the features of the online system. In subsequent years, AIR proposes to focus training on the updates to each system and providing basic refresher training on each system. Each year by March 1, a training plan and schedule will be provided for the summer. A proposed timeline for Year 1 training can be found in Exhibit I.C.2.b-3.

**Train-the-Trainer Sessions**

AIR will offer a series of train-the-trainer sessions during which two or three representatives from each district will attend a full-day session and receive the instruction and materials needed to train additional school personnel on the new online testing system. AIR will conduct full 1-day train-the-trainer sessions annually and will work with USOE to determine locations for this regional training. During testing, these representatives will serve as resources in the district, answering questions and demonstrating the system to faculty and staff as needed. To minimize scheduling impacts, AIR proposes to provide this training during existing USOE or district professional development sessions.

In addition, AIR proposes to develop an online training module on how district and school

**Exhibit I.C.2.b-3: Proposed Timeline for Year 1 Training**

<table>
<thead>
<tr>
<th>Training Type/Title</th>
<th>Proposed Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. User Roles Reference Chart</td>
<td>6–8 weeks before testing window opens</td>
</tr>
<tr>
<td>2. Webinar for District Administrators</td>
<td></td>
</tr>
<tr>
<td>3. Webinar for Technology Coordinators</td>
<td></td>
</tr>
<tr>
<td>4. Conduct Onsite, Face-to-Face Trainings</td>
<td>4 weeks before testing window opens</td>
</tr>
<tr>
<td>5. Online Test Administrator Certification Course</td>
<td>3–4 weeks before testing window opens</td>
</tr>
<tr>
<td>6. Online Practice Tests Available</td>
<td>3–4 weeks before testing window opens</td>
</tr>
<tr>
<td>7. Webinar for Test Administrators</td>
<td>3–4 weeks before testing window opens</td>
</tr>
<tr>
<td>8. Webinar on Online Reporting System</td>
<td>3–4 weeks before testing window opens</td>
</tr>
<tr>
<td>9. Webinar on Learning Point Navigator</td>
<td>3–4 weeks before testing window opens</td>
</tr>
<tr>
<td>10. System-Specific Tutorials Available for Online Testing System, TIDE, Online Reporting System, and Learning Point Navigator</td>
<td>6–8 weeks before testing window opens</td>
</tr>
<tr>
<td>11. Conduct Train-the-Trainer Sessions</td>
<td>3–4 weeks before testing window opens</td>
</tr>
</tbody>
</table>
staff can train test administrators. This training would present step-by-step instructions on how to train test administrators to administer online tests using the new system. As an extra cost option, USOE may choose to produce CDs with train-the-trainer kits composed of webinar PowerPoint presentations for each system, user guides for each system, informational brochures, and other printed materials as well as PowerPoint instructions for trainers.

**Face-to-Face Trainings**

Face-to-face onsite training will consist of five full-day live workshops annually. Up to 500 Utah educators will participate in the onsite face-to-face training, each attending a comprehensive one-day session with breakout groups addressing the topics described above in an interactive manner. The training locations will be held throughout Utah on up to five dates chosen by USOE. AIR will bear all costs for this training.

**I.C.2.c Test Administration Requirements**

**Overview of the Testing Process**

AIR recognizes that the testing process is managed by educators and that every minute spent on administrative tasks is a minute taken from educating students. Therefore, we have designed a testing process that minimizes the administrative and technical burden on school personnel. AIR's two key innovations that reduce administrative burden without sacrifice are

- implicit registration; and
- a pure Internet solution.

Implicit registration means that students do not need to be explicitly registered for a testing session. As described below, through our Online Administration Interface, we have eliminated the clerical task of assigning students to sessions. Instead, at the time of testing, the test administrator (TA) starts a testing session with as few as two mouse clicks. The system assigns a session ID, and student login is simply a request to enter a session. Students log in with identifying information and the session ID. From his or her workstation, the test administrator authenticates the students and admits them to the session. In this way, the roster of students to be tested builds itself as students log in. Nobody has to sit at a computer creating these rosters.

Implicit registration is supported by our data exchange process, which is detailed in Section I.C.3.b. Through completely automated systems, student information systems exchange information with our systems, so that each student is known to the system. Student eligibility to take a test is determined implicitly through a series of configurable rules. For example, a simple eligibility rule might state that students who (a) are in grade 5 and (b) have not taken a grade 5 math test are eligible to test. Below, we discuss further data exchange, eligibility, and implicit registration.

The second innovation that reduces burden on schools is the implementation of a pure Internet solution. Our system has no need for proctor caching or local caching servers. Such caching mechanisms require extra hardware at schools, intermittent reinstallation, and a higher level of technical support. Our system is purely Internet-based, robust, and nearly maintenance-free. Before the beginning of the school year, schools install our secure browser. The secure browser is simply a special build of the Mozilla (Firefox) browser that is modified to protect the security of the test and support certain accommodations. It can 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. Once the browser is installed, it generally does not change throughout the school year. Changes or updates to the test happen on our servers and are transparent to the schools.

The pure Internet solution is robust. Every time a student answers a question, the response is saved to our servers. For longer responses such as essays, the system can be configured to save periodically (say, every minute) or whenever the student presses the “save” button. If the system is unable to reach the server, the student is stopped from testing, so no work is lost. The system is designed to make multiple attempts to reach the server, so 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.

Caching occurs on the students’ machines (in memory, not on disk) in real time. This ensures a seamless testing experience in which students typically see no delays between items.

Contrast this approach with systems requiring caching servers. The caching servers sit in the schools or districts, providing many additional points of failure. The likelihood of failure of such servers is much higher than the likelihood of failure of our industrial, fully redundant servers hosted in professional, secure hosting facilities. Hence, a caching server approach requires more technical expertise within schools or districts, requires more maintenance, and increases the likelihood of failure. Our ability to deliver tests without extra hardware or expertise has contributed to our record of success wherever we implement online tests.

Test Administration Plan

AIR will deliver an annual test administration plan. The plan will cover

■ communication;
■ user management, enrollment, and registration;
■ preparation for testing, including technical preparations and training;
■ test administration;
■ scoring; and
■ use of results for accountability, teaching, and learning.

Communication begins well before testing, as the new testing system is introduced. It is important to educate educators, learners, families, and other stakeholders about computer-adaptive testing (CAT) before moving to the system. Section I.C.2.b above outlines our plan for communication and testing in preparation for the implementation of the system.

Implicit registration, described above, requires that students be known to the system before testing. Student data are loaded into our systems through automated or user interfaces. We typically establish a secure FTP site into which statewide data systems deposit updates to enrollment files each night. Similar data exchanges can be established at the district or school level as well, depending on USOE’s preference and the capabilities of statewide systems. We also offer an option for schools or districts to directly upload enrollment information using a spreadsheet template or a fixed-file format. These processes are discussed at length under Data Exchange.
Preparation for testing happens at three levels:

- Technical preparation
- Training
- Data exchange

Technical preparation generally involves only installation of the secure browser (or secure browser app in the case of tablets). The secure browser is typically released in July and installed only once before school starts. Installation is discussed under Technical Requirements, but many low-effort mechanisms for installation are offered, and we support many different configurations, such as thin clients.

Although the Test Administrator Interface is simple and intuitive, we recommend that USOE require all test administrators to take a customized online training course. This interactive course, which takes about 20 minutes, teaches administrators to operate the software, builds confidence to avoid frustration in the real test setting, and offers an opportunity to remind test administrators of USOE policies. We recommend that all students participate in practice tests and use item-type tutorials as necessary, prior to the day of testing. This practice ensures that students are familiar with the interface before they test.

Data exchange, discussed above and below, is the process by which our systems learn about the students who may be testing. Data exchange must take place prior to testing, and we support near-continuous updates to data, at USOE discretion.

The introduction to this section provided a brief description of the test administration process, and we discuss that further in the next section.

Scoring happens immediately for all items that are machine-scored or partially machine-scored. Section I.C.3.d, Scoring, details our plans for scoring test items. There we discuss a plan to machine-score most items and use automated scoring (AI scoring) for some longer written-response items, along with a plan for targeted human review of a substantial number of cases.

The process does not end when the test is scored. The data flow immediately into our Online Reporting System. The reporting system enables educators to explore their students’ performance on the test. As discussed at length in Section I.C.3.e, the reporting system provides a significant amount of diagnostic information, particularly at the classroom and student group levels, to help teachers focus and differentiate their teaching.

The data are integrated with Learning Point Navigator, our instructional support and formative assessment system. Educators can navigate directly from relevant findings in the reporting system to formative assessments and learning activities that will help them probe into which students need extra help. Learning Point Navigator is described in detail in Section III, where we describe our proposal for formative assessment.

**Tools Available to Monitor Testing Progress, Update Student Data, and Ensure Accuracy**

Our Test Delivery System interface for test administrators enables TAs to control all aspects of administration in a test session. This system can be configured according to USOE preferences. The TA Interface allows administrators to securely approve students for testing, test multiple grade levels and subjects in a single session, update individual student accommodations, and pause and resume testing for individual students.
Our testing model does not require schools to preregister students for tests. Configurable eligibility rules "calculate" the tests that students are eligible to take on the basis of data uploaded through our Test Information Distribution Engine (TIDE), discussed in the next section, Test Delivery System, under Data Exchange. For example, students may not be allowed to test out of grade level, or an uploaded attribute may be required to indicate a student’s eligibility to “challenge up.”

When students log on to a session, they are offered a choice of tests to take. Each student sees an individualized choice of tests for which he or she is eligible, for which the student has not exhausted the opportunities, and that the test administrator is offering in the session. Hence, a single session may contain a class of grade 4 students, some of whom are completing a math test and others who are starting a reading test. A grade 5 student may have joined the group to retake a science test as well. The grade 5 student would see only the grade 5 science test as an option, while most of the grade 4 students would see reading and math (assuming that they had opportunities left for these subjects).

TIDE contains all student demographic and required accommodation information, and states, districts, and schools can upload and edit this information according to the rules set by USOE. TIDE can process information that test administrators enter directly through the user interface or that is uploaded through fixed-format files or spreadsheets. The system supports configurable validation rules that evaluate and enforce sophisticated, client-specific rules for uploaded information. Different levels of validation enforcement are available, ranging from giving a simple warning to rejecting a record to rejecting entire files. When uploading information, the system logs and reports back any data that violate the validation rules.

Test Delivery System: Test Administrator Interface

Test administrators use the TA Interface to create and manage test sessions. The 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, which is summarized in Exhibit I.C.2.c-1. Students can test only when the administrator is logged on and has a session active. After a stu-
dent pauses a test, the TA must re-admit her or him to the session.

**Creating a Test Session**

Test administrators create a session simply by selecting the tests that will be available in the session. When students join the session, they can select from among the tests that (1) are offered in the session and (2) they are eligible to take.

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 must admit them to the session.

Once admitted to the session, the student is given a final chance to confirm that he or she has requested the correct test. Upon confirmation, the test begins.

**Test Settings and Tools for Administrators**

During the authentication process, test administrators have an opportunity to review or adjust the “test settings.” Test settings include any embedded supports or accommodations to which the student may require access.

The availability of access features, embedded supports, and accommodations is completely configurable. USOE 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. USOE will have an opportunity to determine which roles have authority to adjust specific settings and to adjust them through file uploads.

Exhibit I.C.2.c-2 shows our online interface through which test administrators review or adjust individual student accommodations before approving a student for a test.
The TA Interface provides tools for monitoring test progress and ensuring that each student has access to the appropriate accommodations. The sample list of accommodations in Exhibit I.C.2.c-2 is not exhaustive, and AIR will work with USOE to develop an appropriate list for Utah students and administrators. Although students can be preregistered for specific accommodations, the TAs can also change these accommodations, such as the language of the test, the default print size, or other accommodations that Utah may select.

We generally recommend that clients offer a limited number of accessibility tools to all students through the student interface. Instead, we recommend having adults set these accommodations, rather than allowing students to select accommodations from a list. This reduces distractions and allows students to focus on the primary task of demonstrating their knowledge. The list of accommodations available to TAs is completely configurable, and USOE may opt to limit or expand the authority of TAs to set accommodations according to state policies.

Monitoring the Test Session

Once testing begins, the TA can monitor student progress and, if necessary, stop one or all students from testing. The system is designed to automatically notify a TA of events that require 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 I.C.2.c-3 shows the TA interface during an active test session with five approved students.

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 at an earlier time.

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 for authorization to fulfill the request and to retrieve the paper from the authorized printer.

Workflow Summary and Resolving Anomalies

We have described the general workflow above. Exhibit I.C.2.c-3 illustrates the key activities associated with the periods before, during, and after testing.

Each step in the system has automated procedures in place to detect anomalies. We discuss each of these below.

Detecting Anomalous Data During Data Exchange to AIR Systems

As we discuss below, the TIDE system incorporates a powerful and flexible capability for validating data on input. We will work with USOE to identify validation rules and the actions to take when records or files violate these rules. Such actions may include issuing a warning, rejecting a record, or rejecting an entire file.

Detecting Anomalies During Test Administration and Scoring

As described above, AIR’s system includes QM, the quality monitoring component of the Test
Delivery System. As each test is completed, the entire test is rescored, data are gathered about the item responses and test composition, and a series of quality control checks are made. These checks happen in milliseconds and issue warnings or failure notices to which our project staff immediately attend. This feature ensures that if anything goes awry, we can address it before bad data flow downstream.

**Handling Human-Reported Anomalies**

In every state, we sometimes encounter issues where one student tests under multiple state identifiers or multiple students use the same state identifier. Often this can be detected only by the educators who see unexpected results. When these reports come in, AIR’s help desk has tools to address and correct the issues. These are some of the available corrective measures we take:

- Merging or partially merging records associated with different identifiers
- Splitting tests taken under the same identifier across two corrected identifiers
- Invalidating tests, effectively eliminating the student’s opportunity to take that test (like a paper test invalidation)
- Resetting a test, effectively restarting a test for a student who may have started the wrong test or for some other reason requires a reset
- We note that the system will prevent students from being presented with items they had seen on an invalidated or reset test.

**An Overview of Our Continuous Improvement Process**

Whenever issues arise, AIR conducts a root cause analysis. When we have identified the root cause of the problem, we implement software, policies, or procedures that ensure that the problem will not be repeated.
In addition to this “reactive” mode, two proactive annual activities mitigate risks:

- AIR’s senior leadership team meets annually to review risks and identify procedures or software tools to help mitigate those risks.
- The leadership of AIR’s Computer and Statistical Sciences Center meet annually to review software and procedures to identify improvement opportunities.

Opportunities for improvement and associated plans to exploit those opportunities, including investment in staff, software, and hardware, are generally finalized before January 1. We are happy to share our continuous improvement plans with our clients, with the understanding that these plans represent sensitive AIR-confidential information.

**Supporting Changing Technology and Phasing Out Dated Technology**

To date, AIR has committed to supporting new versions of supported operating systems that are released by May 1. This is driven by our typical delivery of the annual secure browser for each operating system in July. Apple, however, often releases new versions of its operating systems late in the summer or early in the fall. We will work with Utah to agree on a support plan for these operating systems that are released later in the cycle. For example, last year we released a new secure browser supporting the new OSX (10.7) systems in January. We plan on a similar release this year.

Users can access systems other than the secure test using other browsers. We try to support new versions within 45 days of release. However, recently Mozilla and Chrome have adopted a very frequent release schedule. Again, we will work with Utah to agree on an appropriate support schedule.

For older systems, AIR has adopted an “End of Support Plan” for operating systems running our secure browser to provide clear guidance to districts planning hardware purchases. This plan helps schools and districts manage support requirements and allow both product planning and information technology planning in organizations based on knowledge of the support timelines. This plan also helps leverage advantages that are available to newer technologies that are otherwise not possible with a requirement to support legacy platforms.

AIR will support an OS for 10 years after release or until the software company stops supporting its OS, whichever is earlier. Decisions regarding whether to discontinue support for an OS will include stakeholder input. Status changes take effect every summer to avoid disruptions in testing during the testing window.

In instances where AIR continues to support an operating system after its manufacturer has discontinued support, we may provide limited support. When providing limited support, we will not provide updates to the predeveloped secure browsers. User support will be provided. However, these browsers will not be updated or enhanced. One year after the company has withdrawn support from an OS, AIR will not take into account future developments considerations based on support for this OS.

Currently, all supported operating systems receive full support.
I.C.2.d Help Desk Support

Easily accessible, prompt, and accurate customer support is essential to success. AIR commits to providing a toll-free customer support line and e-mail for state users, LEA users, educators, and administrators Monday through Friday from 8:00 a.m. to 5:00 p.m. Mountain Time when summative tests are not being administered and from 7:00 a.m. to 7:00 p.m. during the summative testing windows. AIR has competitive average response times, and we will work closely with USOE to define the details of help desk support requirements, including acceptable hold times, response times, and other metrics.

The help desk will be available to answer all questions about any aspect of Utah’s testing system. Examples of types of support include installing the secure browser, using the system, resolving system access issues, processing special cases such as test invalidations or test resets, and resolving issues involving high network latencies. The help desk staff will be able to identify questions that are matters of policy and belong to USOE and those that are to be answered by AIR. Policy questions will be directed to USOE.

AIR implements a tier-level approach for the help desk to resolve questions from callers:

1. The first tier consists of providing scripted answers for resolving routine queries.
2. The second tier consists of a help desk representative calling a member of a user support team for further investigation.
3. The third tier consists of contacting a subject matter expert as described below.

Subject Matter Experts

AIR recognizes that technical and highly detailed questions arise throughout the school year. Some of these questions will be handled through help desk scripts. Other, more difficult questions, however, may require additional follow-up from a subject matter expert.

All calls and e-mails are logged in to a central system, where senior project management monitors them daily. In our experience, most help desk calls are from users who are unfamiliar with the system and need assistance. Successful district and school staff training as well as thorough system documentation and support available online in the testing system should keep these requests to a minimum. The few help desk calls that require technical assistance tend to relate to problems with the user’s local infrastructure, such as firewall problems and lack of connectivity to the Internet. The technical support manager reviews the issues to determine the impact to the program and escalates unresolved issues to the Project Management team.

AIR will provide weekly reports to USOE during peak times. Reports will include volume, wait time, status of each inquiry, and, if the inquiry is resolved, details of the resolution.

Training and Supervision of Customer Service Representatives at the Help Desk

AIR’s training of customer service representatives at the AIR help desk is based on our experience in assisting new users in and with varying degrees of technical knowledge on how to use an entirely new online testing system. Experience has shown that callers may not know enough about the online testing system to articulate the specific system issue they are encountering. Therefore, it is essential that our customer service agents be able to guide callers through a series of questions designed to diagnose the problem quickly.
In addition, our past experience reveals that many callers simply want someone to walk them through the step-by-step instructions provided in the system documentation. AIR’s customer service agents are trained to provide this level of support to educators and administrators wherever required.

Finally, analyses of past help desk call logs show that customer queries tend to be divided into three distinct categories:

1. Questions related to preparation for online testing (e.g., scheduling online test sessions, training test administrators on online procedures, installing secure browsers)
2. Questions during the beginning of the test window (e.g., using the system, setting accommodations, understanding participation counts, understanding how to create class rosters)
3. Questions during the ongoing administration (e.g., system access, test invalidations, test resets, network latencies)

Based on the type of customer queries we typically receive, AIR’s customer service representatives must complete a rigorous three-part training program:

1. Training on the functionality of each component of the online testing systems as well as the paper testing program, including the policies and procedures set forth in user guides and manuals
2. Training based on past logs of customer queries taken from each stage of an online administration: preparation, startup, and ongoing administration
3. Training on how to use the detailed Visio scripts that enable the customer service agent to ask a series of questions to callers to quickly diagnose the system and understand the issue with which the caller is experiencing difficulty

In addition to providing rigorously trained customer service agents, AIR’s help desk is managed by a full-time supervisor responsible for monitoring response accuracy. The supervisor regularly audits calls to the help desk to ensure that help desk agents are providing accurate information. When a question of possible inaccuracy arises, the supervisor obtains a copy of the phone recording and can confirm whether the information provided was correct or not.

Should a system-wide issue arise, AIR’s policy is for the project director to report any issues directly to USOE. AIR will work with USOE to notify districts of any system-wide issue within 10 minutes of discovering the issue.

Item Tracking System Support

AIR provides live training sessions for two different types of USOE users for the Item Management System. USOE users will have two roles in the Item Tracking System (ITS): They will conduct client reviews (or Utah may conduct the reviews on paper if preferred), and they may access the reporting functions to check on the inventory of items and progress of development. ITS trainers will train USOE staff to configure their computers and access ITS. We will offer a walkthrough of each function and take questions. ITS support staff and the project team will always be available to help Utah users.

I.C.2.e Paper-Based Administration

Utah requires a paper form to be administered to a small number of students each year. Paper-based administration is intended for

- students who require Braille;
- students who require large print;
- students who cannot interact with a computer due to a disability or other documented reason; and
- students who need a paper form because of a local emergency (e.g., school fire).

We note that in some states incarcerated students cannot have access to computers, and this group represents another constituency requiring paper testing. Including these students, we have three use cases for paper tests.

In the first three cases, we offer a solution that is currently in use in and which meets the highest standards of comparability and equity for students with disabilities.

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 approves the print and retrieves the document from the printer or embosser.

For students who cannot interact with the computer, we offer two options:

- If USOE would like to offer these students the benefits of an adaptive test, a scribe can enter the student’s responses.
- If USOE is satisfied with a fixed-form test, the student can respond on paper, and a designated scribe can enter the data at a later date through our Data Entry Interface, which essentially allows the test administrator to go through the test for the student.

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.

Large print and print on demand come at no additional cost because they add no additional costs on our end. An adaptive Braille test requires annotating and quality-checking automatically generated Braille for a large pool of items, which does raise our costs. To provide a price comparable to others and consistent with the RFP, our price includes only the Braille preparation of one fixed Braille form per grade and subject. However, we are happy to negotiate an additional price that will provide students who are blind and read Braille the same number of opportunities and quality of adaptation as other students in the state.

Additionally, we note that items shared with and will have Braille already associated and will not produce higher costs if no changes are made to the items. Similarly, Smarter Balanced plans to Braille all of the Consortium’s items. Embossers are not included in the price for Braille. An appropriate embosser costs approximately $5,000 and provides value throughout the year as educators use it to expand the range of material to which students have access.

We strongly recommend that Utah offer the real-time adaptive paper versions wherever possible.

We recognize that in two use cases the computer is not usable (a local emergency and incarcerated students). For these cases, we propose to deliver a single form per grade and subject for use through the course of the contract.

We provide additional details of our proposal below.

**Item Selection for Paper Forms**

Where the real-time adaptive paper option is used, the item selection occurs exactly as it does for the online adaptive assessment.
Our system does provide the ability to generate “customized item pools.” Each item can be tagged for accessibility. For example, some inherently visual items may be inaccessible to students who are blind, some types of interactive items may not be meaningfully administered on paper, etc. Our system supports and respects accessibility limitation tags. Custom item pools are constructed from the full pool by trimming from the pool any items that conflict with the student’s access profile and any items the student has seen before. Hence, the same algorithm selects items from a customized pool that is limited to those items the student can access.

In constructing fixed forms, we propose to ensure comparability by simulating a student taking an adaptive test. We will simulate three students whose proficiency is exactly at the proficiency cut score and assemble bookmaps to match the sequence of items the simulated students received. USOE will have the opportunity to select from among these three forms for the fixed form to be used.

By using the same item selection process used for all students, we ensure optimal comparability, recognizing that access limitations may restrict the item pool as described above. This procedural approach to comparability is the best possible approach. The small numbers of students to be tested on paper preclude a formal comparability study.

Capturing Data from Paper-Based Administration

Above, we describe how the system supports data capture when students are using the real-time adaptive paper (print or emboss on demand) system. Either the student or a scribe enters the responses in real time.

When fixed forms are administered on paper, school-based scribes can transcribe the data into our system. Students respond on the consumable paper form, and school personnel will be responsible for entering that data into the test delivery system.

The system supports this function through our Data Entry Interface, which allows a test administrator to log in as a scribe for a student and run through the test item-for-item, copying the student responses into the system. We recognize that this will mean an increased burden for a small number of educators, but it has the benefit of providing immediate feedback and including the students affected in the same educational workflow as their peers.

Producing the Consumable, Fixed-Form Paper Booklets

AIR’s production system offers several key features that will allow AIR to facilitate materials production and meet the timeline and requirements of Utah’s assessment system:

- **Item Tracking System (ITS) integration.** As described in Section I.C.1, ITS is a full-featured test development system that manages test content from conception through deployment. ITS integrates with AIR’s Test Delivery System (TDS) to publish online assessments, as well as with AIR’s production system for development of hard-copy test forms. ITS ensures consistency in content among all test formats and real-time documentation on item attributes, item use histories, related statistics, and accommodated formats.

- **System-enforced version control.** During the item development phase and for online assessments, version control is enforced by ITS, which shepherds each item through the development and review process, ensuring that each required review takes place and that only items passing those reviews advance to subsequent levels.
Version control for hard-copy test form production is enforced by AIR's production system, including established steps enumerated below.

- **Experienced staff.** AIR's production team includes 12 graphics specialists, nine of whom have B.F.A. or M.F.A. degrees in graphic design, computer design, or art. They offer years of experience with different types of items and stimuli, test booklets, answer documents, special versions/accommodations, manuals, and online test delivery systems. The production staff's experience with other large state testing programs allows them to accurately estimate time on task for materials production activities to develop schedules, staff appropriately, and meet client needs. They are experienced in working with various style guides, item and test specifications, and state-specific requirements.

**Test Booklet Development/Production**

Items in the fixed-form booklets will be formatted as closely as possible to the online versions, given resource and technology constraints.

AIR's quality assurance process is designed for accuracy and efficiency and is predicated on all participants understanding what is expected at each stage, having sufficient time to accomplish the objectives, and applying changes in a systematic way.

AIR’s proposed test booklet production process starts with the creation of test maps (also referred to as bookmaps). The test map is built in ITS and initiates the production of printed test forms. The process proceeds in three steps:

1. The 1x1s are generated based on the test map (bookmap).
2. Blackline 1 is drafted.
3. The documents are brought to blueline (camera-ready copy).

Step 1 is entirely automated within ITS. ITS houses destination templates that define the format of the 1x1s and automatically generates these documents based on the test map. (AIR defines 1x1s as test items printed one per page according to the test map sequence.) At this stage, items are proofread by editorial staff as well as test development staff at AIR and USOE. Additionally, they are reviewed to verify that all edits from previous rounds of review have been correctly implemented. Any changes required at this stage are entered directly into ITS to ensure consistency across all item uses.

Blackline 1 is a semi-automated process. With the appropriate destination template defined and 1x1 approval, ITS generates a Quark-readable document in the specified format. Through this integration, items are automatically styled with fonts, graphics, spacing, and other formatting specifications outlined by USOE's style guide. Our production staff may adjust page layout, including instructions, borders, and other elements, to meet USOE guidelines. At this stage, reviewers check the document layout and formatting. Should an egregious error be found in the content of an item, changes must be entered into ITS and the item must be re-exported to ensure consistent item use across all test forms. Changes to blackline 1 require a second blackline proof. Changes to subsequent blackline proofs require sign-off by senior management.

The final quality assurance step prior to printing is the blueline, or camera-ready copy, review stage. During this step, AIR and USOE staff review proofs from the print vendor, verifying that the file to be printed matches the previously approved blackline proof. At AIR, in addition to reviews by test development and
forms production staff, two members of the technical team, who have not seen the items previously, independently take the test. This process forces a close look at the items and gives a final opportunity to verify the keys.

During the production and review process, test booklet blacklines are accompanied by answer document blacklines, which are produced on high-quality laser printers for quantities this small. The item sequence is based on test maps and corresponds directly with test booklets.

All blacklines in AIR’s production queue are controlled by an electronic version control server system that ensures that only the current version is immediately available to our production staff, preventing version control errors. Like AIR’s ITS, which controls and tracks all changes to items, this production system maintains historical records (including all older versions), which senior production staff can access if necessary.

Each blackline after blackline 1 and the blue-line (camera-ready copy) are automatically compared with the immediately preceding version, using a PDF comparison tool that highlights all changes. This step has proved useful for identifying unintended changes made during the revision process. Such changes are difficult to detect because they can appear anywhere in a document and may be subtle. The PDF comparison tool highlights these changes so that differences between versions can be mapped to an intended revision. All materials delivered will go through this process, ensuring that USOE will receive error-free materials for review and that any changes requested by USOE are implemented promptly and accurately.

At each of the review stages, proofs will be accompanied by proof tickets that identify the document being reviewed, its review stage, the scheduled and actual delivery dates, and the return date. Sign off by USOE will be required at each stage before we proceed with subsequent steps.

Final source files will be made available to USOE in electronic form no later than 30 days after the close of the test window.

**Quantities**

AIR recognizes that no more than 30 students will require Braille, 75 students will require large print, and 20 students will require other printed forms for each subject assessed.
I.C.2.f. Test Administration Manual (TAM) and User Guides

Test Administration Manual (TAM)

The TAM’s target audience will be test administrators and test coordinators. The TAM will provide all the information test coordinators and test administrators require. It will first describe USOE and relevant policies and then introduce each system component that is visible to these users and to teachers. The manual will cover all steps from setup through reporting. A preliminary outline of contents for the TAM appears in I.C.2.f-1. We will work with USOE to identify and develop the final outline and content for the TAM.

The overview section will describe the Utah assessment system. The new testing system will depart from the prior system in important ways. The system is adaptive, and both the test administrator and student experiences will differ from the old system.

The Test Administration section will likely be one of the most heavily used. This interface is the primary tool test administrators use to create and manage test sessions; it is an invaluable reference for new test administrators and a trusted resource for those familiar with the process. This section of the manual will provide easy-to-follow instructions about

- defining roles and responsibilities;
- setting up testing sessions;
- monitoring sessions;
- verifying student information;
- assigning accommodations; and
- starting and pausing tests.

This section will also prepare test administrators to handle common student questions by familiarizing administrators with the Student Interface and describing what students will see, the tools they will use, and how those tools will function.

The bulk of the TAM will cover test administration using the online system, but the manual will also provide brief directions on administering tests on paper (either paper forms or print-on-demand) and administering special forms. The manual will direct users to detailed information about these topics on the portal website.
Exhibit I.C.2.f-1: Preliminary Outline for Test Administration Manual (TAM)

<table>
<thead>
<tr>
<th>(1) Overview of USOE</th>
<th>(5) Paper testing, special versions, accommodations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Purposes and goals of the system</td>
<td>(a) When paper testing is appropriate</td>
</tr>
<tr>
<td>(b) General operational procedures</td>
<td>(i) Ordering paper tests</td>
</tr>
<tr>
<td>(c) Key dates</td>
<td>(ii) Summary of procedures</td>
</tr>
<tr>
<td>(d) Additional resources</td>
<td>(iii) Where to get more information</td>
</tr>
<tr>
<td>(e) Security requirements</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(2) Registration of students, teachers, and test administrators</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) District interface to USOE Student Information System</td>
<td>(b) Special versions</td>
</tr>
<tr>
<td>(b) Data transferred from USOE Student Information System</td>
<td>(i) Availability</td>
</tr>
<tr>
<td>to Test Information Distribution Engine (TIDE)</td>
<td>(ii) Who qualifies</td>
</tr>
<tr>
<td>(c) Features of the USOE User Management Interface (for features not handled by USOE Student Information System)</td>
<td>(iii) Where to get more information</td>
</tr>
<tr>
<td>(d) What to do if a student is not in the system</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(3) Administering tests</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Overview of multiple opportunities and policies</td>
<td>(5) Paper testing, special versions, accommodations</td>
</tr>
<tr>
<td>(b) Establishing sessions</td>
<td>(a) When paper testing is appropriate</td>
</tr>
<tr>
<td>(i) How to...</td>
<td>(i) Ordering paper tests</td>
</tr>
<tr>
<td>(ii) Best practice</td>
<td>(ii) Summary of procedures</td>
</tr>
<tr>
<td>(iii) Practices to avoid</td>
<td>(iii) Where to get more information</td>
</tr>
<tr>
<td>(c) Assigning accommodations</td>
<td></td>
</tr>
<tr>
<td>(d) Authentication of students</td>
<td></td>
</tr>
<tr>
<td>(i) Security reminders</td>
<td></td>
</tr>
<tr>
<td>(ii) Authentication and logon procedures</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(4) Test results</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Overview of reporting system</td>
<td>(b) Special versions</td>
</tr>
<tr>
<td>(b) Authorizing teachers to use the system</td>
<td>(i) Availability</td>
</tr>
<tr>
<td>(c) Using the system</td>
<td>(ii) Who qualifies</td>
</tr>
<tr>
<td>(i) Goals and objectives</td>
<td>(iii) Where to get more information</td>
</tr>
<tr>
<td>(ii) Navigating the system</td>
<td></td>
</tr>
<tr>
<td>(iii) Setting up classes and other rosters</td>
<td></td>
</tr>
<tr>
<td>(iv) Reviewing reports</td>
<td></td>
</tr>
<tr>
<td>(v) Using the data to improve instruction</td>
<td></td>
</tr>
<tr>
<td>(d) What to do if...</td>
<td></td>
</tr>
<tr>
<td>(i) I cannot remember my password</td>
<td>(5) Paper testing, special versions, accommodations</td>
</tr>
<tr>
<td>(ii) Students are in the wrong classes</td>
<td>(a) When paper testing is appropriate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(6) Troubleshooting</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
User Guides

The Assessment Program at AIR has a dedicated team of technical writers, editors, and graphic designers who are continually involved in developing, editing, and enhancing documentation for AIR online testing systems, including the Test Delivery System, Test Information Distribution Engine, Online Reporting System, and more.

The goal of each user guide will be to motivate teachers to make the best possible use of the USOE system and the data it provides. The guides will provide how-to information to enable users to access and navigate the systems they are authorized to use.

Each user guide will begin with an overview of the system, much like the one presented in the Test Administration Manual. The guide will include screen shots and descriptions of the students’ testing experience and how the standards are measured. It will walk the reader through the online reporting system, identifying ways in which the data could be used to inform classroom instruction or curricular decisions.

AIR will draft each of the guides and then work with the USOE to revise it until USOE is satisfied with the quality and contents of the guide.

Test Information Distribution Engine (TIDE) User Guide

Most students will be registered through the Test Information Distribution Engine (TIDE), so no action on the part of test administrators will be required. Test coordinators will learn to upload student information. In some cases, new students will need to be introduced into the TIDE system, and the test administrators will need to know the appropriate procedures. The TIDE User Guide will begin with a high-level overview of the registration process to orient the reader and will continue with the various special cases that the test administrators are likely to encounter.

Online Reporting System (ORS) User Guide

The ORS User Guide will walk users through the online reporting system. It will begin by describing how the data in the system can support the users 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.

Online Technology Requirements User Guide

This document is primarily for technology coordinators and IT personnel. It provides information on hardware and software requirements for online testing, as well as network installation of the secure browser and guidelines for online testing environments.

Production of User Manuals and 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 USOE, during which we discuss issues of both design and content. We view the drafting process as iterative. Especially with new materials, we understand that it is not possible to make all final decisions prior to a first draft. 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 multiple drafts of each product prior to entering the production process.

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 I.C.2.f-2 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 I.C.2.f-2: Overview of Our Editorial Review and Revision Process

Process for Publishing Documents (User Guides, Brochures, etc.)
I.C.3: Test Delivery System

AIR’s online Test Delivery System (TDS) offers USOE four key benefits that form the foundation for a successful, sustainable, state-of-the-art assessment:

- The system has a very small footprint within schools, with minimal requirements for hardware, software, bandwidth, and technological expertise.
- The architecture supports continued innovation by allowing the introduction of new item types within the same framework without disrupting the existing code base.
- The system can be accessed by nearly all students because it supports a broad and growing range of configurable accessibility features and accommodations. AIR continues to invest in extending these features to make the system accessible to all students.
- The system enables the management of complex, over-time administrations.

Our system’s robustness and capabilities have been proven statewide in , and, each of which administers an online adaptive test that includes technology-enhanced items. Our experience delivering standards-based adaptive assessments has led to the development of tools and processes that help us and our clients manage their assessments. For example, our test simulation tool enables our psychometricians to optimize the adaptive engine for specific item pools and blueprints. Our clients know how their tests will perform before students start taking the tests. Other tools support the real-time monitoring of system and network performance in every test session and monitor participation to help assessment coordinators know who has, and who has not, tested.

I.C.3.a. Overall Approach

The AIR online testing system is made up of a set of seven integrated modules, along with our common login system (CLS), 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 I.C.3.a-1 provides a brief summary of each module and its role in the overall system.

Here, we describe the system in two main parts. First, we describe the key components of the system and their capabilities.

System Description and Capabilities

Exhibit I.C.3.a-2 provides a schematic of the overall system. Beginning near the top, we see our Item Tracking System (ITS). ITS is a full-featured test development system that manages test content from conception through deployment. It was discussed at length in Section I.C.1.

The TDS draws data from ITS to publish tests. To administer tests, TDS also needs information about students and TAs, including authentication information. The TIDE system gathers data from districts, schools, or the state and stores those data in the Roster Tracking System (RTS), a flexible database system shared by TIDE, TDS, and the reporting systems. RTS will maintain data about the educational networks in Utah, 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...
Exhibit I.C.3.a-1: Summary of Online Test Delivery Components

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
</tr>
</thead>
</table>
| Test Information Distribution Engine (TIDE) | TIDE is responsible for:  
- Student registration  
- Gathering of demographic data  
- Materials ordering |
| Test Delivery System (TDS) | TDS is the Test Delivery System. The TA interface (TDS-TA) provides the interface through which test administrators establish and manage testing sessions and authenticate student users. The student interface (TDS-S) is the testing system as it appears to the student, on which students take tests. TDS delivers tests to students, records responses, and forwards data to downstream systems. |
| Quality Monitor (QM) System | The Quality Monitor 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 and the Online Reporting System. |
| 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) | 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. |
| Learning Point Navigator (Navigator) | Navigator integrates with the reporting system to provide resources targeted to identified areas of weakness or to extend standards-aligned topics for students who are excelling. The system provides immediate access to standards-aligned professional development, lesson plans, and student activities. Both educators and students (and their parents) can access the system to obtain or use learning resources. |
| Online Portal | The online portal provides a customized one-stop shop for educators and other stakeholders and provides access to the other systems as well as practice sites, training sites, tutorials, and other resources. |

administer tests and can allow them to log in and define sessions.

After a test is administered to a student, TDS passes the resulting data to our Quality Monitor system (QM). QM rescores tests, checks that the tests meet blueprint, captures statistics on items, and runs a host of extensive quality checks. QM 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 QM identifies an anomalous test result, members of the project team are immediately emailed, and the results are held until the results are verified.

From QM, data are passed to the Database of Record (DoR), which serves as our consistent repository of data. The extract system, which will interface with USOE’s systems to deliver test and item data, will draw data from DoR. DoR will send data to Response Bank, our data system that supports the online testing system, as well as to our Score Reporting team.

The components of the system share a common login system (CLS), enabling a user to log in to any one of the systems (excluding ITS) and access any other system for which he or she is authorized. CLS allows the definition of customized roles, and AIR will work with USOE to define a sufficient number of differentiated roles to support workflows at the state and in schools and districts throughout the state.

Test Delivery System

AIR's TDS has five major components:

- The Test Administrator Interface (TA Interface), from which test administrators manage testing sessions
- The Student Interface, through which students take the test
- The Item Selection Engine, which includes the adaptive algorithm, the ability to
Exhibit I.C.3.a-2: Summary of Online Test Delivery Components

Existing Utah Systems

**Utah Item Repository**

<table>
<thead>
<tr>
<th>Item Tracking System (ITS)</th>
<th>TIDE</th>
<th>Test Delivery System (TDS)</th>
</tr>
</thead>
</table>

**UserWare**

- ITS manages item development and maintains the item content and data. Users access ITS through a secure web interface.
- TDS delivers tests containing selected and constructed-response items to students.

**Repository**

- Item Tracking System (RTS)
  - RTS tracks demographics and relationships among districts, schools, teachers, classes, and students, as well as user rights.

**Data Bank/Reporting Manager**

- Response Bank/Reporting Manager holds student scores and scale information.

**Test Delivery System (TDS)**

- TDS Data
  - TDS Data monitors and reports test performance.

**Teacher Workstation**

- Teacher Workstation

**Student Workstation**

- Student Workstation

**Utah Systems**

- Utah Systems

**Single Sign-on Authentication System**

- Single Sign-on Authentication System

**DoR Maintain and Archive**

- DoR maintains and archives data. Updates to scores (for example) are propagated to other systems.

**DoR Monitor (QM)**

- DoR Monitor (QM) performs real-time data validation and maintains and archives data.

**Update to Scores (for example)**

- Update to scores (for example) are propagated to other systems.

**The Asynchronous Scoring Framework**

- The Asynchronous Scoring Framework, which supports the integration of an extensible collection of machine-scored items and human-scored items in a single, adaptive test.

**Administration and Monitoring Tools**

- Administration and monitoring tools, which allow monitoring of sessions and system health and performance.

The Student Interface is designed to operate well on older and lower-end computers and to do so with minimal bandwidth usage. At the same time, the system supports a broad and growing range of item types and accessibility features. These are described below.

**Test Administrator Interface**

The TA Interface is discussed at length in the previous section under Test Administration Requirements in the discussion of the testing workflow. Every aspect of the test administrator interface simplifies the administrative burden. Specifically:

- There is no need to establish testing sessions in advance. At testing time, the test administrator creates a session with as few as two mouse clicks and the system returns a Session ID. Students then log in with identifying information and the Session ID, requesting admission to the session. From the TA Interface, the test administrator authenticates the students and admits them into the session. The test administrator can view or change test settings and accommodations for students at the time of testing.

- The core functionality of the test administrator is accessible through a single, intuitive screen. This includes setting up sessions, admitting students, adjusting test settings, monitoring student progress and requests, and pausing or ending sessions.

We recognize that test administrators are often teachers, and each teacher administers tests...
infrequently. Because the system is used infre-
quently, it is unreasonable to expect TAs to become “expert users.” The system is simple and intuitive enough that TAs can operate it even if they only interact with it every couple of months or so.

Please see Section I.C.2.c for a more detailed description of the testing workflow and the TA interface.

**Student Interface**

Students take tests through the Student Interface, which is essentially a secure website that is accessed through the AIR secure browser (see below). The AIR secure browser is the only software needed to take a secure test, and it is simply a secure build of the Mozilla (Firefox) browser.

AIR’s TDS works on mobile devices, although some accessibility features are currently available only on computers (including netbooks and such). AIR tests are compatible with WebKit browsers (Safari and Chrome) and can be accessed using these devices. The previous iPad operating system (iOS Version 5) includes some security limitations that must be controlled by policy rather than electronically.\(^1\) These limitations have been addressed with the release of iOS 6. We recommend that Utah limit mobile use to devices that can be appropriately secured.

For mobile devices, AIR is developing a separate app targeting iOS 6 that will enforce stricter security using new features available with Version 6. We also intend to release a similar app targeting Android devices by spring 2013.

AIR’s TDS can deliver a variety of machine- and human-scored item types. These include multiple-choice items, natural language constructed-response items to be machine- or human-scored, items that capture student speech, machine-scored items to which students respond graphically, drag-and-drop items, simulations, and others.

The 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. 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-and-pencil test.

**Secure Browser**

Students can only access secure tests using AIR’s secure browser. The browser can be “branded” for USOE, with an icon and logo of your choosing. Non-secure tests, such as practice tests and some formative tests, can be accessed with ordinary browsers. The secure browser operates in a full-screen mode, disables access to other applications, and prohibits navigation outside the test. The browser

---

\(^1\) The versions, and earlier of iOS have two security limitations. First, there is no way to stop students from switching applications. However, we can reliably detect when the browser loses focus and we can stop the student from testing at that point. The devices also have a screenshot capability. While this capability can be administratively disabled, it cannot be disabled through software, and software cannot currently detect whether it is disabled. If Utah decides to allow the use of iOS Version 5 (or below) devices, these security risks will have to be controlled through administration policies. These limitations have been addressed with the release of Version 6.
The system can deliver a large and growing set of both machine-scored and human-scored item types:

- Multiple-choice items.
- Machine-scored “propositional response” items, in which students respond with one or more phrases or sentences. For example, English language arts (ELA) items asking for a summary of a passage or an enumeration of supporting details often fit into this category.
- Machine-scored graphic response items in which students respond by dragging objects or drawing points and lines. This technology can be used not only for simple drag-and-drop items but also for a wide range of much more open-ended item types. Exhibit I.C.3.a-3 shows two quite different items of this type.
- Machine-scored equation-response items, in which students use a graphical equation editor to enter equations, which are machine-scored by a sophisticated symbolic algebra scoring engine. Test developers can create rubrics that offer partial credit based on the mathematical characteristics of the student response.
- Simulation items, which can score students on their interactions with the simulation. For example, which data they choose to generate, how they respond to simulated situations, and most any other interaction that test developers can imagine.
- Machine-scored essays.
- Human-scored, open-ended items in which responses are offered in writing, drawn, or spoken and are captured by the system.
- Collections of items of different types around a theme, simulation, or common stimulus.

Our graphic response technology is unique within the industry. Using a single, consistent, and simple response mechanism, the system supports many different item types. Exhibit 1.3a-3 illustrates two items implemented in this system.
Our simulation technology is also unique within the industry. A simulation display consists of three panels: an input panel, an animation panel, and an output panel. The input panel uses standard user input widgets to provide a consistent interface across the entire test. Student inputs are translated into two types of outputs: instructions to the animation and data to be output to an output display table. The animations can be graphics, Flash, or HTML-5. All of the calculations are done outside the animations and passed into the animation control. Our test developers have tools to define the calculations. Using this approach, we ensure that those with the content knowledge, the test developers, can directly control the interactive content. There is no need to involve software developers. Our animators (who are analogous to graphic artists for paper tests) create animations that dynamically “display” the data passed into them by playing the correct animations in the correct ways.

One critical characteristic of our technologies is that they are developed entirely by our Test Development team, rather than by programmers. This enables us to develop the items at approximately the same cost as human-scored constructed-response items, allowing for richer tests with more constructed-response items.

### Interface Design, Tools, and Accessibility Features

In this subsection we describe

- student navigation of the test,
- available tools and accommodations,
- help functions and tutorials, and
- capture of USOE information.

#### Student Navigation

The Student Interface is designed to be simple and intuitive. When a student is taking the test, his or her attention and concentration should focus on the content of the items, rather than navigation of the system. Therefore, the interface minimizes the number of controls the user must access and uses a consistent paradigm across tools.

The basic interface is illustrated in Exhibit I.C.3.a-4, where readers will notice very few controls, just navigation buttons on the bottom right, a few test-specific tools on the bottom left, and position tracking on the top toolbar. Simple business rules help keep the interface uncluttered—for example, the “End Test” button does not appear until the last item on the test.

Clients can choose to have most features set remotely by the test administrator or other adult working on a separate workstation. Utah will have the option to have accessibility features (background color, large-print, text-to-speech, etc.) for a given student set in advance through an upload, set in real time by the test administrator through the TA Interface, or available on the student interface for students to choose (although this last option works against overall accessibility by cluttering the screen and requiring student attention).

Highlighting, strike-through, text-to-speech, and other capabilities that operate on a section of the item or stimulus all work identically: Select the desired object or region, click the right mouse button (or ctrl-click for Mac) to instantiate a context menu, and choose the desired action from the menu.

We have often been asked why we use context menus to help students access tools. This approach provides the most usable and accessible interface. First, students who cannot use a keyboard often use assistive devices such as switch arrays that generate a limited number of keyboard commands. Maintaining compatibility with these devices precludes creating
Exhibit I.C.3.a-4: What Students See While Testing

Keeps track of which question the student is on and which he or she has marked. The student can go back to questions he or she has recently answered.

The look and feel and screen layouts are configurable.

Items about a common stimulus can be shown together or one at a time.

Navigation buttons use simple, intuitive icons.

Intuitive icons present tools available to the students. We configure which tools appear.

A wide variety of item types are available.

1.4 million online test administrations. The only test makes available to students offline is the writing assessment, which 12,828 students took in paper form last year. In the 2011–12 school year, 96,763 students sat for assessments. Of that total number, only twelve completed a paper version of the test.

Alone in the industry, AIR delivers real-time, adaptive Braille tests and . The adaptive tests draw on the full pool (excluding inherently visual items) and delivers these items in Braille that meets Braille Authority of North America (BANA) standards—something that standard refreshable Braille displays are unable to do because (1) none currently available renders Nemeth Braille, the standard Braille for math and science, and (2) none is currently capable of delivering tactile graphics. We deliver a cost-effective solution that meets these requirements.

Our system supports Braille readers and also our more standards-compliant approach. Specifically, we render items to desktop embossers (which cost between $5,000 and $10,000 each) that can integrate Braille and tactile graphics. The items simultaneously render on a reader-accessible screen, and the student will be able to navigate to response spaces to provide answers. This approach is more compatible with an adaptive test than an approach that distributes hard-copy tactile...
graphics to be selected and handed to a student during the test.

AIR’s approach to accessibility favors Universal Design over separate accommodations. Each of our accessibility features can be configured for universal access (all students can access the feature) or for access controlled by any one or more of the following user types: TA, AC, district coordinator, or state user. For example, in order, each student can choose to zoom (large print) on any item he or she chooses. The TA can set the background color for a student from the TA Interface. AIR offers text-to-speech on math and science tests as a universal design feature but requires that a TA turn it on (from the TA Interface) after confirming that the student has the necessary equipment on his or her computer (headphones). For reading tests, the TA reserves the right to turn on text-to-speech for the state, so TAs and ACs can view the assignment of this modification, but only state users can change it for a student.

In general we recommend that most features, even those that are universally available, be controlled from the TA workstation rather than the student interface. The reason for this recommendation is that the sheer number of features can create a distraction in the student interface. Many students become distracted changing background colors, using text-to-speech voices, playing with masks and guides, etc. Therefore, too many student choices can actually reduce access to accessibility features.

Exhibit I.C.3.a-5 summarizes by category many of the features, tools, and accommodations that are available on the system. With few exceptions, which are discussed below, our system supports all the tools and accommodations requested in the SOW. Please see Section II.B.2 for details. Exhibit I.C.3.a-6 shows how our currently deployed system supports the seven principles of universal design. We note that our writing tests support a host of tools and features that are selectable by USOE, in accordance with the SOW.

Help Functions and Tutorials

AIR’s online tests provide four critical features that support students as they use the test:

- Before testing, we provide practice tests with features that mirror the operational assessment.
- Before testing we provide online public, interactive tutorials that teach students how to use our graphic response items. Other tutorials can be added.
- Each test begins with a Help screen showing how to navigate the system. This screen is accessible throughout the test.
- Each item can be associated with a unique viewable tutorial, which we generally assign by item type.

The practice tests generally represent a version of the actual test and are based on released items. If a sufficient number of released items are available, the practice tests can be adaptive; otherwise, we configure them as fixed-form tests.

Exhibit I.C.3.a-7 diagrams our interactive graphic-response tutorial. Studies in our cognitive laboratory confirmed that young students (grade 4) who had little access to computers at school and none at home were able to successfully navigate graphic-response items after working independently through the tutorial. The tutorial efficiently guides the student through every maneuver ever required by any item type using this response mechanism.

Item Selection Engine

AIR’s system offers a mature, robust adaptive engine. We believe that it is unique within the industry in that we employ a standards-based...
Exhibit I.C.3.a-5: Available Features That Support Access to the System

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
</table>
| Visual accessibility features   | ■ Zoom/large print is available.  
■ Change background color/contrast is available.  
■ Text-to-speech synthesizer can be used.  
■ Audio files for instructions, items, and other user support materials are provided.  
■ Secure passage print facility is available.                                                                                          |
| Attention accessibility features| ■ Test can be paused at any time and restarted and taken over many days.  
■ Passage highlighting is available.  
■ Strike-through is available for multiple-choice options.  
■ Mark-and-return can be used for item review.  
■ Navigation options (Back, Next, Pause, End Test) are distinct and intuitive.  
■ Item notes allow students to jot down ideas about items or passages. Test can be reviewed before ending it.  
■ Messages are available informing students of status of test.                                                                            |
| Agility accessibility features  | ■ Keyboard navigation makes the test device independent.  
■ Area boundaries for mouse-clicking multiple-choice options mean students can click anywhere on the selected response text or button.  
■ Graphic-response items can be navigated 100% using the keyboard.                                                                     |
| Language accessibility features | ■ Any language that is necessary can be supported.  
(Refer to our cost proposal for a pricing model when a language needs to be supported.)                                                |
| Reference accessibility features| ■ Subject-assistive tools include periodic tables, calculators, and formula sheets.  
■ Help section explains how the system and its tools work.  
■ Practice tests and tutorials familiarize students with the online testing system.  
■ Performance report is available at the end of the test for the student.                                                               |

adaptive approach. By this we mean that our adaptive algorithm, which remains the only one to have passed federal peer review, will enforce sophisticated blueprints to ensure that each student sees the full breadth of the content, regardless of his or her performance level. The blueprint can control the distribution of items across content classifications (e.g., strands), enforce a broad distribution across benchmarks, and enforce a distribution across Depth of Knowledge classifications, item types, and other features of items.

AIR offers the tools necessary to support the nuanced complexity of realistic standards-based blueprints. The performance of any algorithm in optimizing item difficulty while meeting a complex blueprint depends on the interaction between the blueprint and the available item pool. Our algorithm parameterizes many of the internal calculations, allowing us to fine-tune its behavior for any given blueprint–item-pool pair. Our psychometricians work with our simulation tool and the actual item bank to simulate thousands of students taking the test and optimize the algorithm to meet the blueprint and optimize measurement. Below we offer a brief description of the adaptive algorithm.

**Overview of Adaptive Algorithm**

Early adaptive algorithms viewed content as completely exchangeable, so the only attribute used to select items was difficulty. Most algorithms now use one mechanism or another to ensure content coverage, at least at a broad
### Exhibit I.C.3.a-6: Features Reflecting Specific Universal Design Principles

<table>
<thead>
<tr>
<th>Universal Design Principle</th>
<th>Supporting Features</th>
</tr>
</thead>
</table>
| **1. Equitable use** | - At USOE's discretion, we can provide the same accessibility features for visual, agility, language, reference, and attention to all students.  
- Practice and training sites are available ahead of time.  
- Navigation buttons have color and graphical indicators so that they are understood by all students.  
- Keyboard navigations or shortcuts are available for students who have trouble working with the mouse.  
- Any language necessary for the student population is available. |
| **2. Flexibility in use** | - Zoom/large print and subject-assistive tools can be used at any time at the student's preference.  
- Area boundaries are provided for mouse-clicking multiple-choice options.  
- Tests can be paused at any time and restarted and taken over many days. Test administrators can create sessions ahead of time or just in time; sessions can include any combination of tests. |
| **3. Simple and intuitive use** | - Interface is clean and simple.  
- Messages inform students of status of test.  
- Navigation buttons have color and graphical indicators so that they are understood by all students.  
- The Help tool is available at any time during the test.  
- Test instructions are simple.  
- Font face and type are appropriate to age group. |
| **4. Perceptible information** | - Background color/contrast can be changed.  
- Text-to-speech synthesizer is available.  
- Audio files are available.  
- Keyboard navigation can be used to start and end a test.  
- Keyboard-navigable multiple-choice response options are available. |
| **5. Tolerance for error** | - Logon system incorporates multiple validations and confirmations.  
- Students can go back to review/change answers during a testing session.  
- Messages inform students of status of test.  
- End Test button appears only when the last item group is displayed, thereby preventing students from inadvertently ending the test prematurely.  
- Mark-and-return is available for further review of item.  
- Any marked or all items can be reviewed before ending test.  
- AIR's operational tasks include a tool that performs test invalidations and test resets.  
- On the Test Administrator Interface, when the Approve All button is clicked, approving all students to take a test, a warning message is displayed.  
- Passage printing requires approval by the test administrator. |
| **6. Low physical effort** | - Highlighting allows students to focus on certain portions of test.  
- Strike-through allows students to focus on correct answer choice.  
- Keyboard shortcuts and area boundaries around multiple-choice options support agility accessibility.  
- Approve All button is available for test administrators.  
- Refresh Page feature is available for test administrators.  
- If students are waiting for approval, a message immediately comes up on the Test Administrator Interface demanding his or her attention. |
| **7. Size and space for approach and use** | - This refers to the physical characteristics of the testing environment and will be covered in the administration guide. |
level. AIR’s algorithm is designed to meet complex blueprints that incorporate both hierarchical constraints (e.g., distributions of items from within a content hierarchy such as strands and benchmarks) and non-hierarchical constraints that may cut across strands (e.g., item type, depth of knowledge). The features included in the blueprint are configurable and can reflect any recorded characteristic of items.

The algorithm seeks to optimize three criteria:

- Match to a complex blueprint
- Classification of students on content strands or other reporting categories
- Precision of overall estimates

The relative importance of each of these objectives in item selection is configurable separately for each test, so USOE may choose the mix that provides optimal results for each item pool.

Details of our adaptive algorithm appear in Section I.C.3.c, immediately following the section on data exchange.

**Asynchronous Scoring Framework**

Our system incorporates a modular, asynchronous scoring framework. The engine is modular in that it can accommodate new types of responses to be scored with new “engines” to score them without disrupting the current code base. The architecture enables us to “plug in” new scoring engines as we develop new item types and new response mechanisms. These engines fit within the existing framework, which manages matching items to rubrics, accepting requests to score items, and returning scores to the testing engine.

The scoring engine is asynchronous so that the system never has to wait for an item to be scored to administer the next item. Some item types may take longer to score than others. For example, graphic response items usually score in milliseconds, whereas a natural language item response may take a second or more. The same scoring framework manages human-scored items, which may not return scores for a matter of hours, days, or weeks.

Each time the adaptive engine selects an item, it draws all the currently scored items to get the working estimate of student proficiency.
Human-scored items or items that have not yet been scored are simply not used in the calculation. This approach provides the flexibility required to integrate a range of item types and scoring engines into an adaptive test, even if some of the scoring technologies require a time lag.

**Administration and Monitoring Tools**

AIR’s TDS provides a broad range of tools for monitoring testing activity as well as system health. State users and district and school test coordinators (TCs) have access to participation reports. These reports provide real-time information about the testing status of students within the user’s jurisdiction. A school TC, for example, has access to information about students in his or her school, while the district TC has information spanning the entire district. The reports provide information about current testing activity and are a powerful tool for monitoring progress toward getting all students tested. Users can filter or sort the data to identify students who have (or have not) completed any opportunities, students who have (or have not) started a testing opportunity, students whose unfinished tests are at risk of expiring if they do not complete them, and a host of other useful administrative information. Exhibit I.C.3.a-8 highlights some features of the participation reports.

State users also have access to our “latency” report, which includes information about how long various parts of the system take to deliver items. For example, it reports the average time between the server receiving an item response and (1) the database responding to the server request; (2) the overall system responding with the next item; and (3) the time it takes between sending the next item and the client machine receiving it. These measures tell instantly and accurately how the various system components are performing. We also capture client latencies, the time students wait after completing one item before the next item is fully rendered on the client screen. Client latencies benefit from the prefetch mechanism, so students experience a very responsive system.

Latency statistics are available at each level from the state through the school. Because latency reports put a substantial demand on the database system, they are not available in real time. Instead, they are calculated each night and made available for the prior day each morning.

**QM, DoR, and Reporting**

Each time a student completes a test, TDS transmits data from that test event to our Quality Monitor (QM) system. The transmitted data include

- student, school, and district identifiers;
- raw and scored student item responses;
- accommodation records; and
- initial scores (those reported to the student).

Upon receipt of the student record, the QM system

- checks that all operational items are active;
- verifies scoring of the student responses;
- validates the 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.
QM’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 online reporting system.

QM’s quality reporting interface accesses 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.

Data from the system can be extracted and delivered through an automated system our Data Extract Interface (DEI) with whatever frequency USOE desires, up to once per day. We will work with USOE to establish

- a customized data delivery format,
- a set of data validation rules for the files sent to USOE, 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.

State personnel as well as educators throughout the state have immediate access to the data through the online reporting system. That system is described in detail in Section I.C.3.e, Reporting.

**Online Portal and Training Site**

The online portal is the primary entry point for all users of the test system (other than students taking operational tests), including TAs, TCs, system administrators, and students taking practice tests. Each of these user groups has a dedicated page that provides the tools and resources needed for their specific role. In addition to providing downloadable secure browsers, the portal provides users with important news and information throughout the year. The portal from the System is shown in Exhibit I.C.3.a-9. AIR designers will work with USOE staff to design a portal.
AIR will develop the following materials for posting on the online portal:

- User guides for each public-facing component of the system
- Administration guides and manuals
- Practice tests composed of released items and of approximately the same length as the operational tests
- Item type tutorials to teach students to respond to various types of items
- E-learning modules to
  - teach TCs to operate the TA Interface,
  - teach educators to operate the reporting system and Learning Point Navigator (see Section III for a description),
  - teach TCs to operate TIDE, and
  - teach users to access the secure browser.

AIR will post other resources and materials to the portal as needed.

**Summary**

AIR’s system is modular but appears as an integrated system to users. A single sign-on system allows users to log in once and navigate between the systems. The common login system is notified when the user is no longer accessing any secure site, and it automatically logs the user off to protect data from unauthorized users.

The system offers

- integrated registration, demographic data collection, and materials ordering through TIDE;
- training for TCs, TAs, educational data users, and students;
- systems for secure student authentication and test administration;
- the ability to administer a wide range of item types, scenarios, and simulations;
- robust accommodations to make the test accessible to nearly all students (with plans to expand this to all students);
- monitoring tools to help monitor
  - system performance,
  - student and school participation, and
  - student performance (through the reporting system).
- data capture
  - demographics through TIDE,
  - administration information such as accommodations through TDS, and
  - student responses and scores through TDS.
- ongoing data quality assurance and periodic reports of statistical indicators of test quality; and
- delivery of data on any schedule (up to nightly) desired by USOE.

We are proud to offer this robust suite of services. Below, we describe the procedures that ensure the quality of our services and the security and safety of your data.
I.C.3.b Data Exchange

Data exchanges will be handled by three separate components of our system.

1. AIR’s Test Information Distribution Engine (TIDE), mentioned in our description of the overall testing process, handles uploads of files from the state or districts to our systems.

2. Outgoing, warehouse-ready data flowing to USOE are handled by our Data Extract Interface (DEI). In our description of the delivery system we describe how the data flow from the test delivery system to our Quality Monitoring system (QM) and into the Database of Record (DoR). The DFG packages and quality checks these data as they flow to USOE.

3. Outgoing data to educators, students, and families flows from QM to our reporting system, where a real-time analysis tool empowers educators to slice and dice the data along meaningful dimensions.

We describe TIDE, which gathers incoming data, and DEI, which processes outgoing data in this section. The reporting system is described at length in Section I.C.3.e.

Incoming Data: TIDE

TIDE provides an integrated system for

- gathering and managing student enrollment and registration data;
- gathering and adjusting data about students, including student demographics and accommodations or needs profiles; and
- adding, editing, and deleting users and setting them with different authorizations based on roles.

Because the TIDE system is integrated with all the other online systems such as test delivery and reporting, edits made in TIDE are reflected in real time across all systems.

TIDE offers interfaces to

- define Utah’s schools and districts and their associated adult users (such as test coordinators and test administrators) and
- upload, validate, and review student information.

We discuss these below.

Defining the Network of District, Schools, and Users

AIR or the state can define districts, schools, and users in the system. TIDE supports adding and editing for school and district information in flat files (.CSV, fixed-width, or Excel). TIDE can be programmed to connect to a secure FTP location at a predetermined frequency and import any updates to school or district data. TIDE will validate the files placed on FTP against validation/data errors and can be programmed to reject the entire file or certain records based on the severity of the errors. After the file is imported, TIDE can send a summary of the import to a predetermined set of recipients. In addition to the file-based imports, TIDE also supports a user interface to quickly modify, add, or delete a small number of entities. The file uploads prove far simpler when many entities require modification.

TIDE also manages the users in the system. TIDE can be configured to reflect the user roles that are used in the field. The roles set in TIDE will define access for users in different online systems. TIDE uses a hierarchy model that reflects the user hierarchy in the field.

We propose that Utah give us a list of districts and schools as well as an initial list of district test coordinators. We generally rely on district
test coordinators to assign user rights for others in the district as well as school-level users. TIDE allows users to be added one at a time through a user interface or multiple users to be added using an Excel or CSV template. Once users are added, TIDE will notify them through emails. TIDE also allows viewing and modifying existing users in the system.

**Student Registration**

TIDE supports uploading of student information either using automated data exchanges between the state and AIR or allowing the state users, LEA staff, or school users to upload student information directly into TIDE.

**Automated Upload by Student Information Systems (SIS)**

LEA SIS can upload and update their local agency’s student information. TIDE can be scheduled to talk to a secure FTP (SFTP) location where the SIS can place nightly student files in a predetermined file layout. TIDE services will validate the file present in the folder and import records from that file. TIDE can be configured to provide an import summary with validation and error logs and frequency distribution of the imported data. This information can be emailed to a selected set of recipients or placed in a secure location accessible to USOE. The file layout used for data upload is configurable and can be updated to account for changes in student data attributes from year to year.

**Web-Based Student Upload**

Should the LEA staff decide to upload student information directly into TIDE, the TIDE website has a step-by-step interface that walks 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-upon format. After the file is identified, the system scans it to ensure that the data match the format and allows the user to preview a few records of the file being uploaded. The file then goes through validation to ensure that the data in the file conform to the business rules set in place in TIDE. Any data format or validation errors are reported to the user in real time.

The process of uploading student information is shown in Exhibit 1.C.3.b-1.

**Validating Student Preloads and Error Reports**

The files are validated using configurable business rules. Project staff will collaborate with USOE to define the validation business rules. These rules can trigger one of three outcomes:

1. Reject the entire file
2. Reject only the offending records
3. Issue a warning

Users receive error reports and graphical indications of where in the file the problems were encountered. These messages are usually sufficient to enable district users to correct the file formats. When they are not, our help desk is always available.

The validation rules present on the upload interface are the same as that on this interface and provide helpful validation messages to the user. TIDE can be configured to have multiple validations done on each data element being added or uploaded. For example:

- Validation rules prevent LEA or test coordinators from editing information for students that are not present in their school/district.
- TIDE can also be configured to have validation rules that verify the SSID number and the student demographic information.
through an existing web service hosted by the USOE SIS.

- TIDE role-based access limits what roles have access to viewing and editing student demographic information.

- When districts upload data, it is common for multiple districts to claim the same student. TIDE can enforce business rules that prevent a district from claiming a student until that student is released by his or her current district. TIDE can also allow the most recent claim to hold, or more specialized business rules can be implemented.

AIR will work with USOE to identify custom business rules that will ensure the accuracy and validity of your student data.

**Reviewing and Correcting Student Data**

Moving to an online testing system enables Utah to move the error resolution process “upstream.” When a testing program uses paper, answer documents may be returned with any number of errors in the report identifiers or demographic data. AIR’s test delivery system ensures that only students known to the system can test. Errors such as duplicate or poorly formed identifiers can be flagged and corrected at pre-registration time, rather than after students test.

TIDE provides an interface to add students and modify the data in the student record. These modifications can be made through file uploads, but the interface also provides a convenient interface mechanism for making these corrections.

In addition to supporting the automated and website-based student upload, TIDE also supports manual adding and editing of individual student information. TIDE’s record change system allows users to update student information reliably and efficiently before, during, and after an assessment, and changes are reflected in different online systems in real time. This is particularly helpful, for example, when new students have not been properly enrolled in the system but are present in the class when testing begins. The test coordinator can quickly ensure that the student will be able to test. Exhibit I.C.3.b-2 illustrates a portion of this interface.
In addition to adding and editing students, TIDE also provides robust search features to search for students based on different attributes. In addition, the search feature has business rules to limit the search function to certain user roles and limit the search results to the user's jurisdiction (school or district).

Districts can use these features to modify student data, including demographics, at any time. Business rules can dictate that the most recent student characteristics present in TIDE are the ones reported back to the districts. In essence, districts have from the beginning of the registration window until a set date shortly before final data files are delivered to the state to amend student data.

Student Needs Profiles and Accommodations

TIDE can accept information about the special test settings or accommodations individual students need. USOE may choose to use this feature to enable schools or districts to identify required student accommodations prior to testing. As described in the Test Delivery section, USOE may also choose to allow test administrators to set this information at testing time.

User can either choose to upload student accommodation information into TIDE or use the interface shown in Exhibit I.C.3.b-3 to edit the accommodation information. The accommodation 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 USOE.

Users can add students, move students between different schools, or edit student data.
Rostering Students for Reporting: The Public Reporting Field (PRF)

The TIDE preload process can also read the PRF information and associate students with teachers. This association is one way of defining student class rosters. The online reporting system can then be configured to provide reports and aggregates to teachers based on this roster information. In addition to supporting the import of PRF information from the student upload file, TIDE can also support an upload feature just to create classes for teachers. In addition, TIDE has a roster management feature that will allow the management of class rosters so that users can add and remove students from these rosters. This roster management feature can be configured to validate rosters with the CACTUS ID web serv-
ice before committing the rosters into the TIDE database.

**Test Assignments**

As we discussed in the Test Delivery section, our systems support both implicit and explicit test registration. Most of our current clients prefer implicit registration (in which students are admitted to tests based on eligibility criteria rather than an explicit registration). This eliminates the time-consuming administrative task of explicitly assigning students to tests. However, TIDE can be configured to read the test assignment from a predetermined file layout that can be securely transferred to AIR’s secure FTP or TIDE can pull the test assignment files from USOE’s secure FTP site. This test assignment information is accessible on student search results and can be exported to Excel or CSV. Based on their access rights, users can view and/or edit the test assignment information on the TIDE website.

**Accurate and Timely Reports**

The Reporting section details our plans, including quality assurance for accurate and timely reporting. Each year we update our quality processes as part of our continuous improvement effort, and will share updated plans for quality assurance with USOE each year.

**Provide Warehouse Ready Assessment Data**

Recall from the overview of the test delivery system that data flows from this system through the Quality Monitor (QM). QM, in turn, passes this data to the Database of Record (DoR) and the reporting system. From the DoR, data flow to client systems.

AIR will work with USOE to define the specific file format to be transferred between DoR and the USOE data warehouse through our Data Extract Interface. Most of our current clients use a fixed or delimited file format, though we can also transfer data through SIFF or some other XML format. File formats are defined in a layout file that specifies the data element, acceptable values, and validation rules. Each file is automatically validated before delivery.

We can deliver data as often as requested. Most clients choose to receive nightly deliveries of all tests completed during that day. All clients receive a final end-of-year data file that incorporates final demographic edits, results of appeals and verifications, and tests that were not completed.

The specific mechanism used to accomplish these transfers will depend on the specific capabilities and restrictions on USOE’s systems as well as USOE preference. In most states, we transfer these files through a secure FTP mechanism.

Another alternative is more of a real-time link- age through a web service. Web services entail substantial overhead, and real-time transfers could tax state systems if records are passed individually at the time of completion.

We will work with USOE to define the data format, periodicity of delivery, date of the end-of-year delivery, and data delivery mechanisms.

**Annually Review Contents, Formats, and Delivery Methods**

We recognize that needs sometimes change. We also recognize a rapid, national movement to define and adopt assessment interoperability standards. Therefore, AIR commits to annual review of the content, format, and delivery methods of data. This will include both the data we receive from USOE and the data that we deliver to you.
Security

All of our systems conduct data transfers on a Secure FTP (SFTP) connection. In addition to using encrypted login and file transfer mechanisms, any access to this SFTP site is limited only to SIS data services IP addresses such that only the SIS services are able to access the TIDE SFTP folders.

Login information generated during user addition is communicated securely to the individual users. Passwords are not emailed to users. Instead, upon addition to TIDE, users are sent a secure link, one that expires after a configurable time, to set their password. Passwords in the TIDE databases are encrypted.

All communication with any of the online websites happens over SSL. This ensures that passwords and secure student are not passed over the network in clear text.

Readers are referred to Section I.C.4.b for detailed information on physical, network, and software security.

I.C.3.c Adaptive Engine

AIR’s adaptive algorithm has a well-proven track record of delivering tests that ensure that each student has an opportunity to show what he or she knows and can do across the breadth of the curriculum. We begin our presentation by clarifying a couple of terms. The delivery of an adaptive test depends on two things: the test blueprint and the adaptive algorithm. The test blueprint defines the characteristics and content of the items that will be delivered. The algorithm refers to the sequence of procedures that select the items according to the blueprint. It is the blueprint that specifies the range of standards to be administered to each student and the other characteristics of the items that make up each student’s test, such as depth of knowledge and item type. The algorithm is the procedure that ensures that the blueprint is matched, and that those items are of the appropriate difficulty.

Our algorithm seeks to optimize any or all of the three following criteria:

- Match to a complex blueprint
- Classification of students on content strands or other reporting categories
- Precision of overall scores

The emphasis of each of these objectives is separately configurable for each test.

The flexibility of our blueprint specification differentiates AIR’s adaptive algorithm from others available in the K–12 market. A blueprint can be specified on multiple dimensions, specifying both minimum and maximum numbers of items on each dimension. For example, a blueprint might specify that each test should contain between 9 and 12 items on Strand A. It might also specify that the test include no more than two items each from Benchmarks A-1, A-2, and A-3 to ensure that each test, and the set of tests as a whole, tests the breadth of content of the assessments. On an orthogonal dimension, the blueprint might specify that at least four items represent high cognitive complexity (e.g., Depth of Knowledge Level 3), regardless of the strand from which they come. Still another entry might specify the number of items of each type to appear on each test.

The second objective of the adaptive algorithm is classification of students on content strands or reporting categories. If schools and teachers are going to use the immediate feedback from the tests to guide instruction, the tests must tell them more than whether or not the student is proficient. It should tell them where the student needs help.

The adaptive algorithm can improve information about student proficiency on each content
strand or reporting category. As any psychometrician will testify, summative assessments provide very imprecise information about performance on content strands. Another way to say this is that the test leaves substantial uncertainty about student performance on content strands. We have long suggested that our clients present individual content strand information as falling into one of three categories: clearly above proficient, clearly below proficient, or uncertain. The adaptive algorithm selects items to reduce the “uncertain” range, giving more students a clear classification of above or below on content strands.

Finally, the algorithm seeks to maximize the precision of the overall score.

With each item answered, the engine recalculates the student’s score both for the overall test and for each content strand or reporting category, based on all the prior responses.

By simultaneously optimizing these three objectives and sticking to grade-level content, the algorithm
- covers the range of content specified in the state standards,
- provides actionable diagnostic information,
- offers precise overall scores, and
- has received peer-review approval for NCLB.

In general, it is not typically necessary to include off-grade items to cover a broad ability range. In fact, in most item pools there is substantially more variation in item difficulty within grade than across grades. We recommend building robust, on-grade item pools to cover broad proficiency ranges, and we recommend against going off-grade. However, our system is able to deliver standards-based, adaptive tests that go across grade levels. We sought a system that had the potential to go off-grade, and we successfully accom-

modated this while still ensuring that each student saw the full on-grade curriculum.

In , we currently offer assessments that cover the grade-level content with a sufficient number of items to cover the curriculum, while allowing a limited number of items to go off grade level along well-defined learning progressions for students who are struggling or excelling. The system is designed to meet current peer-review requirements. We implement this blended design entirely through blueprint specifications—it did not require modification of our standards-based adaptive algorithm.

**Description of the Algorithm**

As described above, the blueprint defines the characteristics of the items to appear on each test, specifying the minimum and maximum number of items to be delivered in each category. These categories generally include content (e.g., strand, cluster, standard), Depth of Knowledge, item type, and other salient subject-specific characteristics. Usually, minimums and maximums in the lower levels of the hierarchy are established to ensure that each student sees the full breadth and has an equal chance of seeing each standard, rather than forcing an item on each standard. The de facto standard for peer review has been that at least half of the standards in each reporting category be covered, though this is entirely a state-specific, blueprint issue. In cases where every student does not see every standard, we ensure that each student has an equal chance of seeing items on each standard and prevent systematic content differences across the range of proficiencies.

The algorithm is based on a multidimensional objective function. Each time an item is selected, every item in the pool has a value, and this value derives from the item’s utility in terms of fulfilling the blueprint and its match to student proficiency. Actually calculating this value
directly for each item is prohibitively expensive in terms of time. Instead, the algorithm goes through a three step process, which trims the search space, limiting the number of items for which the actual value of the objective function must be calculated.

The algorithm goes through a four step process:

- First, identify the $k$ items or item groups (for example, a group of items following a common passage) that provide the most value in terms of matching the blueprint.
- Next, sort the $k$ items in an order based on the weighted average of match-to-blueprint and contribution to performance. The weights are configurable, and are generally set during the pre-deployment simulations.
- Select randomly from the top $n$ candidates (this limited application of random selection helps control item exposure).
- If (a) the selection is an item group; and (b) the group contains more items than it is set to administer with the stimulus, then sort the items according to a configurable mix of blueprint match and performance match, and select the top $m$ items, where $m$ is the number of items designed to be administered with the stimulus.

The variables $k$, $n$, and $m$ are all configurable, as are the weights associated with each element in the blueprint and the relative weights of the components of the objective function. This configurability provides a powerful mechanism for tuning the algorithm for the item bank for each test.

**Algorithm Simulation Tool**

The adaptive algorithm and a complex blueprint have dozens of adjustable parameters. Examples include balancing the weight given to one objective or strand versus others and details of the search-trimming algorithm. The optimal values for the parameters vary depending on the item pool, specifics of the blueprints, and their interaction.

AIR has developed a sophisticated, interactive simulation tool for our psychometricians. This tool allows the psychometricians to adjust these parameters or evaluate recommended adjustments to the blueprints. In only minutes, we can try out a new configuration and evaluate its impact on the match to blueprints, precision of measurement, and item reuse. (The algorithm can be set to allow reuse of items when only a reused item will meet blueprint—a very rare circumstance that arises when item pools are deficient.)

The simulator outputs a broad range of diagnostic statistics, which include:

- estimates of bias at the overall and reporting category levels;
- standard errors and distributions of standard errors;
- match to blueprint;
- number of unique standards administered to each student;
- item exposure;
- number of item groups administered; and
- item reuse across opportunities (typically zero unless the pool is insufficient).

**Enumerated Adaptation Principles**

AIR’s adaptive engine will facilitate complete and robust measurement of the domain defined by the blueprint. This algorithm ensures that this measurement covers the full extent of the achievement spectrum, within the limitations of the item pool.

AIR will work with USOE content experts to design blueprints that accurately reflect the
full depth and breadth of the Utah Core Standards. Our item development section describes how we will develop an item pool sufficient to support these blueprints.

The algorithm adapts based on psychometrically defensible principles, as confirmed by the technical advisory committees in the states in which we work and federal peer reviewers, as well as our own psychometric experts.

Specifics requested in the RFP follow:

- Termination rules. While we can use a variety of termination rules, in statewide tests we have found that the real constraint is content coverage—most tests could terminate based on precision considerations before the range of the curriculum was covered. This, of course, would be a poor design, as students would not be measured on the full curriculum. Therefore, we generally recommend a fixed-length test, with the length reflecting the length necessary to cover the standards.

- Adaptive decision rules. These are discussed at length above. In summary, the adaptation is highly configurable, and we recommend that precision be subordinate to full coverage of a robust blueprint.

- Item exposure and release. Our system includes several mechanisms to mitigate exposure issues. Simulations in advance of deployment allow us to fine-tune the algorithm’s parameters to address exposure issues. We will work with USOE to identify appropriate item release, retirement, and replacement rules.

- Number of test pool items for each assessment. Adaptive tests are widely misunderstood on this count. Often psychometricians will suggest that an adaptive test requires up to 10 times as many items as would be administered. From a statistical point of view, any increase over the number of items to be administered on a single test will improve precision over the fixed form test—just not optimally. Pool size influences precision of measurement at the ends of the proficiency distribution and item exposure. In our section on item development we address these issues in the context of optimizing the system given budgets and test window length. In general, having at least two or three forms worth of items (appropriately distributed) beyond the number of items that will be administered usually meets most requirements.

- Functional differences in the algorithm across subject areas. Our algorithm functions the same across subject areas; however, the blueprints vary independently across subject areas. Of course, subjects that group multiple items associated with a common stimulus make use of the final step in the algorithm, while subjects that do not have no need for the final step. We expect that all subject areas will likely include item groups.

- Degree of constraint. We discussed the degree of constraint at length above. The blueprints should constrain the algorithm so that every student sees the full breadth of content, the full depth of knowledge, and full range of other salient characteristics of the content domain.

- Utilizes a selection process that measures all standards for all students. As described above, this is controlled by the blueprint. We discuss this topic at more length below.

- Provides students access to all domains required within a given year. Our system absolutely meets this requirement. We discuss this at length below as well.

With the Utah Core Standards it is possible to measure each standard for each student. The number of standards varies across grades and subjects, and USOE has the opportunity to
make different blueprint decisions across grades and subjects. However, to improve measurement Utah may want to consider blueprints that (a) ensure that each student sees the majority of the standards; and (b) ensure that the standards seen by students do not vary systematically according to student performance. This will provide incentives for the appropriate consequences within schools—every student might see every standard, so educators must teach all the material to all of the students. At the same time it will provide opportunities for the system to adapt to student performance without requiring an enormous item pool. This may prove particularly attractive in the case of ELA tests, where items are grouped by prompts. This grouping imposes an additional set of unavoidable constraints. AIR will work with USOE to design optimal blueprints.

The issue of which students see which standards is related. We strongly agree with USOE that students should not be prevented from seeing some topics because of poor performance on other topics:

- Learning in general (including mathematics) does not require a linear learning progression common to all learners.
- Items on a given topic do not cluster tightly around given points on proficiency scales because the same topic might cover tasks that are more difficult or easier.
- One or a few items contain too little information to make any judgments about whether a student can perform on that topic.

These three reasons support our commitment to standards-based adaptive testing. The first is simply a recognition that both curricula and students vary, so little can be inferred about what a student knows and can do on something that has not been tested.

The second addresses a widely circulated, but inaccurate, myth. Some organizations seek to characterize a measurement scale by the topics covered by items appearing on that point in the scale. In fact, items do not work this way. Exhibit I.C.3.c-10 presents item difficulties on one narrow content standard from one state. The proficiency scale is marked according to the state’s proficiency levels. As Exhibit I.C.3.c-10 reveals, this one narrow topic is measured across the entire proficiency distribution. We note that over 18 different standards are measured in the narrowest slice around the Proficient cutscore. Items on a common topic do not generally share a common difficulty.

Finally, a few items cannot possibly cover the breadth of most content standards. A few items cannot determine whether a student can perform on the standard or not. Basing subsequent item selection on data this thin is folly.

In summary, our approach supports the principles for adaptive testing set forth by the Consortium for Citizens with Disabilities, an umbrella group encompassing most national advocacy groups for students with disabilities (and other exceptional students):

1. Every student is tested on the full range of grade-level content with no discernible differences in the content assessed.

Exhibit I.C.3.c-10: Distribution of Item Difficulty of Items Measuring One Topic
2. Every student is tested on items measuring the same mix of cognitively complex skills, with no discernible difference—regardless of student proficiency.

3. Every student is tested on items reflecting the full range of other aspects of the grade-level curriculum as may be appropriate for the grade and subject.

4. Students are tested on items that provide the best measurement possible within these constraints.

These four principles serve the objective of making sure that every student can accurately demonstrate his or her academic skills and knowledge across the entire grade-level curriculum. AIR’s adaptive algorithm supports blueprints that meet these goals, and we propose to work with USOE to design Utah blueprints that adhere to these principles.

USOE will receive results from our simulations, and we welcome your psychometricians to join us and participate in this important psychometric activity.

I.C.3.d Scoring

This section describes our plans for item and test scoring, including field-test, pilot, and final operational items. In the broadest terms:

- With the exception of the extended constructed-response items in reading and mathematics, all items will be machine-scored with explicit machine-scoring rubrics;

- The extended-response items in reading will be human-scored during the operational field test. Those scores will be used to train Autoscore, our artificial intelligence (AI) essay-scoring engine. In subsequent operational testing, Autoscore will be used to provide an interim score. A subset of responses (approximately 20–25%) with a higher risk of having been mis-scored will be reviewed by human scorers.

The extended response items in mathematics will be human-scored.

Below, we describe the entire process. We begin by describing our process for item scoring in a discussion that covers these topics:

- Types of machine-scoring rubrics
- Refining machine-scoring rubrics through rubric validation
- Rangefinding for human-scored and automated essay-scored items
- Handscoring of responses for training AI engines and scoring human-scored items
- Overall item scoring framework
- Item scoring on the operational test

We then describe our process for test scoring, how item scores are aggregated and reported, and the quality assurance processes that are in place.

**Item Scoring**

AIR has a collection of machine-scoring engines for scoring different types of responses. Most of these engines apply explicit machine-readable rubrics, which are created by test developers in the process of creating items. One scoring engine, Autoscore, which scores more extensive responses, uses a statistical process to evaluate scores. Autoscore’s implicit rubric is derived from the statistical analysis of a set of training papers that have been human-scored. Finally, because the mathematics standards clearly require human-scoring to validly measure some of the standards, we have teamed with Data Recognition Corporation (DRC) to provide human-scoring services on this contract.
Creating and refining machine rubrics for scoring engines with explicit rubrics

AIR has five scoring engines that use explicit rubrics:

- **Multiple-choice scoring engine**, which takes an option identifier as the key.
- **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.
- **Equation scoring engine**, which evaluates the characteristics of the equations entered as the student response against an explicit test developer-created rubric.
- **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.
- **Simulation interaction scoring engine**, which evaluates a sequences of trials in a simulation item against an explicit test developer-supplied rubric.

Rubrics for each of these scoring engines is 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. We typically recommend selection of 45 responses for each item for review by the committee.

Exhibit I.C.3.d-1 demonstrates some features of this software. We select the item responses to disproportionately represent anomalous responses. Specifically, the sampling algorithm identifies examinees who scored well on the multiple-choice items but poorly on the constructed-response item being studied, as well as those who did poorly on the multiple-choice and well on the studied constructed-response item. Given these guidelines, the selection is random, ensuring representation of all responses. The balance of the sample comprises those responses fitting neither of the other two categories. By selecting equal numbers of cases from these three strata, we over-represent anomalous responses.

The committee is able to review the student responses, note observations about each response, designate a consensus score for each response, and select 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 our Item Tracking System with the results. Exhibit I.C.3.d-1 highlights some of the features of REVISE.

AIR test developers make the recommended changes to the items, and REVISE rescores all of the responses with the revised rubrics. Users can then review every changed response (or a sample of them) to evaluate whether the revision had any unintended consequences. The sample brought to the committee (or any other existing sample) can also be reviewed to
evaluate the impact of the rubric changes on those responses.

Final revisions, along with sample responses and a report of effectiveness of rubric revisions on the committee sample(s), will be communicated to USOE for final determination of which changes to implement. The ITS preserves every version of the rubric, so it is always possible to revert to an earlier version.

We plan a validation study of the equation and proposition items. For this study, a random sample of 500 cases from each item will be handscored, and discrepancies from the machine score will be reviewed and resolved. We will report a validity rate for each item.

Creating and refining human-scoring and Autoscore rubrics

Automated essay-scoring relies on accurate human scores for the engine training process. Each ECR item will be drafted with a rubric, and the rubric will be refined through a traditional rangefinding process. Handscoring requires refined rubrics, along with training materials and sample papers.

Handscoring Material Needed

Effective scoring of the ECR items will require the following material, which will be obtained through rangefinding:

- A scoring guide that includes the item exactly as viewed by the student, the scoring rubric with any additional scoring notes or criteria, and two fully annotated exemplars at each score point
- A training sets of 10 responses
- A qualifying sets of 10 responses
- 10–20 validity/recalibration responses

Rangefinding Preparation

DRC's content specialists and scoring directors will prepare for rangefinding meetings by using their Imaging Rangefinding Viewer to access student test responses and, when applicable, they will prepare sample responses from both modes—paper/pencil and computer-based. Having worked in the field for many years, DRC's proposed handscoring specialists are well aware of the nuanced differences between score points and the importance of selecting rangefinders that show the distinctions, however small they may be, among the different score points.

DRC's handscoring staff will use the scoring guidelines defined by each rubric to select a representative sampling for each rubric score point. They will be sure to include responses that take a variety of approaches toward addressing the task given in the item, seeking, for example, responses that take a novel or unexpected approach. The responses will be assembled into sample sets and duplicated for rangefinding. DRC will make hard copies for all rangefinding committee participants, including USOE staff. Each piece of material will be numbered in order to track it for a security log that DRC will keep. Additionally, DRC will ensure that receipt of boxes of material prior to the meetings will be managed by an on-site permanent employee of DRC. Prior to rangefinding meetings, all material will be kept in a locked storage area at the meeting location.

Conducting Rangefinding Meetings

DRC will open the rangefinding process by reviewing and discussing the rubric of each item under consideration. When an understanding of the scoring guidelines has been established, participants will individually score responses. Then, as a group, participants will compare scores and discuss the responses.
until a consensus is reached. Facilitators will move through each of the rangefinding sets until there are a sufficient number of responses for each item. Responses with a high level of agreement and responses that illustrate key scoring concepts (such as how to score responses with novel approaches or responses that help illustrate the cutoff between two score points) will be selected as exemplar responses.

During discussion of each response, DRC’s facilitator will take detailed notes to document the rational for assigning scores. These notes will later be transformed into official annotations for scorers to use as part of their training material.

### Post Rangefinding Activities

Once the meetings are over, AIR and DRC staff will meet with USOE staff to finalize which responses should be used for training, qualifying, and validity/recalibration. Once handsoring materials are compiled for each item, DRC will post these on a secure site accessible only to USOE staff and wait for official sign-off.

AIR will be responsible for all costs associated with meeting locations and any type of audiovisual equipment that may be needed. Additionally, a DRC meeting planning team member will be present at all rangefinding sessions to provide general support for committee members and USOE staff.
**Presented by the American Institutes for Research**

**Educator Meeting Logistics for Rubric Validation and Rangefinding**

AIR and DRC understand that we will be responsible for all administrative and logistical arrangements and costs for each of the committee meetings. DRC is well known in the assessment community for providing high-quality, responsive, and flexible meeting planning services to large-scale assessment clients. Its meeting planning and coordination experts are qualified and prepared to provide first-class meeting services for the USOE project.

Well-organized, coordinated meetings and associated administrative processes, combined with appropriate and comfortable meeting facilities and appealing meals and refreshments, help promote successful assessment programs. Over the years, DRC has gained considerable experience coordinating all types of meetings for large-scale assessment clients in more than 15 states.

DRC’s meeting coordinators work closely with other project team members to ensure that all administrative processes and meeting arrangements exceed client expectations. DRC’s staff is skilled at negotiating contracts with hotels and meeting facilities to obtain optimal rates and ensuring that meeting costs are within project budgets. DRC provides full-service meeting administration services to clients as well, including overall organization and coordination for meeting participants; developing and distributing letters and email communications to attendees; tracking participant registration; and processing invoices, travel vouchers, stipends, and teacher substitute reimbursement forms to ensure proper payment. DRC’s meeting coordination staff excels at providing exceptional customer service to meeting attendees. In 2011, DRC planned 80 meetings for more than 2,600 participants in 12 states.

DRC will provide the following logistical services for the Rangefinding Educator Committee meetings:

**Meeting Location**—DRC will work with USOE to identify easily accessible and cost-effective location(s) for the in-person meetings (i.e., a city with a major airline hub and the capacity to meet the logistical demands of large-scale, multi-group meetings). Meetings will be held at a hotel, conference center, or similar suitable location. Where possible, meetings will be scheduled concurrently and in the same venue to reduce costs.

**Facilities**—DRC will make all facility arrangements, including securing appropriate meeting rooms and supplying all technology, audiovisual equipment, and supplies needed to conduct the meetings. DRC will also take measures to ensure that the meeting rooms are secure.

**Meeting Planning Documentation**—Prior to each meeting, DRC will provide a Meeting Planning Document to USOE that outlines the parameters of the meeting, participants, logistical information, required materials and resources, and other details as required by USOE. Upon USOE approval, DRC will produce meeting schedules, agendas, and participant lists for each meeting. At the conclusion of each meeting, a follow-up report containing meeting notes, action items, a summary of results, and any other requested information will be provided to USOE.

**Meeting Facilitation and Management**—DRC’s project manager, along with a DRC meeting planner/administrative assistant and appropriate content staff from DRC and AIR, will be present at meetings as needed to provide overall meeting management and support to USOE.

**Signed Agreements**—DRC requires signed security agreements from all committee meeting participants and retains the agreements for the duration of the contract. At the beginning
of all rangefinding meetings, each participant will be asked to sign a confidentiality letter specifying the confidentiality agreement and security regulations. The agreement will also outline ownership regulations. DRC will ensure that no confidential materials related to the project will be released without USOE explicit approval.

The original copies of these forms will be retained in secure storage and copies will be made available to USOE upon request.

**Secure Materials**—During the rangefinding meetings, secure materials will never be left unattended. DRC staff will monitor the security of all testing-related materials throughout the entire process. All materials sent to meetings are sent through a secured mailing process and have tracking documentation. DRC will number each set of materials used during the meetings so that any missing material will be immediately noted when materials are checked in and out each meeting day. When materials are not in use, they will be kept in a locked space. DRC staff is vigilant about maintaining security at these meetings.

DRC will work with USOE to ensure that USOE's preferences and requirements for meetings are understood and met. All prices associated with meeting planning services and deliverables have been provided in the Cost Proposal.

**Educator Committee Travel Costs**

**Travel Arrangements**

AIR and DRC will make all necessary travel arrangements including securing car rentals, air transports, and hotel rooms for participants. DRC will reimburse participants for travel-related expenses including airfare, car rental, hotel accommodations, meals, and other relevant expenses including tolls, parking, cab fare, and vicinity mileage for personal vehicle use.

DRC will work with USOE to determine the most suitable plan for meals, whether it be DRC-provided breakfast, lunch, and dinner options at each meeting, OR to reimburse participants for their expenses to obtain their own meals. If the latter is chosen, DRC will ensure that meeting locations afford a variety of close-by, cost-effective restaurant options.

**Communication with Participants**

DRC assumes that USOE will identify the committee members who will participate in each meeting. Based on participant lists provided by USOE, DRC will prepare and distribute meeting notifications/invitations and logistical information to committee members, USOE attendees, Achieve attendees, and consultants. DRC will also coordinate any follow-up communications to meeting participants after the meetings are concluded, as required.

**Consulting Fees and Substitute Reimbursements**

We understand that AIR will be responsible for determining eligibility for consulting fees and substitute reimbursements, processing payments to committee members, and maintaining records of all transactions. Financial support to committee members will include the payment of a $100 per day consulting fee for each meeting day not paid through the participant's employer. Substitute teacher reimbursement of $100 per day will be paid to school districts for each day a substitute is needed for a teacher attending a USOE meeting.

**Handscoring**

For both the field test and the operational test, handscoring will be required. This section details our procedures and quality assurance
processes during handscoring. Handscoring will be led by our proposed partner, DRC.

DRC brings a tremendous amount of experience scoring student responses from short and extended constructed-response items, performance tasks, and writing prompts for many state assessments, and DRC has total confidence in its ability to perform a superior job for USOE. DRC has 30 years of experience in providing accurate scores for millions of student responses for numerous clients, including large-scale assessments for 

DRC’s full-time permanent content specialists have many years of experience in conducting range-finding meetings and preparing quality training materials, including selection of exemplars for each rubric score point and accompanying annotations. They have annotated responses for all grades, including advanced mathematics subjects and literary analysis. DRC’s handscoring staff is on-site during scoring, which is advantageous in working with room scoring directors to interpret daily and cumulative scorer statistics. They are able to identify potential scoring issues and proactively intercede, as needed.

DRC’s clients will confirm that DRC consistently meets its range-finding and handscoring deadlines while maintaining focus on quality throughout all processes associated with handscoring. DRC believes that experienced personnel, precise training materials, and thorough quality control measures are all essential elements in its continuing success in the field of handscoring and would welcome the opportunity to provide USOE with this excellent service.

### Scoring Centers

DRC proposes to use physical scoring centers (not distributed scoring) to accomplish the handscoring for the field test and operational test. DRC wants its content specialists who attend range-finding to be on-site during scoring so they can quickly respond to any issues/questions that may arise once scoring starts.

DRC operates seven handscoring centers throughout the United States, occupying over 315,000 square feet of office space. These sites are owned or leased year round and have been in operation for a minimum of three years. Each site is ISO 9001:2008 certified in Performance Assessment Scoring; DRC believes that this sets it apart from its competitors in terms quality.

### Security

DRC understands that security is paramount to the success of this program. DRC’s scoring centers are secure facilities. Access is limited to staff and to visitors accompanied by authorized staff. Doors to the building remain locked at all times. Scorers are given electronic keycards to enter the building. Their keycards allow access only to the scoring rooms and break rooms; warehouse spaces and libraries of training materials cannot be accessed by the scorers. The scorers’ keycards allow access only during work hours. The keycards can be remotely disabled by the site facilitator.

Student responses are kept secure at all times. All readers are made aware that no scoring materials are to leave the scoring center, and they must sign legally binding confidentiality agreements stating that they are aware of the secure nature of the work before training and scoring begins.
Most of the training material is scanned in and presented to scorers on DRC’s Image Handscoring System. These online training materials cannot be printed or copied from the scorers’ image-scoring stations. Access to the scorers’ computers is controlled by usernames and passwords. Furthermore, each scorer has a unique and secure ID and password that must be used to log into DRC’s Image Handscoring System in order to prevent unauthorized access. Scorers cannot copy imaged responses from within DRC’s Image Handscoring System.

Some of the handscoring materials may be distributed to scorers in a paper format. DRC proposes that these secure materials be printed on light green paper in order to make them instantly recognizable as secure material. When the materials are assigned to the scorers, an ID number associated with the scorer will be written on the material so that each scorer’s secure material can be tracked. The secure hardcopies will be accounted for daily by signature when distributed and collected, and they will be maintained in locked storage that can be accessed only by authorized staff.

Confidentiality Agreements

DRC has every scoring center employee sign a confidentiality agreement prior to viewing secure items or student responses. DRC often uses confidentiality agreements that have been written by specific state department clients but will use its standard agreement if none is supplied. The standard agreement states, “I understand that, as an employee for Data Recognition Corporation, I will come in contact with materials (scoring rubrics, training materials, student work) that are secure and confidential. By signing below, I agree to:

- Leave all training materials and student responses at DRC.
- Refer requests for information about particular projects to DRC management.
- Refrain from commenting on students’ performance with non-project related personnel.
- Refrain from producing, in part or in whole, student responses.
- Respect, in general, the privacy of the students whose work I evaluate.

Failure to abide by this agreement can result in termination of employment with Data Recognition Corporation.”

Scoring Site Staff

DRC offers experienced personnel who are unsurpassed in the industry. Its scorers are dependable, accurate, and instrumental in meeting strict reporting deadlines. Team leaders are articulate and task-oriented, and the scoring directors are strong content leaders who are exceptional facilitators of rangefinding activities and trainers of handscoring practices.

Scorers, team leaders, and scoring directors are selected who are concerned with the task at hand and, most importantly, flexible. They are valued for their experience but, at the same time, are required to set aside their own concepts about student performance and accept the scoring standards of the client’s program.

With numerous colleges and universities near its scoring sites, DRC is able to draw upon excellent and well-educated labor pools. This, in turn, allows DRC to tailor its handscoring staff to each client’s specific program. All scorers and supervisors will have, at minimum, a four-year college degree and a demonstrated ability to write. At the personal interview, scorer candidates are asked to demonstrate their writing proficiency by responding to an open-ended DRC on-demand writing prompt. Additionally, any potential employee who interviews for a handscoring mathematics position will also be asked to solve two high
school mathematics questions and to show work.

DRC has a human resources coordinator dedicated to recruiting and retaining scorer staff. DRC does not use a temporary employment agency. Applications for scorer positions are screened to create a large pool of potential scorers. In this process, preference is given to candidates with degrees emphasizing the appropriate content areas and experience in scoring large-scale assessments.

**Scorer Training**

DRC recommends rangefinding enough responses to create the following handscoring material:

- A scoring guide that includes the item exactly as viewed by the student, the scoring rubric with any additional scoring notes or criteria, and two fully annotated exemplars at each score point
- Two training sets of 10 responses
- Two qualifying sets of 10 responses
- 10–20 validity/recalibration responses

DRC will provide team leaders who will assist the scoring directors with scorer training and monitoring. The scoring director will oversee comprehensive team leader training lasting two to three days depending on the content. Team leader training will follow the same procedures used in the scorer training (detailed below) but will be more comprehensive due to the monitoring responsibilities required of the team leaders. To promote item-level scoring consistency, it is imperative that each team leader imparts the same rationale for each score assigned. Consistent training by the team leaders will result in scorers assigning accurate scores for the appropriate reasons.

Once the team leaders have qualified, they will prepare for the arrival of their teams of scorers. Training of scorers will begin with a presentation and discussion of the item, rubric, and exemplars by the scoring director or team leader. Next, the scorers will practice by scoring the responses in the training sets. Afterwards, the scoring director or team leader will lead a thorough discussion of each set.

After the exemplars from the scoring guide and all training sets have been discussed, scorers must demonstrate their ability to apply the scoring criteria by qualifying (i.e., scoring with an acceptable agreement with true scores on the qualification papers). Any scorer who does not qualify will not be allowed to score actual student work.

DRC insists that scorers must be present for all training sessions or they will not be allowed to score that particular set of items. Staying on schedule is of utmost importance in order to report results in a timely manner, and DRC has enjoyed on-time successes as a result of its carefully implemented and managed scheduling procedures.

**Handscoring Procedures**

DRC’s training, scoring, qualifying, and monitoring processes, described below, are the best in the industry. All of these processes have been used for years to score numerous large-scale assessments and will be upheld under this new contract:

- Pairs of scorers will be seated in ergonomically adjustable chairs at long rectangular tables. There will be two imaging stations at each table. Each workstation will include a large, adjustable flat-screen monitor for clean image reproduction and easy viewing. Each scorer will be assigned a unique ID number and password to ensure security.
The scoring director will explain, in detail, how to use the computerized handscoring system. The scoring director will also provide a written document containing the instructions in the Imaging Handbook, created specifically for DRC scorers. The scoring director is physically in the same room as all of the scorers for one item.

The student responses are separated for scorers by grade and subject. Images from specific items will be sent, in randomized order, to designated groups of scorers certified to score those responses. Only qualified scorers will have access to student response images. The scorers will read each response and enter the correct scores. After scores are entered, a new response image will appear. No student demographic information appears on any image.

Ten percent (10%) of responses will receive two independent readings to monitor interrater reliability. If the two readers’ scores do not match (exact scores for a 2-point item and adjacent scores for 4- and 6-point items), a third and sometimes a fourth reading will be required. DRC's imaging system ensures that these responses are properly routed to scorers who are certified to score that item. Scorers do not know whether they are reading a response for the first or second time: all first and second readings are “blind.” Third and fourth readings are typically done by scoring directors.

Ongoing quality control checks and procedures will monitor and maintain the quality of the scoring sessions. If any unusual data are observed, DRC will investigate and resolve any issues.

Routing and scoring sets of imaged responses continue until all items have received the prescribed level of first and second readings.

DRC’s Image Handscoring System allows for on-demand retrieval of specified images (e.g., specific batch files, specific grades, specific students) should the need arise during or subsequent to the handscoring process.

**Image Handscoring System**

Now in its tenth year of operation, DRC’s Image Handscoring System has proven to be highly efficient and completely accurate for scoring large-scale assessments. This dynamic system allows scorers to score items online, increasing efficiency by eliminating the routing of paper and eliminating the possibility of lost or destroyed student answer documents. Instead, imaged responses are electronically routed to geographically dispersed DRC scoring centers. Responses are allocated to scorers through a custom dealer program, ensuring that each scorer is assigned a random workload that allows the project to be processed in the most fair and efficient manner.

Within the Image Handscoring System, each open-ended item’s position within the student answer document is defined through the use of the programmatic item definition application. For each item, the system also requires the definition of the possible score values, the possible nonscore values, the applicable scoring rule (e.g., 10% of responses are read twice), and the scoring center at which the item will be read. The Image Handscoring System functionality also requires scorers to forward all nonscorable responses to the scoring director. Only the scoring director is able to assign nonscorable codes.

The system provides the scorers with the ability to view full-page images from a variety of perspectives, such as zoomed in/out, flipped, and/or rotated, in order to accurately read student responses. Images remain intact with the various viewing capabilities and cannot be
# Exhibit I.C.3.d-2: Handscoring Quality Control Reports

<table>
<thead>
<tr>
<th>Report</th>
<th>Report Specifics</th>
</tr>
</thead>
</table>
| Scoring Summary Report  
(Inter-rater Reliability, Score Point Distribution, and Production Rates) | DRC's Scoring Summary Reports show inter-rater reliability (agreement rates), score point distribution, and production rates for every reader and every item being scored. DRC also has an Item Summary Report that summarizes the inter-rater reliability and score point distributions at an item level, which is useful for a high level overview of handscoring quality control. |
| **Inter-rater Reliability (Agreement Rates)** | Monitors how often scorers are in exact agreement and ensures that an acceptable agreement rate is maintained based on all of the responses that are independently scored twice. This report provides daily and cumulative exact and adjacent inter-scorer agreement. The calculations for this report are as follows: |
| ▪ **Percent Exact**—Total number of responses by scorer where scores are equal, divided by the number of responses that were scored twice |
| ▪ **Percent Adjacent**—Total number of responses by scorer where scores are one point apart, divided by the number of responses that were scored twice |
| ▪ **Percent Non-Adjacent**—Total number of responses by scorer where scores are more than one score point apart, divided by the number of responses that were scored twice |
| **Score Point Distribution Report** | Monitors the percentage of responses given each of the score points. For example, for the open-ended writing responses, this daily and cumulative report shows how many 1s, 2s, 3s, and 4s a scorer has given per analytic trait for all the responses he or she has scored prior to the time the report is produced. It also indicates the number of responses read by each scorer, so that production rates can be monitored. |
| **Production Rates** | Monitors the number of reads performed by each scorer and the number of reads completed for each item. |
| **Item Status Report** | Monitors the progress of handscoring. This report tracks each response and indicates the status (e.g., “needs a second reading,” “complete”). This report ensures that all discrepancies are resolved by the end of the project. |
| **Responses Read by Scorer Report** | Identifies all responses scored by an individual scorer. This report is useful if any responses need rescoring due to potential scorer drift. |
| **Validity Report** | Provides validity results at an item level and for each individual scorer. |
| **Read-Behind Log** | Used by the team leader/scoring director to monitor intra-rater reliability. Team leaders read a random selection of scored responses from each team member. If the team leader disagrees with the scorer's score, remediation occurs. This has proven to be a very effective type of feedback because it draws from items live-scored by a particular scorer. |
modified by the scorers. Additionally, the Image Handscoring System functionality applies a set of process rules and client-defined read-behind criteria.

Each handscoring site is connected to the main DRC operations facility with multiple T1 transmission lines. The operations facility has secure database servers and multiple applications that support the scanning, editing, scoring, and handscoring processes. Database backups and archived images are stored off-site on tape media for disaster recovery purposes. Each DRC scoring site has a server and a local area network (LAN). Scorers, team leaders, and scoring directors connect to the LAN via hundreds of PC workstations and use locally resident software to view and score student responses. Authorized on-site DRC personnel (e.g., content specialists, project managers) can access the LAN to view images of any student document.

Quality Control of the Image Handscoring System

Software quality assurance analysts test the imaging system to verify that all handscoring programs are compliant and in place for performance assessment personnel prior to the transfer of production images. Images produced from test scan files are randomly distributed to handscoring computer terminals, where quality assurance analysts score the test images using project-specific handscoring criteria and specifications. Throughout this testing cycle, multiple quality checks are executed to ensure that the data integrity for each student record is intact and accurately reflected in the scoring database.

Quality Control of Handscoring Procedures

Accurate and consistent results are the backbone of all handscoring activities. Exhibit 1.C.3.d-2 summarizes quality control reports for handscoring. The following methods used by DRC guarantee scoring quality.

- Inter-rater reliability. Ten percent (10%) of all responses will be independently read by two scorers for the purposes of monitoring inter-rater reliability. All readings are “blind”; scorers cannot tell whether a response has or will be scored by another scorer. In order to monitor scorer reliability and maintain an acceptable level of scoring accuracy, DRC will closely review reports that will be produced daily. These reports document individual scorer data, including scorer number, number of responses scored, and exact agreement rates. DRC will investigate any issues and resolve any problems identified by the reports. For nonadjacent scores, third (and even fourth) readings may be necessary for 4- and 6-point items.

- Read-behinds by team leaders and scoring directors. DRC’s imaging system allows team leaders and scoring directors to determine read-behind rates (frequency of monitoring) for each scorer. DRC typically has a ratio of one team leader for every eight scorers. If, however, a scorer needs clarification on scoring rules or is scoring tentatively, DRC will increase the read-behind percentage. The imaging system randomly selects which images the team leader will read behind.

- Validity responses. Validity responses are distributed periodically throughout a scoring session to re-focus the scorers on USOE’s scoring criteria. They have been prescored by rangefinding committee members and/or by USOE staff and DRC content specialists. Validity responses will be used throughout the scoring process to monitor scoring accuracy by comparing each reader’s scores to predetermined scores. Because the validity responses are “seeded” to readers during scoring sessions, they are a powerful tool for monitoring readers.

- Handscoring quality control reports. DRC is proud of the scorer quality control
reports developed for its handscoring projects. Please be assured that USOE will have full online access to handscoring reports. DRC’s handscoring quality control reports are described in the following table.

Monitoring Scorers

During the handscoring process, the scoring directors will meet several times a day with their team leaders to review daily statistics. If scoring patterns become apparent among scorers, each team leader will deal with these issues on an individual basis. As mentioned earlier, DRC’s imaging system allows a team leader to determine read-behind rates (frequency of monitoring) for each scorer, but please be assured that a minimum of 10% will undergo a read-behind from a team leader.

DRC will also monitor the inter-scorer agreement. If a scorer falls below an acceptable rate of agreement, the team leader will retrain the scorer. If a scorer fails to improve after retraining and feedback, DRC will remove the scorer from the project. If necessary, DRC will remove all assigned scores given by the scorer in question. The images will then be re-dealt and rescored.

DRC does not report on scorer performance after the fact, as some contractors do. DRC believes that scorers with less-than-acceptable scoring patterns must be identified immediately and those patterns corrected. DRC has worked diligently to devise effective monitoring reports and procedures to ensure both detection and correction. DRC will provide USOE online access to its daily handscoring quality control reports.

Overall Item Scoring Framework

When students are taking an online adaptive test, item responses must be scored in order for the system to adapt. Upon each student response, the appropriate scoring engine is called through an asynchronous scoring framework.

AIR’s system incorporates a modular, asynchronous scoring framework. The engine is modular in that it can accommodate new types of responses to be scored with new “engines” to score them without disrupting the current code base. The architecture enables us to "plug in" new scoring engines as we develop new item types and new response mechanisms. These engines fit within the existing framework, which manages matching items to rubrics, accepting requests to score items, and returning scores to the testing engine.

The scoring engine is asynchronous so that the system never has to wait for an item to be scored to administer the next item. Some item types may take longer to score than others. For example, graphic response items usually score in milliseconds, whereas a natural language item response may take a second or more. The same scoring framework manages human-scored items, which may not return scores for a matter of hours, days, or weeks.

This allows the adaptive engine to adapt to the greatest extent possible, given the scores that are on hand at the time an item must be chosen. It provides students with a smooth, uninterrupted testing experience and enables AIR to offer an ever-expanding array of measurement options.

Item Scoring on the Operational Test

On the operational tests, both summative and interim, the vast majority of items will be automatically scored through the real-time, asynchronous scoring framework. There are two exceptions:
The ECR item in the ELA performance task, which will be scored by the essay-scoring engine, but evaluated and potentially routed to human scoring based on heuristics that we use to identify suspect scores.

The ECR item on the math performance task, which will always be human-scored.

Given the lower-stakes nature of the interim assessment, we propose to use only the automated scores (no human follow-up) for the ELA interim assessment. We propose to send all math ECR responses for human scoring for both the interim and summative assessments.

### Test Scoring

Test scoring is the process of assigning overall scores, performance levels, strand scores, and other scores to completed tests. We propose to assign maximum likelihood scores, as discussed in Section I.D, Psychometrics. The process through which these scores are assigned is highly automated and results in error-free scores.

The scoring process begins before deployment of the online system. Above, we discussed the process through which our test administration simulator is used to tune the adaptive algorithm before deployment. As part of this process, we configure the test scoring engine. Automated checks ensure that the resulting scores are accurate and free of bias. In addition to the automatic checks, our psychometrists review and replicate a sample of the data from each scoring configuration.

This process ensures that the scores coming from our automated system will be accurate.

When students compete tests, the entire record is passed to Quality Monitor (QM), a component that assigns scores, checks data from each completed test for completeness and accuracy, as well as conducts ongoing analysis to detect breach, cheating, and item drift. QM is discussed in Section I.C.3. These data are then transferred to our Database of Record (DoR) and real-time reporting system, both of which were discussed earlier. Please see the section on Data Exchange for a discussion of the quality checks around the data to be delivered to USOE.

### A Meaningful Scaled Score

Please see Section I.D, Psychometrics, for a discussion of how we propose to establish a scale that is meaningful across grades and substantially linked to the progression of standards. In particular, please see the section on standard setting.

### Quality Assurance Process

We discussed our scoring processes above, and the quality assurance steps are integral. As that discussion revealed, our approach is distinguished by four features that ensure quality and robustness:

- Our process is highly automated, reducing the opportunities for human error.
- Where processes cannot be automated, they are independently replicated.
- Our online system includes a real-time quality monitoring system that flags potential issues.
- Our REVISE system for reviewing student responses lets committees and clients see the responses exactly as the students did when they entered them.

Our general approach to quality assurance is based on three tenets:

- All humans, no matter how committed or talented, make mistakes. Therefore, no high-stakes system can allow a failure by
a single individual to cause error in the system.

- Systems that are not used do not really exist, so all staff must know and be individually committed to using those systems.
- No system is flawless, so eliminating errors requires constant improvement.

In response to this belief we have built systems that

- use multiple layers of review when human knowledge, talent, innovation, and insight are required by the product or service;
- train staff so that everyone knows the systems and understands their import, and build systems that are self-enforcing (for example, the ITS software will not allow reviews to be skipped);
- continue to build new systems and infrastructure to deliver a better product, faster, more reliably, and less expensively.

We remain committed to continue building systems with these characteristics. Our partner DRC shares this view and commitment to structured systems and continuous innovation.

AIR’s online testing system includes a real-time quality monitoring component. Each time a test is completed, responses to the items are added to aggregates. These aggregates can be used to flag selected-response items when the key(s) is not the most common response, items that very few or very many students get correct, and other potential red flags.

This provides a backup system in the unlikely event that pre-deployment quality assurance and control mechanisms let problems pass through. If this were to happen, the quality monitoring system can identify and correct problems early. Under operational conditions, where item parameters are available in advance, this system can give early warning of item drift that may signal a compromised item.

The quality monitoring system is also able to flag suspicious activity that may reflect testing irregularities.

Finally, our REVISE system supports committee review of student responses. The system can automatically select samples of student responses according to a variety of different sampling designs. Some of these sampling plans are explicitly designed to oversample responses that have the greatest likelihood of being anomalous. REVISE presents the responses exactly as the student saw them and provides the ability to modify the rubrics and view the intended and unintended effects of the modifications. When rubric changes are made during the committee process, REVISE will automatically rescore the data according to the new rubric. REVISE prevents these data from being exported until our content experts have reviewed the results of the rescore, enforcing our quality assurance procedures.

These four features—automation, replication, real-time monitoring and the REVISE process—will all contribute to error-free scoring of USOE tests.

### Scoring Methods for Item Types

The introduction to this section described the scoring engines that will be used for all items and item types on this assessment. Further discussion of item types and examples can be found in Section II.A.3.

### Scoring Decisions

These scoring decisions are made in part during item development and review and in part during rangefinding or rubric validation. Those processes are discussed in detail above. Field-testing helps validate those decisions by demonstrating the progressive difficulty associated with successive score points. This aspect
of the validation is discussed in Section I.D, Psychometrics.

I.C.3.e Report System

Introduction

The RFP envisions a reporting system that meets the needs of stakeholders while fostering communication among parents, educators, and students and contributing to instructional decision-making. AIR’s proposed system is designed to provide timely, relevant reports to various stakeholder groups while guiding those stakeholders to make valid, actionable interpretations of the data. Our reporting system is designed to leverage the information available in computer adaptive test results to help bridge the gap between interim assessment results and targeted formative assessment.

Details and Samples of Reports

To help guide valid instructional decision-making, the reporting system must provide meaningful, actionable reports. As test results come pouring in from the computer adaptive tests, the reporting system updates every report in real time so that stakeholders have immediate access to the information they need about students’ strengths and weaknesses.

Navigation in the reporting system is designed to mirror 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 content strand 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.

Each navigational step changes the reporting display, providing richer context when interpreting a class’s or individual student’s performance. The system, of course, contains many reports. However, the interface design encourages users to think about the substantive, educational questions to which they need answers and access information from that perspective. Users find the system, which is operational in , to be intuitive and user-friendly.

Benchmark Reports

The reporting system leverages the information from the computer adaptive interim and summative tests to make suggestions for formative assessments that may be relevant for a group of students. The adaptive nature of the test presents an opportunity to offer the kind of diagnostic information that has eluded summative and interim assessments in the past. Educators want very precise information about the bits and pieces of knowledge or skills that students have or have not mastered. This sort of fine-grained information is actionable; it tells educators what to teach and lets them know where they are succeeding. On fixed-form tests, this detailed reporting has proved elusive. A typical fixed-form test may test a benchmark with only one or two items. Even when aggregated, these data reflect the benchmark only narrowly because they reflect only one or two ways of measuring the benchmark.
An adaptive test offers a tremendous opportunity for benchmark-level data at the class, school, or district level. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given benchmark. A mechanism for aggregating them that accounts for the differences in student overall proficiency and differences in the difficulty of the items can provide a robust measure of a class’s, school’s, or district’s mastery of that benchmark. Here, we propose just such a measure.

Specifically, we propose to report benchmark information at aggregate levels (e.g., school, class, group of students). We propose that this information be presented in terms of relative strengths and weaknesses so that a teacher sees whether a class’s performance on a particular benchmark is better than, worse than, or about the same as their performance on the test as a whole. We currently report this information in , and for individual benchmarks.

To make this report actionable for educators, the reporting system provides direct links from a single benchmark on the report to relevant assignments, content, and activities in the formative assessment system. A teacher may observe that his or her class is performing worse than expected on a given benchmark, given their performance on the test as a whole. The teacher can immediately link to the formative assessment system, create an assignment for his or her class based on the content of that benchmark, and then see how individual students perform on that assignment. This link between the interim and summative assessment results and targeted formative assessments gives educators the tools they need to investigate how individual students’ gaps in knowledge or misconceptions are translating into test performance. Exhibit I.C.3.e-1 shows a sample benchmark report for a class where relative strengths and weaknesses are shown for each benchmark. The formative assessments are accessed with the [Find Resources] button.

Subject Summary Reports

Aggregated subject reports show average performance for the state, LEAs, 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

Exhibit I.C.3.e-1: Benchmark Report for Computer Adaptive Test
Exhibit I.C.3.e-2: Subject Summary Report for a Teacher’s Classes

Subscore Summary Reports

Aggregated subscore reports follow the layout of the subject summary reports, displaying subscore performance data for the state, LEAs, schools, teachers, and classes. Exhibit I.C.3.e-3 shows a sample subscore summary report.

Trend Reports

Scores for the state, LEAs, schools, teachers, classes, and students can be plotted on a trend report to illustrate how performance has changed over time. Scores can be plotted across years (e.g., from spring administration to spring administration) and within a year (e.g., from fall administration to spring administration). Multiple trend lines can be plotted at once to see how changes in performance compare between students or groups of students. Subscore trends can also be added to see whether there are differences in performance within a subject. Exhibit I.C.3.e-4 shows a sample trend report with two students plotted on it.

Individual Student Report

Individual student reports summarize a student’s performance in an organized, easy-to-understand document that can be distributed to educators, parents, and students. The student’s performance is plotted against cut scores on a barrel chart that provides detailed explanations of each performance level. Multiple test opportunities can be shown on a single report, and a student’s subscores and comparison data for the state, LEA, and school are provided in separate tables. The report can be exported as a PDF document, and users can batch print multiple students’ reports simultaneously. Exhibit I.C.3.e-5 shows a sample student report.

Participation Reports

Computer-based testing can provide more flexibility to schools in terms of when they administer tests; however, schools still need test participation information to manage their testing schedule, allocate testing resources, and prioritize testing. The reporting system’s Test
Exhibit I.C.3.e-3: Subscore Summary Report for a Teacher’s Classes

Management Center provides multiple tools to generate up-to-the-minute reports showing students’ test status. Users can set testing schedules, monitor testing progress across schools and districts, and even track students’ participation based on their performance on previous tests. Exhibit I.C.3.e-6 is a sample of the summary statistics report from the Test Management Center. This report provides a high-level view of how many students have completed each test opportunity and the percentage of students proficient by opportunity and overall.

Meeting System Requirements

Data Review and Correction

For data to be actionable, they must be accurate. For this reason, AIR has extensive quality control procedures in place to allow for data review and correction prior to reporting results. In the unlikely event that data need to be modified after a test administration, the online reporting system can be quickly updated during non-business hours, minimizing any interruption in the field.

Though results from the computer adaptive tests seem to flow into the
reporting system instantaneously, the results actually must pass through a configured quality assurance process before being reported. The testing system can also be set up with test codes for certain testing events (e.g., incomplete tests), and the reporting system can be configured with specific reporting rules for such cases (e.g., suppress results).

AIR will work closely with USOE to define all such configurations and reporting rules in a reporting specifications document that clearly states all reporting rules, system configurations, and display options. As the reporting system is built, AIR’s experienced Quality Control team carefully vets every page in the reporting system to check for proper functioning and display of results across all supported browsers.

Once the reporting system is built, AIR conducts a User Acceptance Testing period where

USOE has the opportunity to test the functionality of the site against its specifications before the site is launched. AIR provides detailed checklists to help guide the USOE through each report and every available function to ensure that everything meets expectations. AIR reporting staff work closely with their USOE counterparts during this time to address questions and issues.

In the rare event that there is a reporting issue following a test administration, AIR’s production control processes allow us to react immediately to suppress affected sections of the reports, make the correction, and update the reports following an expedient review by our Production Control Board. AIR can also provide communication to the field through the portal notifying users of any changes. AIR’s rigorous quality control process means that corrections following a test administration are extremely rare; however, we are prepared to take corrective actions with the live reporting system should the need ever arise.

**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), which allows educators and administrators to seamlessly move between authorized programs without having to reenter their login information. Secure access rules can also be set up to allow parents to view their child’s individual student report electronically. When a user logs in for the first time, he or she must agree to the terms of a Non-Disclosure Agreement.
As with all of our systems, student data are protected by the same physical, network, and software security as our test delivery system. Please see the extensive security section in the test delivery section for more details.

### Transfer of Return Files to USOE

Student data files are provided to USOE on a regular basis and are transferred via a secure FTP site. USOE can specify how frequently the data files are posted (e.g., nightly, weekly) as well as the data file layout.

### Immediate Access to Results

For tests delivered via computer, the reporting tool displays results within seconds of the test being completed. Users are able to immediately view individual student’s results and aggregate reports for their rosters, schools, and districts. Students can also see their score immediately at the end of the test, though this can be suppressed if desired.

Additionally, the Test Management Center tools will provide real-time information on test participation. For example, if a student needs to pause his or her test, the reports will immediately show that this student’s test has been paused.
Translation of Reports

USOE may want the individuals student reports translated into up to five other languages. Our system supports this.

System Capabilities

Exportable Data

The reporting system’s Retrieve Student Results tool allows users to export a raw data file for an LEA, school, teacher, or roster. These data files can be configured to include students’ identifying information, student demographic information, test participation status, and performance information at the subject and subscore levels. The files can be exported for import into spreadsheets or other software or any other delimited file format so that they can be easily loaded into other software platforms. The reporting tool will also temporarily store the file for a set period of time so that the user can access it later.

In addition to the raw data files in the Retrieve Student Results tool, users can export any table on any report in the reporting system. An “Export” button is available on every report, and the user can immediately export the table as an Excel file.

Customizability

All text and measures in the reporting tool are configurable, and AIR reporting staff will work closely with USOE to define all text and measures as part of the specifications development. The reporting system can accommodate different kinds of data, and various reporting elements are available (e.g., barrel charts, bar charts, trend charts) to help best display those data.

The user can also customize his or her reports with some of the system tools. Users can choose to show or hide comparison data on their aggregate reports. Users can also choose to show or hide individual columns in a table. In addition, a user can sort most columns in ascending or descending order to group reports by things like score, performance level, or demographic subgroup.

Drill Downs

The report navigation is designed to allow the user to drill down through the different dimensions of their 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 want to switch from subject-level reports and drill down into the subscore reports or benchmark 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 couple of clicks without having to go back to the beginning and drill down again.

Aggregation and Disaggregation

In addition to the expected aggregated reports for an LEA, school, teacher, and class, the reporting tool allows users to create custom rosters of students so that they can be tracked and reported on separately. For example, a teacher may have a class of students, and be tracking his class’s test scores; however, he also has a group of students in his class who are receiving extra help after school. He can create a second roster of only those students to report on their aggregate performance from the group. This may help the teacher to begin to answer questions about the effectiveness of the extra help or the strengths and weaknesses
of this subgroup of students. All aggregated reports are calculated from the student level up, so as new students test, reports for the roster, teacher, school, and LEA are updated instantaneously.

All aggregate reports in the reporting system can also be disaggregated by subgroups defined by USOE. For example, reports can be disaggregated by demographic characteristics like gender, ethnicity, or ELL status. The reporting system can also be configured to disaggregate by multiple characteristics at once (e.g., gender by ethnicity).

**PDF Copies of Reports for Stakeholders**

An individual student's report can be printed or saved as a PDF file. The PDF format makes electronic distribution easy and ensures that the report will print properly regardless of the user's browser. Users can batch print or save multiple students' reports in PDF format. The reporting tool can also PDF versions of students' written responses to performance tasks.

**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 amount of 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.

As mentioned previously, the adaptive tests also support benchmark reporting for groups of students. Besides providing an easy-to-understand report of a student's relative strengths and weaknesses by benchmark, this report also links directly to relevant formative assessment materials. By bridging the gap between interim assessments and formative assessments, the user can focus his or her interpretation of the results to a single student's capabilities and then use that information to make targeted decisions about instruction and intervention. See Exhibit 1.C.3.e-7 for interpreting and navigating the reporting system.

The reporting system comes with an intuitive user guide that provides step-by-step instructions for every report, information about each measure included in the reporting system, and screenshots 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.

**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; however, the user can select a previous year and view the same reports for the students that he or she had previously. As long as the user is authorized to use the reports, the user can access any of his or her reports from any test administration with the click of a button, even if the students from those years have changed schools or graduated.
Exhibit I.C.3.e-7: Navigating the Reporting System

Observe: My students are struggling in Grade 5 Math.
Ask: Which of my students are struggling in Grade 5 Math?

Investigate: Use the reporting system to see the performance of different groups in Grade 5 Math.

Observe: The students in one group are not performing as well as expected.
Ask: Are they struggling in particular areas of math or the subject as a whole?

Investigate: Use the reporting system to see the performance of the group in each Math standard and benchmark.

Observe: Some benchmarks are relative weaknesses for these students.
Ask: Which individual students need help in these areas?

Investigate: Assign benchmark-aligned activities to students; review their work and identify the ability of individual students.

You can use the formative assessment activities in Learning Point Navigator to assign relevant activities to a group of students and identify individual learning needs.
Similarly, users can access previous year’s scores for their new students. For example, a middle school teacher may begin teaching a class of students who have come from multiple elementary schools. Though the middle school teacher is not authorized to view the school reports for the elementary schools, the teacher can create a report based on the student’s previous test scores. This allows the teacher to assess the strengths and weaknesses of the class and of individual students based on their previous scores and adjust instruction accordingly.

**Test Management Center**

The Test Management Center allows users to track test participation in real time to help plan their testing schedules. The reports can also be filtered by student performance. For example, a user may want to plan testing for students who were not proficient on their first interim assessment. The user may decide that those students should not test yet because the school is planning additional instruction. High-level reports are also available so that school and district users can see the percentage of students who are completing each test.

**Student Search**

Sometimes a user just wants an individual student’s report and wants it quickly. The student search lets users search by student ID, student name, or any partial student ID or name. The search results provide a list of any students who match that criterion, as well as a table with links to the individual student reports for any tests that student has completed. As always, users can only see students that they have access to, though users can also search students from previous years. This allows a user to see an individual student’s results with a minimal number of mouse clicks.
I.C.4: Technology Requirements

I.C.4.a. Application Installation & Maintenance

AIR’s computer adaptive testing engine relies on a low-maintenance, low-bandwidth software model unique in the industry. USOE, districts, and individual schools will conduct virtually no maintenance on our secure system. Because the only software needed is a secure build of the Mozilla browser, AIR conducts maintenance at the server level and there is no need for schools to re-install software after an update. We support operating systems that are at least ten years old, and our system does not require extra hardware or complex software for local caching servers or proctor caching servers.

Below, we describe

- bandwidth requirements;
- testing, deployment to production, and maintenance of the production environment; and
- cutover and parallel processing.

This approach has served us well by shielding our tests from third-party software changes during the testing window. Each summer, we issue a secure browser update to incorporate important changes and deliver it for testing the following school year.

Overview

Secure Browser

Our system avoids dependance on third party software. Students can only access secure tests using AIR’s secure browser. The browser can be “branded” for USOE, with an icon and logo of your choosing. Non-secure tests, such as practice tests and some formative tests, can be accessed with ordinary browsers. 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. The secure browser enables keyboard combinations specifically designed for test navigation only. The system verifies that the test is being launched in the secure browser and prevents the test taker from continuing if the test is launched in a normal browser window.

Bandwidth Requirements

AIR’s system has been designed to have a “small footprint” within schools, meaning that it places minimal requirements on equipment, bandwidth, and expertise within the schools. Our system is designed to minimize bandwidth usage. In part, this is accomplished through measures taken on the front end to develop items in ways that respect the bandwidth limitations in the real world. For example, our Item Tracking System allows test developers to view different representations of graphics and select the lowest bandwidth representation that maintains the quality required by the particular graphic. We have taken steps to reduce the size of audio files by representing speech in a format that is optimized for voice. Our animators specialize in the optimization of the animations they create in either Flash or HTML-5.

The bandwidth demands are further reduced by our new real-time caching mechanism. This mechanism, introduced three years ago in and currently in use in four states, fetches the next item as a student is working on the current one. The caching is done within the secure browser on the student’s workstation (in memory, not on disk), so no additional hardware, software, or support is required.
Psychometric studies show that our prefetch mechanism does not perceptibly impact standard errors of measurement. Suppose a student is working on item number 19. The system notices that too few items remain cached (say, zero), and it begins the prefetch process. While the student works, the system is using responses to items 1–18 to obtain a current score estimate, a score that will be used in the selection of item 20. Item 20 is selected and delivered to the browser’s memory, or cached. We have conducted studies on the impact of our prefetch mechanism on psychometric properties of student scores. We investigated a range of configuration settings that launch the prefetch process when there are 0, 1, 2, 3, 4, or 5 items left in the cache. No impact is perceptible unless we set this parameter to at least 5, which means that the score used in the selection is 5 or 6 items outdated. In general, we recommend that prefetch initiation be set to 1 for tests with stand-alone items and 2 for tests with item groups (e.g., reading).

With this mechanism, even if a network slows down, the student rarely sees a delay after the test starts up. Currently in the field in operational states, after a student answers an item, the next item appears on the student’s screen in an average of less than one-quarter of 1 second. If a student takes approximately 30 seconds to answer a question, the next item has that long to load before the student sees any delay. Exhibit I.C.4.a-1 summarizes the required bandwidth per student with these very conservative assumptions—current performance in, and exceeds these expectations. Conservatively, even if 90% of the capacity on a T1 line is used for other purposes, a school could simultaneously test at least 50–100 students without apparent delay.

### Testing, Deployment to Production, and Maintenance of the Production Environment

Each release of every one of our systems goes through a complete testing cycle, including regression testing. We are pleased to share information about our testing processes with USOE upon request.

Each release, and every time we publish a test, the system goes through User Acceptance Testing (UAT). During UAT, we provide our client with login information to an identical (though smaller scale) testing environment to which the system has been deployed. We provide recommended test scenarios and constant support during the UAT period. For Utah, we propose to begin UAT a full 4 weeks prior to opening of the testing window. Issues identified within the first 10 calendar days of testing will be resolved and the fixes returned to production by the 14th calendar day of testing for final system check.

Deployments to the production environment all follow specific, approved deployment plans. Teams working together execute the deployment plan. Each step in the deployment plan is executed by one team member and verified by a second. Each deployment undergoes shakeout testing following the deployment.

<table>
<thead>
<tr>
<th>Percent of Bandwidth Utilized for Other Purposes</th>
<th>Amount of Bandwidth Available</th>
<th>Minimum Number of Students Supported Without Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 @90% capacity</td>
<td>125 kb/s</td>
<td>15–30</td>
</tr>
<tr>
<td>T1 @80% capacity</td>
<td>250 kb/s</td>
<td>30–60</td>
</tr>
<tr>
<td>T1 @60% capacity</td>
<td>500 kb/s</td>
<td>60–120</td>
</tr>
<tr>
<td>T1 @20% capacity</td>
<td>1,000 kb/s</td>
<td>120–240</td>
</tr>
</tbody>
</table>
This careful adherence to deployment procedures ensures that the operational system is identical to the system tested on the testing and staging servers. Upon completion of each deployment project, management approves the deployment log.

During the course of the year, some changes may be required to the production system. Outside of routine maintenance, no change is made to the production system without approval of the Production Control Board (PCB). The PCB includes the director of AIR’s Assessment Program or the chief operating officer, the director of our Computer and Statistical Sciences Center, and the project director, in this case. Any request for a change to the production system requires the signature of the system’s lead engineer. The PCB reviews risks, test plans, and test results. In the event that any proposed change will affect client functionality or pose risk to operation of a client system, the PCB ensures that the client is informed and in agreement with the decision.

- The PCB approves a maintenance plan that includes every scheduled change to the system.
- Deviations from the maintenance plan must be approved by the PCB, including server or driver patches that differ from those approved in the maintenance plan.
- Every bug fix, enhancement, data correction, or new feature must be presented with the results of a quality assurance plan and approved by the PCB.

The project director is responsible for representing USOE on the PCB and ensuring that you are informed.

We have an emergency procedure that allows rapid response in the event of a time-critical change needed to avert compromise of the system. Under those circumstances, any member of the PCB can authorize the senior engineer to make a change, with the PCB reviewing the change retroactively.

Typically, deployments happen during a maintenance window, and we schedule the deployments at a time that can accommodate full regression testing on the production machines. Any changes to the database or procedures that in any way might affect performance are typically subjected to a load test at this time.

**Cut-Over and Parallel Processing**

AIR maintains multiple environments to ensure smooth cut-over and parallel processing. With hosting at our site in Washington, DC, we maintain multiple development environments and a test environment. At Rackspace, our hosting provider, we maintain a staging environment and the production environment.

The production environment runs independently of the other environments and is only changed with the approval of our PCB. When we are developing enhancements, they are developed and tested initially on the development and test environments in DC, before being deployed to the staging environment in Rackspace.

The staging environment is a scaled-down version of the production environment. It is in this environment that UAT takes place. Only when UAT is complete and the PCB signs off is the production environment updated. In this way, the system continues to function uninterrupted as testing takes place in parallel until a clean cut-over takes place.
Hosting and Maintenance

Utah data and systems will reside on servers at Rackspace, our hosting provider. Rackspace has been identified by the Gartner group industry leaders in the Magic Quadrant of a two-dimensional chart ranking hosting providers on innovation and ability to execute. Over nearly a decade working with Rackspace, we have found the company to be both reliable and responsive.

Maintenance will be scheduled within agreed upon maintenance windows. These windows will occur between 2200 and 0600 MST/MDT or at other times as agreed by AIR and USOE. Should problems arise, we will document them, including risks and benefits of the proposed solutions. These communications are part of our PCB approval process described in detail above.

Availability Guarantees

This section outlines availability guarantees that will warrant that students, proctors and administrators will be able to complete their respective tasks at any time except the designated maintenance windows. AIR will negotiate particular service levels agreement, but our typical guarantees in similar statewide testing projects include these:

- System always up and running during designated hours, except due to circumstances beyond our reasonable control (e.g., war, weather emergency, etc.)
- Guarantees that any issues that arise and affect testing will be ameliorated within one hour
- Rapid response to help desk inquiries with transparent reporting on all status and resolution inquiries
- Less than one second average wait time between items or item groups
- Complete and accurate display of every item delivered
- Uncompromised security and confidentiality of student data

AIR will be pleased to negotiate the particular metrics around these or other service level agreements during contract negotiations.

Method of Deployment

As described above, the only software required at the schools or districts is the secure browser. The secure browser can be installed by using any of the network management tools described in the proposal. We will provide appropriate packages for distribution through network distribution tools, as well as for installation on individual machines.

At Utah’s discretion, we can also provide a version of the browser that can be installed on single machines that does not require any administrative permissions. We have found that some smaller schools (often charter schools) lack sophisticated network distribution systems, and have limited access to their own machines.

When installation packages are employed (excluding the “administrative permission free” version) they can be configured to eliminate older versions of the browser without a separate action.

When uninstalled, our secure browser uninstalls cleanly, leaving no trace.

System to Monitor Operation

AIR’s test delivery system provides a broad range of tools for monitoring testing activity as well as system health. State users and district and school test coordinators have access to
Participation Reports. Participation Reports provide real-time information about the testing status of students within the user’s jurisdiction. A school test coordinator, for example, would have access to information about students in his or her school, while the district test coordinator would have information spanning the entire district. The reports provide information about current testing activity, as well as providing a powerful tool for monitoring progress toward getting all students tested. Users can filter or sort the data to identify students who have (or have not) completed any opportunities, students who have (or have not) started a testing opportunity, students whose unfinished tests are at risk of expiring if they do not complete them, and a host of other useful administrative information. Exhibit I.C.4.a-2 highlights some of the features of the Participation Reports.

State users also have access to our “latency” reports, which include information about how long various parts of the system take to deliver items. For example, it reports the average time lapsed between the server receiving an item response and (1) the database to responding to server requests; (2) the overall system responding with the next item; and (3) the time it takes between sending the next item and the client machine receiving it. These measures tell instantly and accurately how the various system components are performing. We also capture client latencies, which indicate the time that students wait to see the next item. The client latency is the elapsed time between completing one item and the next item fully rendering on the client screen. Client latencies benefit from the prefetch mechanism, so students see a very responsive system.

Latency statistics are available at each level from the state through the school. Because latency reports put a substantial demand on the database system, they are not available in real time. Instead, they are calculated each night and made available for the prior day each morning.

As described above, our Quality Monitor (QM) as well as our server monitoring software generate alerts that go to our network engineers. Alerts that are likely to have consequences for the field can be automatically emailed to USOE, though we recommend that clients rely on our project teams to inform them. Our monitors are quite sensitive and generate many “early warning” alerts that turn out to reflect no problems in the field.
I.C.4.b. Security and Integrity

The system adheres to strict, industry-standard security procedures, and each element of the system is redundant, minimizing the risk of failure. We discuss each of these below.

System Security

All our systems meet strict standards for security of their data and the systems themselves. Access to AIR production servers is strictly limited to a few network and senior software engineers. Each engineer, and his or her level of access, must be approved by our Production Control Board (PCB), as described above. Access to the servers requires at least a 10-character password incorporating three of the following four character types: mixed case, alphanumeric, numerals and symbols. After access rights have been granted, any changes on the production servers must be approved by the PCB.

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

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

Below, we describe key security procedures that will protect Utah’s items, ensure confidentiality and privacy, and enforce Utah’s public records laws, FERPA, and other federal laws.

Physical Security

Utah 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 keycard 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 are keycard 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 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, which 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 128bit 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 Utah’s public records laws, FERPA, 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 Family Educational Rights and Privacy Act (FERPA) differently, and we support customized interpretations. Our system is designed to support these interpretations flexibly. Some states limit a school’s access to data collected while the student attends the school, limiting access to historical data on students who transfer into the school. Other states provide the full history of data to the school or teacher who has jurisdiction over the student at any point in time. Similarly, while some states provide each teacher with access to information about all students in the teacher’s school, other states limit access to those students taught by the teacher. Our systems can be configured to support all these scenarios and more.

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 that 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. For example, one client allows district personnel to view only test results administered in their district, while others do not impose this restriction. We will work with USOE to establish the business rules governing data access in the system.
In ITS, users' rights to data are governed only by project and role. Users with different roles have differential ability to perform different types of reviews, view different reports, or invoke different administrative functions.

**Security Audits**

AIR conducts periodic security audits, including an electronic security audit of every new system deployed. Our chief security officer conducts the security audit, and security staff pose challenges to our software systems to emulate security threats. Physical security is periodically audited by our Director of Program Management DeeAnn Wagner.

**Robustness**

We have designed our TDS 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 follow:

- 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 max 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 internal redundancy, 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. The systems are run with full transaction logging, enabling us to restore the system to its state immediately prior to a catastrophic event.

All backup media are stored in a secure, fireproof, offsite location. Our hosting provider ensures that all media sent offsite are shipped in locked, water-resistant, and impact-resistant containers. The offsite vendor does not have direct access to individual media containing customer data at any point or any time during transport. The offsite storage location is audited on a regular basis to ensure proper physical security, media management, and location tracking are meeting the hosting provider’s industry standard guidelines.

**I.C.4.c. Supported Devices and Operating Systems**

Not only does AIR’s system (specifically the Student Interface) currently support all the devices and operating systems mentioned in the RFP, it also supports operating systems that are more than 12 years old (e.g., Windows 2000). AIR is committed to having a small technology footprint at the districts and schools; thus, part of this footprint is to be
### Exhibit I.C.4.c-1: Operating Systems

<table>
<thead>
<tr>
<th>Supported Operating System on AIR’s Student Interface</th>
<th>Operating System Version</th>
<th>AIR Capability and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows</td>
<td>2000 XP, Vista, 7, 8 Windows Server 2003 and 2008</td>
<td>AIR supports all Windows operating systems in the RFP and more.</td>
</tr>
<tr>
<td>Mac OS X</td>
<td>10.4–10.8</td>
<td>AIR supports all Mac operating systems in the RFP.</td>
</tr>
<tr>
<td>Linux</td>
<td>Fedora 6 (K12LTSP 4.2+) Ubuntu 9–12 SUSE Linux Enterprise Desktop 11 with Gnome</td>
<td>AIR supports all Linux operating systems in the RFP and more. We are still testing the SUSE configuration but commit to supporting it.</td>
</tr>
<tr>
<td>Android</td>
<td>4.0</td>
<td>AIR supports Android 4.0.</td>
</tr>
<tr>
<td>iOS</td>
<td>5.0, 6</td>
<td>AIR supports iOS 5.0 and 6. We do not recommend allowing iOS 5 for secure tests due to security limitations. We are currently finishing security testing on iOS 6, but it appears to be securable.</td>
</tr>
</tbody>
</table>

backward compatible as well as forward-looking. AIR realizes that many districts and schools have old systems but many have new systems too. Thus, the exhibits illustrate support for a 12-year-old operating system such as Windows 2000 on one end and a 2-month-old operating system such as Mac 10.8 on the other. Exhibit I.C.4.c-1 summarizes supported operating systems. Exhibit I.C.4.c-2 summarizes supported devices.

#### Continuous Upgrades

AIR recognizes that as updates are made to operating systems, our secure browsers need to be upgraded as well. AIR will work with USOE on an upgrade policy that, as much as possible, minimizes disruption and setup on the districts and schools in the middle of the school year. Thus, AIR usually upgrades its secure browser at the beginning of every school year.

In line with this upgrade, AIR conducts extensive platform testing of test content on several operating systems to ensure that test content, including graphics and videos, is rendered in the same way across all machines. We also ensure that all background colors and invert colors render test content meaningfully across all operating systems. In addition, AIR tests its secure browser on all supported operating systems and devices and ensures that both functionality and security are in no way compromised.

#### I.C.4.d Hosting

AIR will install a system capable of handling 30,000 concurrent users, more than 1.2 million summative tests, and more than 500,000 students participating in three or more tests. If more tests than this will be administered, AIR can do so free of charge up to the capacity of our designated Utah servers.
## Exhibit I.C.4.c-2: Devices

<table>
<thead>
<tr>
<th>Device Supported by AIR’s Student Interface</th>
<th>AIR Capability and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standalone desktops and laptops that are less than five years old</td>
<td>AIR supports these desktops and laptops, assuming they use operating systems listed in Exhibit I.C.4.c-1.</td>
</tr>
<tr>
<td>Netbooks</td>
<td>AIR supports netbooks. However, we recommend a screen size of 11.6 inches to enable the student to have a seamless testing experience with the test and item layouts.</td>
</tr>
<tr>
<td>Standalone desktops and laptops that are more than five years old</td>
<td>AIR supports these desktops and laptops, assuming they use operating systems listed in Exhibit I.C.4.c-1.</td>
</tr>
<tr>
<td>Multiplied workstations</td>
<td>NComputing is supported on computers running Windows XP, 7, and 2008.</td>
</tr>
<tr>
<td>Thin client/hosted workstations</td>
<td>Citrix, VMWare, Terminal Services, and Microsoft RDC are supported on Windows 2003 and 2008 Servers. However, AIR does not recommend using VMWare for two reasons: (a) there is a security issue because students can get out of a session undetected by the Proctor interface, and (b) there is an issue with sound transferring from a &quot;virtual sound card&quot; to a &quot;real sound card&quot; on the local machine, so that any test content with a sound or speaking component may be unreliable.</td>
</tr>
<tr>
<td>Tablets</td>
<td>AIR supports iPad 2 and iPad3, as well as Motorola Xoom and other Android tablets, assuming they use operating systems listed in Exhibit I.C.4.c-1.</td>
</tr>
<tr>
<td>Small form tablets</td>
<td>AIR supports small form tablets using operating systems mentioned in Exhibit I.C.4.c-1. However, AIR does not recommend using these small devices. Due to the very small screens on these tablets, students cannot effectively test on them without being distracted by the constant need to zoom/un-zoom and scroll.</td>
</tr>
</tbody>
</table>

We commit to a mean server response time well below one second. (Typically, our server response times are more than 10 times faster than that.) Mean refresh time is affected by many factors, including the simultaneous demands on the school network, Internet transmission anomalies, student workstation configurations, and other factors outside our reasonable control. However, we will collaborate with USOE and the schools to ensure the smoothest possible testing experience for all students.

To meet this RFP requirement, our plan has three aspects: (a) setting up the hardware and its configurations, (b) making the system extremely fault tolerant, and (c) monitoring system performance proactively and accurately.
Hardware Setup

Our plan is to set up a dedicated server cluster for AIR’s online testing system consisting of at least two web servers and an active/passive database cluster. If a need arises for more web servers in the spring, when testing is usually heaviest, AIR can easily add these servers to the cluster with minimal disruption.

The servers will also be configured in a way that distributes data loads across multiple file groups and separates heavy transactions through efficient resource allocation.

Fault Tolerance

We have designed our system to be extremely fault tolerant. It 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 include these:

- Our hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts in place, the generators 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 a 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. In the event of a failure of any individual disk, recovery is immediate by accessing the redundant data on another disk.

Monitoring System Performance

AIR’s system is distinguished by our ability to continuously monitor performance, both at the servers and at every student’s machine. We aggressively analyze system performance at every layer of the system and continuously design improvements. Our systems enable us to identify schools that are experiencing latency problems, work directly with those schools to identify the sources of these problems, and help the schools resolve them.

Our system captures a wide range of latency data to provide early warning of any potential slowdowns and identify schools or regions that may be experiencing Internet or other delays. Our system captures data on the amount of time that passes after a student answers one item before the next is fully loaded on the examinee’s screen. We also log a host of interim times, such as the time taken by any database requests and the time until a complete request is addressed by the servers.

The latency data (the time from when a user submits an item until the next item appears fully on the screen) form the basis for a set of latency reports that our engineers review frequently. We can use these tools to identify schools that may be experiencing problems
even before they decide to call the help desk. We also have a school latency report that identifies schools that may be experiencing slow response times. These reports can be used to identify schools that may be in need of extra support from our network engineers or schools that may be experiencing problems with their Internet provider.

**Load Testing**

To ensure the fastest possible client-side refresh times, AIR commits to extensive load testing on our actual production machines (to be conducted at night, during maintenance periods, and when the window is closed) at least 90 days before the first operational assessment is delivered to a student on the system.

To get reliable results, our load tests will be run with exactly the same system and network configurations as our production servers and will be located within the same firewall as our production servers.

Our load-test drivers enable us to simulate large loads. Our approach is to test system components with the stress and demand generated by 50,000 simulators users. We will report both server-side and client-side latencies. Results from these load tests will be shared promptly with USOE, and any issues will be resolved. We are confident that USOE will be quite pleased with our load-test results.

Throughout the school year, any changes to the database procedures that may affect performance in any way are always subjected to the type of load test described above.
I.D: Psychometrics

Overview

When done well, computer-based testing is not simply an alternate mode of test administration. As we’ve detailed in Section II.B.2 on item development, AIR’s suite of simulation and graphic response items provides test developers tremendous flexibility to design items that measure student knowledge deeply and in ways that engage student attention to maximize their achievement potential.

Similarly, designing an online adaptive assessment system, when done well, involves more than simply maximizing test information near each student’s level of ability. The central notion of criterion-referenced testing is that test performance directly measures the knowledge and skills that students are expected to achieve by the end of each grade or course. Moreover, these criteria are multifaceted, including not only academic content standards, strands, and benchmarks but also students’ ability to engage with academic content using more complex levels of cognitive processing and to demonstrate knowledge and skills across a variety of response modalities, including writing as well as other forms of symbolic representation, such as equation writing, manipulating variables to achieve different outcomes, and plotting and graphing data.

Regardless of their overall ability level, all students should be provided the opportunity to demonstrate higher-order thinking skills as part of their assessment of academic knowledge and skills. Being able to analyze, reason, and synthesize information is not proprietary to high-ability students. Test questions can and should be designed to measure greater depth of knowledge across the ability spectrum, and all students, regardless of ability, should be administered items that measure those higher-order thinking skills. AIR’s adaptive algorithm enacts complex blueprints that meet test specifications not only for academic content standards but for test specifications that reach across standards, including depth of knowledge, item types, and passage genre, among other such test specifications.

And beyond maximizing test information near each student’s overall ability level, test results should provide more fine-grained feedback about student performance that educators can use to diagnose student strengths and weaknesses to yield more effective instructional strategies for students. To increase the diagnostic utility of strand-level reporting, AIR’s adaptive algorithm also maximizes the likelihood that students’ performance on content strands is above or below the performance standard to better enable educators to identify students with particular strengths or weaknesses within each content strand.

Thus, AIR’s adaptive algorithm is designed to configure student assessments to achieve a balance among multiple objectives.

Balancing Multiple Objectives

Although widespread adoption of computer-adaptive testing (CAT) has importantly benefited from recent advances in computer availability and increased Internet bandwidth, perhaps the critical advance in adaptive testing has been the development of item selection algorithms that allow for adaptation in item selection within the context of criterion-referenced tests. In criterion-referenced tests (CRTs), the content of test administrations matters. Test administration strategies that simply maximize test information near the examinee’s ability will typically not administer a representative sample of the content domain, especially as formalized in the test specifications. What is needed, and what AIR’s
adaptive algorithm provides, is an item selection engine that adapts to student ability within the constraint of a criterion-referenced test blueprint.

In a fixed-form test, the ideal test information function is decided once, as is the blueprint. 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 depend, often, 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.

The implications of the decisions differ between fixed-form and adaptive tests as well. When a fixed-form test misses blueprint by one item, 100% of students are administered a test that does not quite meet blueprint. In an adaptive situation, a test that misses blueprint is more likely to miss it by a small number of items (one or two) for only a portion of the tested students. Extending the ability of the test to adapt has an impact on the precision of scores for higher- and lower-achieving students, the groups that are generally compromised in a fixed-form test.

AIR alone offers sophisticated tools for optimizing algorithm and blueprint parameters to best meet competing objectives. The act of balancing these objectives to establish an optimal test is complex because the objectives interact and they all interact with the available item pool. Most often there are one or more optimal configurations, but finding them can be challenging. Our psychometricians have access to our simulation tool, which enables them to manipulate key blueprint and configuration settings by making changes in a spreadsheet. The simulator then simulates a large number (determined by the psychometricians) of examinees taking the test and reports back many measures indicating performance relative to each objective. The feedback is sufficiently detailed to enable our psychometricians to react with further configuration changes until an optimal mix is reached.

**Item Bank Calibration and Equating**

A derivative advantage of our online adaptive assessment system is that it provides for the possibility of real-time reporting of assessment results for students, classrooms, schools, and districts. In fact, online administration of state assessments almost demands immediate reporting, because consumers of assessment results have come to expect immediate feedback when interacting with online systems. Taking advantage of real-time reporting requires a pre-equating approach to item calibration and equating, which we propose for Utah’s online adaptive assessment system.

Establishing a pre-equated item bank to support adaptive test administration and real-time scoring and reporting also requires balancing many important considerations. Ideally, the
very first operational administration of the online assessment system would be adaptive and would support immediate reporting of test results, which presumes adoption of performance standards for the new assessment. A common approach is to administer an independent or stand-alone field test designed to support calibration and equating of bank items, and to conduct standard-setting workshops based on the field-test administration. While functional, because this approach relies on field-test data, motivational characteristics contributing to student performance in operational test administrations may not be captured in item calibrations or student ability estimates used to help identify performance standards. Nevertheless, this approach provides a seamless and immediate transition to a fully functional online adaptive assessment system with real-time reporting of test results.

An alternative approach is to calibrate items and set performance standards based on an initial operational administration. In this approach, the first operational test is administered online, but not adaptively. Item parameters are calibrated and equated and performance standards established, following the first operational test administration, and scoring and reporting of test results are necessarily delayed. However, because item parameter estimates are based on operational test conditions, they are likely to be more reliable and therefore less likely to drift across subsequent test administrations although other factors, including item exposure and targeted instruction, may also lead to item parameter drift. This approach has typically been implemented using a fixed form, incomplete block design, just like the approach used in paper-based tests. While this approach is adequate and does provide a linking design allowing concurrent calibration of test items, the linking design is artificially limited by a constraint that does not exist in the online world. The incomplete block designs employed in paper-based tests derive from the expense of publishing and printing a variety of test forms. In an electronic world, however, such constraints need not apply.

AIR has developed an item selection engine designed explicitly to calibrate adaptive item banks under operational test conditions. This system is designed to accomplish two important outcomes. First, because it is an operational administration of a criterion-referenced test, students are administered test forms that meet blueprint specifications. Second, to ensure that items are administered to representative samples of Utah students and that all items are administered with all other items in a fully crossed linking design, this item selection engine preferentially weights least administered items for selection from among all items that satisfy blueprint requirements. This algorithm ensures that items are “spiraled” across test administrations statewide, yielding samples for each item that are more nearly representative of the state as whole, including across important demographic subgroups, facilitating analysis of differential item functioning. Importantly, because all items are linked to all other items, parameter estimates resulting from concurrent calibration of all bank items are more reliable than can be obtained with concurrent calibration of common item linking designs that utilize only a sample of items in the bank.

I.D.1. Overall Approach

To maximize the diagnostic utility of the interim assessments and to better prepare students to perform on the summative assessments, we propose to deploy the same adaptive blueprint to administer student assessments for both the interim and the summative assessments. This means that students taking the interim assessments will be administered tests that have not only the same distribution of items across the CCSS and Utah Core Standards but also the
same distribution of items across DoK levels and item types, including machine-scorable constructed-response items and performance tasks. Below, we propose a plan to place all interim and summative items on a common scale. With a common blueprint and items calibrated on a common scale, test scores between interim and summative assessments will be more than predictive, they will be equivalent.

To better maintain the security of the summative test items, once the item banks have been calibrated and equated, Utah may choose to divide the items into separate interim and summative pools. While some states do provide students and schools multiple opportunities to participate in the accountability assessments and therefore maintain only a single item pool to administer tests across multiple occasions, because the single-opportunity summative assessment is the only test administration that satisfies the state’s accountability requirements, there is an advantage to limiting exposure of the items in the summative assessment. Splitting the item banks into interim and summative pools also suggests a strategy for managing items over time. For the initial split of items into interim and summative pools, we propose the following criteria:

- **Divide items into roughly similar-sized item pools.** Although the summative assessment is only one of three opportunities (two interim and one summative), as the high-stakes assessment, the summative item pool should be more robust, providing greater flexibility to meet blueprint while adapting to student ability.

- **Maintain distribution of item attributes across pools.** Item pools should have the same distribution of items across content strands, standards, and benchmarks, as well as non-hierarchical elements such as DoK and item type. The range of item difficulties across pools should also be main-

- **Assign most discriminating items to accountability pool.** The most discriminating items, items with large IRT parameters or with the highest biserial correlations, should be preferentially assigned to the accountability pool.

Once the initial item bank has been established, all embedded field-test items will be automatically deposited into the summative item bank following field-test analysis and subsequent item data review.

Following the summative test administration window each year, items in the summative pool will be identified for transfer to the interim pool. As described in Section II.B.2, AIR performs test integrity analyses throughout the test administration window, including an examination of item fit and item exposure. As part of those analyses, AIR will flag items in the summative pool that show substantial misfit, as well as heavily used items that are at risk for overexposure. These flagged items will be reviewed for transfer from the summative to the interim pools. Items showing unacceptable levels of misfit may be recalibrated prior to transfer to the interim pool. USOE may also elect to maintain some highly exposed items in the summative pool if they measure underrepresented content standards or benchmarks.

The same test integrity analyses applied to the summative assessments will be used to monitor the quality of test administrations in the interim assessments. Even though the interim assessments are not attached to high stakes, fidelity of the interim test scores is necessary to ensure the diagnostic and predictive utility of the assessments. Flagged items in the interim pool will be reviewed for possible release.
Released items will still have a productive life as part of Learning Point Navigator, our formative assessment and learning tool.

**Field-Testing**

While the interim and summative assessments may constitute separate item pools, the diagnostic utility of the interim assessments depends on the equivalence of test scores between the two systems. Therefore, we propose to establish item pools for both assessments as part of a single independent or stand-alone field test. AIR proposes to launch Utah’s statewide online adaptive assessment system by fall 2014. To accomplish this, we propose to establish bank parameter estimates in spring 2014 and conduct standard-setting workshops in summer 2014.

**Option A: Operational Field Test**

As described above, AIR has developed an item selection engine uniquely capable of supporting item bank calibration and equating while providing individual test administrations that meet test blueprints. In this approach, AIR would import the Utah item banks, configure test blueprints for the new Utah statewide assessment system, and provide operational test administrations of the CRT statewide in the spring 2014 assessment window. Although this schedule is aggressive, because Utah already provides a computer-based CRT, test items are electronically available to import into AIR’s Item Tracking System. Item development staff will work with USOE to ensure that items render as intended, as well as to complete any necessary work to ensure alignment of test items to the CCSS and the Utah Core Standards. Following operational test administration, student responses to test items would be calibrated and equated using the 3PL IRT model described in Section II.B.2, and final reporting scales developed and implemented in time for a fall 2014 deployment of the online adaptive interim assessment system.

Option A represents Utah’s best option for the assessment. It allows Utah to retire the current CRT after the spring 2013 administration. In 2014 students will experience a test that automatically selects items for each student to match the blueprint, as the adaptive algorithm will in 2015. During the spring 2014 administration the second consideration, after blueprint match, will be quality of the statistical sample for calibration. During this first administration students will not receive immediate feedback because calibration and scoring will have to wait until the end of the testing window. Over the summer of 2014 we will hold standard setting, and by fall 2014 the interim assessment will be able to go live, with a fully calibrated pool, to be followed by the summative assessment on any date USOE chooses during the 2014–2015 school year.

Our proposed price reflects Option A. If Utah would like to pursue Option B, we are happy to negotiate the details and provide a price.

**Option B: Independent Field Test**

Alternatively, AIR is prepared to administer an independent or stand-alone field test to obtain student response data from which to calibrate item parameters and recommend performance standards. AIR’s standard field-test engine selects items randomly from among the least administered items in the pool. This procedure ensures that items are spiraled across students statewide, reducing sampling inefficiencies due to clustering in schools and classrooms and yielding randomly equivalent samples across all test items. As described below, AIR will design a sampling plan to ensure that a representative sample of schools and students par-
participate in the field-test administration. Moreover, because independent field-test administrations do not need to meet blueprint and because participation in an independent field test is likely to be voluntary, we can minimize the burden on participating classrooms and schools by reducing significantly the number of items administered to each student, ensuring that field tests can be administered within a single class period.

Our price reflects the operational field test and assumes spring 2013 will be the last administration of the current CRT.

**Independent Field-Test Sample**

All independent field tests for English language arts, mathematics, and science will be administered in spring 2014, so that students will have had the opportunity to learn all course objectives identified in the CCSS and Utah Core State Standards. Because course-based assessments (e.g., Biology) are designed to measure the skills and content knowledge at the end of the course instruction, it is imperative that the field test be administered to the students who have had the opportunity to benefit completely from instruction in each course in order to construct an item pool with robust item statistics that reflect a full range of student ability at the end of the instruction.

**Number of Independent Field-Test Items**

Consistent with the RFP, the distribution of items in the field-test pool will be proportional to the specifications in the test blueprints for content domains, content clusters, and standards, as well as item types and depth of knowledge, and representing a wide range of item difficulties. It should be noted, however, that this is not a requirement for AIR’s adaptive algorithm, which can enforce complex blueprints while adapting to student achievement, even if the item pools are not proportional to the blueprint. Only items that are approved by the Content and Fairness Item Review Committees will be included in the independent field-test pool.

To support three test administrations (two interim opportunities and one summative opportunity), we would ideally like to target for independent field-test administration about nine times as many items as specified in the test administration blueprint. Based on the proposed blueprints presented in Appendix A, Exhibit 1.D.1-1 presents approximate numbers of field-test items for each independent field test.

While USOE has relatively large item banks for some subject areas and item types, other subject area item banks, especially ELA, are not sufficient to support both interim and summative adaptive assessments, and the current pools do not currently include the kinds of machine-scorable constructed-response (MSCR) items that AIR envisions for Utah’s statewide assessment system. To bolster Utah’s item pools, especially with respect to MSCR item types that take full advantage of available technology, AIR has obtained commitments from and to participate in item sharing with Utah. Because the item pools are already aligned to the Common Core State Standards, Utah can immediately field-test and implement online adaptive assessment with a full range of machine-scorable constructed-response items. With the initial item pool in place, AIR will work with Utah to carefully develop and implement high-quality machine-scorable items.

**Field-Test Length**

To yield accurate and reliable parameter estimates in the item pool, it is important that parameter estimates be obtained under test conditions that closely approximate opera-
Exhibit I.D.1-1: Number of Items in the Independent Field Test (option B)

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MC</th>
<th>TE</th>
<th>SIM</th>
<th>RT</th>
<th>Total</th>
<th>Target Item Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English Language Arts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G4 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G5 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G6 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G7 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G8 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G9 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G10 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td>G11 ELA</td>
<td>36</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>405</td>
</tr>
<tr>
<td><strong>Mathematics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3 M</td>
<td>32</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>42</td>
<td>378</td>
</tr>
<tr>
<td>G4 M</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>G5 M</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>G6 M</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>G7 M</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>Pre-Algebra</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>Number &amp; Quantity</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>Algebra I</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>Functions</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>Modeling</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>Geometry</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>Statistics &amp; Probability</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td><strong>Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4 S</td>
<td>38</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>49</td>
<td>441</td>
</tr>
<tr>
<td>G5 S</td>
<td>38</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>49</td>
<td>441</td>
</tr>
<tr>
<td>G6 S</td>
<td>43</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>54</td>
<td>486</td>
</tr>
<tr>
<td>G7 S</td>
<td>38</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>49</td>
<td>441</td>
</tr>
<tr>
<td>G8 S</td>
<td>36</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>423</td>
</tr>
<tr>
<td>ESS</td>
<td>43</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>54</td>
<td>486</td>
</tr>
<tr>
<td>Biology</td>
<td>38</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>49</td>
<td>441</td>
</tr>
<tr>
<td>Chemistry</td>
<td>43</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>54</td>
<td>486</td>
</tr>
<tr>
<td>Physics</td>
<td>38</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>49</td>
<td>441</td>
</tr>
</tbody>
</table>

Note:
MC: Multiple-choice items, hot-spot items, word builder items
TE: Technology-enhanced items, including graphic response items, natural language items, equation editor
SIM: Simulations
RE: Research tasks
Exhibit I.D.1-2: Option B Field-Test Length and Sample Size per Item

<table>
<thead>
<tr>
<th>IFT Items per Student</th>
<th>Total Items</th>
<th>Likelihood of Item Appearing in Test Admin</th>
<th>Target N per Item</th>
<th>Target IFT Sample N</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>450</td>
<td>0.062</td>
<td>1000</td>
<td>18000</td>
</tr>
<tr>
<td>25</td>
<td>450</td>
<td>0.062</td>
<td>1500</td>
<td>27000</td>
</tr>
<tr>
<td>25</td>
<td>450</td>
<td>0.062</td>
<td>2000</td>
<td>36000</td>
</tr>
<tr>
<td>30</td>
<td>450</td>
<td>0.074</td>
<td>1000</td>
<td>15000</td>
</tr>
<tr>
<td>30</td>
<td>450</td>
<td>0.074</td>
<td>1500</td>
<td>22500</td>
</tr>
<tr>
<td>30</td>
<td>450</td>
<td>0.074</td>
<td>2000</td>
<td>30000</td>
</tr>
</tbody>
</table>

tional testing conditions. However, because the field-test participation is voluntary, we propose to shorten the field-test length to reduce the testing burden, maximize students’ motivation, and ease the scheduling of the computer labs.

We propose a field-test length of 25 items (one class period) for all grade and subject area assessments, yielding sufficient sample sizes for the item calibration and the differential item functioning (DIF) analyses. Exhibit I.D.1-2 provides the target sample size given the item pool size, the number of items administered to each student, and the target sample size per item. The table shows options for a 25- and 30-item field-test length, as well as sample sizes of 1000, 1500, and 2000 student responses per item. A test length of 25 items administered to fewer than half the students in the state can yield sample sizes of 1000 students per item for a pool of 450 items, which is sufficient to yield reliable item parameters and detect DIF between important demographic subgroups.

**Option B Field-Test Item-Selection Algorithm**

The algorithm employed by AIR’s field-test engine ensures that:

- efficient samples are used for all items by randomly selecting from among the least administered items to ensure that resulting DIF statistics are maximally efficient across all items;
- position effects are averaged out by randomly administering test items across test positions; and
- more robust linkages exist among items, because each item is linked with every other item across hundreds of unique test “forms” to a degree not manageable through a set of fixed forms.

For pragmatic considerations related to system performance, the field-test algorithm limits the item selection to test start-up, instead of selecting each item in real time as with the adaptive algorithm. Therefore, the field-test engine assigns all items at the beginning of each test administration.

Upon test start-up, the algorithm selects a series of items in the following iterative sequence:

1. Identify all the items that were least frequently administered to the stratum.
2. Randomly select an item from among that group with equal probability.
3. Return to step 1 and continue if the requisite number of items has not been met.
The first step initiates a random sequence. Note that when the first student tests, all items may be selected with equal probability. Subsequent selections depend on this first selection because items are sampled without replacement until the entire pool has been administered, at which point the whole set becomes eligible for sampling again. This dependence is analogous to the dependence that occurs when test booklets are spiraled within classrooms. The probability of a student being assigned a particular booklet depends on the availability of booklets left to be handed out.

The probability of administering an item from the pool is constant across individuals in the population. For a single grade and subject, let $n$ represent the number of students, $k$ represent the number of items on the test, and $m$ represent the total number of items in a pool. The expected number of times that any single item $j$ will be administered within the population can be calculated by

$$\frac{n k}{m} = n_j.$$

The corresponding probability that a given student $i$ will receive item $j$ is therefore

$$\frac{n_j}{n} = \frac{k}{m},$$

which does not depend on $n$, the number of students in the sample, and is constant across the population.

From this we see that

- a random sample of students is administered each item; and
- for any given item, the students are sampled with equal probability.

In summary, the field-test algorithm yields a representative, randomized sample of student responses for each item. The field-test algorithm also leads to randomization of item position and the context in which items appear. Field-testing each item in many positions and contexts should render the resulting statistics more robust to these factors.

For the items with stimuli, such as items associated with a simulation and research/essay tasks associated with some reading passages, if an item that is part of a shared stimulus group is selected, then all items in that group will be administered.

**Embedded Field-Test Items**

After the analysis of independent or operational field test data, we will perform a detailed operational item pool analysis, identify any gaps in the item pool, and develop new items to fill the gaps as well as to augment the item pool for the future Utah test administrations. The new items will be embedded in each operational test administration. The embedded field-test (EFT) items do not contribute to the student’s test score and are not administered adaptively. Like the independent field test, the electronic test delivery system will employ a field-test engine to administer a predetermined number of nonoperational test items.

Placement of items within the test administration may be constrained (e.g., no EFT items administered as part of the first 5 or last 5 items in a test administration), but within those constraints, items will be administered to students based on a random selection scheme. This ensures efficient samples for all items because it minimizes effects of clustering in a manner similar to spiraling test booklets. This approach also ensures that position effects are averaged out because items are randomly administered across test positions.

Field-test items will be embedded in all operational administrations to keep the same test length. Because the same field-test item pool
### Exhibit I.D.1-3: Embedded Field Test and Sample Size per Item

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MC</th>
<th>TE</th>
<th>SIM</th>
<th>RT</th>
<th>Total</th>
<th>EFT Items per Student</th>
<th>Likelihood of Item Appearing in Test Admin</th>
<th>N per Item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English Language Arts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td>G4 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td>G5 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td>G6 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td>G7 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td>G8 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td>G9 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td>G10 ELA</td>
<td>54</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>5</td>
<td>0.074</td>
<td>2941</td>
</tr>
<tr>
<td><strong>Mathematics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3 M</td>
<td>48</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>63</td>
<td>5</td>
<td>0.079</td>
<td>3175</td>
</tr>
<tr>
<td>G4 M</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>G5 M</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>G6 M</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>G7 M</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>Pre-Algebra</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>Number &amp; Quantity</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>Algebra I</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>Functions</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>Modeling</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>Geometry</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>Statistics &amp; Probability</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td><strong>Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4 S</td>
<td>57</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>74</td>
<td>5</td>
<td>0.068</td>
<td>2703</td>
</tr>
<tr>
<td>G5 S</td>
<td>57</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>74</td>
<td>5</td>
<td>0.068</td>
<td>2703</td>
</tr>
<tr>
<td>G6 S</td>
<td>65</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>81</td>
<td>5</td>
<td>0.062</td>
<td>2469</td>
</tr>
<tr>
<td>G7 S</td>
<td>57</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>74</td>
<td>5</td>
<td>0.068</td>
<td>2703</td>
</tr>
<tr>
<td>G8 S</td>
<td>54</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>71</td>
<td>5</td>
<td>0.070</td>
<td>2817</td>
</tr>
<tr>
<td>ESS</td>
<td>65</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>81</td>
<td>5</td>
<td>0.062</td>
<td>2469</td>
</tr>
<tr>
<td>Biology</td>
<td>57</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>74</td>
<td>5</td>
<td>0.068</td>
<td>2703</td>
</tr>
<tr>
<td>Chemistry</td>
<td>65</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>81</td>
<td>5</td>
<td>0.062</td>
<td>2469</td>
</tr>
<tr>
<td>Physics</td>
<td>57</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>74</td>
<td>5</td>
<td>0.068</td>
<td>2703</td>
</tr>
</tbody>
</table>
will be embedded in each administration, all EFT items need to be developed and field-test-ready before the testing window opens. We propose to increase the item pool about 50 items embedding 5 field-test items in each test. The 50 items will be developed to replenish any shortages in the item pool and to increase the item pool size. The EFT item pool size may increase if any additional items are available from other secure sources including, possibly, Smarter Balanced item banks or other state item pools that can be shared with Utah. Exhibit I.D.1-3 presents the number of EFT items and the sample size per item for each statewide assessment. Given Utah’s student population, with just five EFT slots per test administration, more than 2,800 student responses will be obtained for each of the 70 items, so additional items can be field-tested even without adding to the EFT set size.

**Item Analyses**

After the testing window closes, all new machine-scored constructed-response items will go through a rubric validation process that is analogous to rangefinding for human-scored items to refine the machine-scored rubrics. For any responses scored using the essay scoring engine, all responses will first be human-scored. Then a subset of human-scored responses will be used to develop the autoscoring algorithm (e.g., identify the regression model that best predicts the human-assigned scores). The remaining human-scored responses will be used as a cross-validation sample to determine whether the “trained” autoscore algorithm successfully predicts human scores in a sample of responses that were not used in the training.

Once the scoring rubrics for all machine-scored constructed-response items are validated, all constructed-response items will be rescored using the final rubrics, and the final data file will be extracted for the item analyses. The item analyses include classical item statistics and item calibrations using the 3PL IRT model for multiple-choice items and the generalized partial credit model for polytomous items. 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).

**Classical Item Analysis**

Classical item analyses ensure that items function as intended with respect to the underlying scales. AIR’s analysis program computes the necessary individual item and overall test statistics for each multiple-choice and constructed-response item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item. Note that for adaptive tests the variable used to compute the biserial correlations and the DIF ability stratification is either the IRT-based ability estimate or the scale score, rather than the total raw score used for paper-and-pencil tests. In a computer-adaptive test, the raw score is not meaningful because each student sees different items. Therefore, where an overall score is needed, we use the IRT-based ability estimate or scale score, which accounts for the different items administered.

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 test-takers 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 is calculated as the correlation between the item score and the student’s IRT-based ability estimate (biserial correlations for multiple-choice items and polyse-
Differential Item Functioning (DIF): DIF analyses are designed to determine whether students at similar levels of ability have different probabilities of answering the same item correctly (or of receiving higher scores in the case of constructed-response items) based on a group membership. A variety of factors may lead to differential item functioning, but DIF may indicate item bias. For this reason, a bias and sensitivity review committee must review all items classified as DIF to determine whether an item is unfair to members of targeted student subgroup populations.

AIR will conduct DIF analyses on all items included in the field test to detect potential item bias for subgroups. The performance on each item by focal group members (e.g., protected ethnic group members, females) will be compared with the performance of the appropriate reference group (e.g., white students, male students). The purpose of these analyses is to identify items that may have favored students in one group (reference group) over students of similar ability in another group (focal group).

The procedures that AIR uses for detecting DIF are the Mantel-Haenszel (MH) chi-square for dichotomous items (multiple-choice items) and Mantel's chi-square for polytomous items (constructed-response items). AIR will calculate the Mantel-Haenszel statistic (MH D-DIF) for multiple-choice items (Holland & Thayer, 1988) and standardized mean difference (SMD) for constructed-response items (Zwick, Donoghue, & Grima, 1993) to measure the degree and magnitude of DIF. The total scale score for each student on the test will be used as the ability-matching variable. The total score will be 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 will compute 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, as well as the MH chi-square, the SMD, and the standard error of the SMD for the constructed-response items. The purification method described by Holland and Thayer (1988) will be included in the DIF procedure. Items will be classified into three categories (A, B, or C) ranging from no DIF to mild DIF to severe DIF according to the DIF classification convention. Items will also be categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group, or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group.

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 will be flagged and will be reviewed by the Fairness Data Review Committee. Exhibit I.D.1-4 details the DIF classification rules.

In addition to the classical item statistics, AIR will provide IRT item parameter estimates and item fit information for all field-test items.

### Calibrating, Scaling, and Equating

For the independent field test, all field-test items will be concurrently calibrated, equating all the field-test items to each other (a common-item design) and creating an initial online item pool on the same scale. The linkages among items are more robust than using a set of anchor items across fixed forms because each item is linked with every other item across hundreds of unique test “forms.” Field-test item parameters in the initial item pool become the operational item pool for the adaptive operational test administrations. For the embedded field test, the field-test items will be calibrated together, fixing the precalibrated operational item parameters and placing the field-test items on the operational scale. The EFT items will be added to the operational item pool for subsequent administrations.

For all online field-test items (in both independent and embedded field tests), we propose a minimum sample size of approximately 1,000 students per item. The resulting item parameters will be sufficiently stable to support accurate equating, stable state-level results, and precise student scores.

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. AIR proposes to continue to use statistical models based on IRT (Allen & Yen, 1979; Birnbaum,

---

### Exhibit I.D.1-4: DIF Classification Rules

<table>
<thead>
<tr>
<th>DIF Category</th>
<th>Flag Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dichotomous Items</strong></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>MH$\chi^2$ is significant and $</td>
</tr>
<tr>
<td>B</td>
<td>MH$\chi^2$ is significant and $</td>
</tr>
<tr>
<td>A</td>
<td>MH$\chi^2$ is not significant.</td>
</tr>
<tr>
<td><strong>Polytomous Items</strong></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>MH$\chi^2$ is significant and $</td>
</tr>
<tr>
<td>B</td>
<td>MH$\chi^2$ is significant and $</td>
</tr>
<tr>
<td>A</td>
<td>MH$\chi^2$ is not significant.</td>
</tr>
</tbody>
</table>
1968; Lord & Novick, 1968) to calibrate the items to a common scale. AIR is experienced with a wide range of IRT models, extending from the simple one-parameter (or Rasch) model (Rasch, 1980; Cohen & Kolstad, 2000) to multi-parameter models (Johnson, 1992; Beaton, Johnson, & Ferris, 1987) and even to cutting-edge multi-component models (Cohen & Kolstad, 2000).

Like the National Assessment of Educational Progress (NAEP), the Utah state assessments use multi-parameter models. These more complex models (essentially, models that acknowledge the possibility of guessing on multiple-choice items and recognize that some items measure with less error than others) do not offer a one-to-one correspondence between raw scores and scale scores. Some psychometricians (Thissen & Orlando, 2000; Thissen, Nelson, & Swygert, 2000; Yen, personal communication, 1999) advocate mechanisms by which raw score conversion tables can approximate the scores that would be assigned on the basis of these more complex models. AIR recommends that scores be calculated using maximum likelihood with boundaries for the perfect scores (all correct or below chance).

### Overview of Some Common IRT Models

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:

\[
L(Z) = \prod_{j=1}^{n} P(z_j \mid \theta),
\]

where \(Z\) represents the pattern of item responses, and \(\theta\) 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), which is used for items that are scored either right or wrong, takes the form

\[
P(x_j = 1 \mid \theta_k) = \frac{1}{1 + e^{1.7 \theta_k - b_j}} = p_{j1}(\theta_k).
\]

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. (The typical formulation of the Rasch model does not include the arbitrary constant 1.7 factor in the exponential—the factor is included here for consistency with the more general models described below.)

The two-parameter logistic model (2PL) takes the form

\[
P(x_j = 1 \mid \theta_k, a_j, b_j) = \frac{1}{1 + e^{1.7 a_j \theta_k - b_j}} = p_{j2}(\theta_k).
\]

The \(b\) parameter is still called the location or difficulty parameter. The \(a\) parameter is often 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.

For multiple-choice models, the three-parameter logistic model (3PL) adds a "guessing
parameter,” so the probability of a correct response is given by

\[ P(x_j = 1|8_k, a_j, b_j, j) = \frac{1-c_j}{1+e^{-1.7a_j e_k-b_j}} = p_j(8_k). \]

The third parameter, \( c \), defines a lower asymptote. In the one- and two-parameter model, the probability of a correct response approaches zero as proficiency decreases toward negative infinity. The third 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 right answer. The guessing 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 generalization of Samejima’s (1972) graded response model. For the smaller-sample tests (such as modified or alternative 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 \( mj \) categories can be written as

\[ P(x_j = i|8_k, b_j = b_{mj-1} = \frac{e^{1.7a_j e_k-b_j}}{1+e^{1.7a_j e_k-b_j}}. \]

Muraki’s generalization adds an item-dependent discrimination parameter as follows (again, Masters’ formulation does not usually include the arbitrary constant 1.7):  

\[ P(x_j = i|8_k, b_j = b_{mj-1} = \frac{e^{1.7a_j e_k-b_j}}{1+e^{1.7a_j e_k-b_j}}. \]

Returning to the likelihood equation, the contribution of each item to the overall likelihood function remains independent of all other items, given \( 8 \). 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.

### Estimation of IRT Models

AIR will use any software requested by USOE to calibrate the data. 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 psychometricians at AIR regularly use Parscale, Multilog, IRTPRO, and other software and are prepared to use any software requested by USOE.

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 construct-ed 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 distribu-
ation in an approach known as *marginal* maximum likelihood estimation. 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 \( \chi^2 \) 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 correct-ly accounted for, the \( \chi^2 \) becomes an \( F \) statistic. \( iAM \) reports fit statistics along with innovative graphics that characterize the item fit, as shown in Exhibit I.D.1-5.

Exhibit I.D.1-5: Item Fit Graphics

The top panel shows a sample fit graph for a well-fitting multiple-choice item, and the bot-
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 shows a well-fitting item as misfit appears only at the tails of the distribution, where data are 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.

**Equating**

As discussed in Section I.C.3.c, for online adaptive tests, we recommend a pre-equated design. Under this design, following the initial calibration and equating of the bank item parameter estimates, embedded field-test items will be calibrated subject to the anchor formed by the operational items as described in Section II.B.2. The Utah student population is sufficiently large to allow for embedded field-testing of large numbers of items while still obtaining numbers of student responses per item to yield highly precise item parameter estimates.

Within the context of adaptive testing, however, highly reliable student ability estimates can be obtained even when item parameter estimates are less reliable than desired. It turns out that with an adaptive test, both aggregate estimates and individual student scores are more precise if there are more items in the pool, even if the parameters for each item are estimated with less precision.

Consider first the aggregate scores. For simplicity of explanation, consider a Rasch model, although the same logic holds for more general IRT models. The equating variance may be approximated by

$$\text{var}(A) = \frac{1}{N} \text{E}(\frac{1}{N} \sum_{j=1}^{N} (\frac{E_j}{A})^2),$$

where $A$ is the linking constant, $N$ is the number of linking items, and each item $j$ has sampling variance $\frac{1}{N} \sum_{j=1}^{N} (\frac{E_j}{A})^2$. In a fixed-form test, the linkage affecting an aggregate is based, at best, on a single form’s items—say 50. In an adaptive test, each test has a different mix of items from the pool. Effectively, each test represents a unique, though not fully independent, draw on the distribution of linking errors. Hence, all the items in the pool may contribute to the linkage. A pool of 400 or 500 items may have 10 times as many items contributing to the linkage, resulting in sampling variance of the
linkage an order of magnitude smaller. The sample size associated with each item parameter estimate could be reduced by the same proportion without affecting the aggregate statistics. Another way to think about this is that an adaptive test yields many draws on the linking distribution, and the standard error of the mean of those draws falls as the number of tests administered rises.

At the individual level, the linking error is typically a small fraction of the standard error of measurement. For many students, the increased precision at the ends of the distribution offsets the increased linking error due to less stable linking item parameters. In designing the field tests for very small states such as Utah, AIR investigated this issue at length, including conducting simulations based on bootstrap samples from real data. The theoretical expectations outlined above were upheld. Finally, we note that our approach has been used for years in Utah, and its system has produced consistent results over time.

### Test Integrity Analyses

Throughout the testing window, a set of quality assurance reports will be routinely generated and evaluated to ensure the quality of Utah’s statewide assessment system. The quality assurance reports provide information on item behavior and blueprint match rates. Additional reports include an analysis of item exposure rates and a cheating analysis report that will flag any unlikely patterns of behavior in a testing session. 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 for each project reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Exhibit I.D.1-6 presents an overview of the quality assurance (QA) reports.

#### Exhibit I.D.1-6: Overview of Quality Assurance Reports

<table>
<thead>
<tr>
<th>QA Reports</th>
<th>Purpose</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Statistics</td>
<td>To confirm whether items work as expected</td>
<td>Early detection of errors (key errors for multiple-choice items and scoring errors for constructed-response items)</td>
</tr>
<tr>
<td>Blueprint Match Rates</td>
<td>To monitor unexpected low blueprint match rates</td>
<td>Early detection of unexpected blueprint match issue</td>
</tr>
<tr>
<td>Item Exposure Rates</td>
<td>To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)</td>
<td>Early detection of any oversight in the blueprint specification</td>
</tr>
<tr>
<td>Cheating Analysis</td>
<td>To monitor testing irregularities</td>
<td>Early detection of testing irregularities</td>
</tr>
</tbody>
</table>
**Item Statistics**

A set of item analyses will be run for early detection of any unexpected issues in the online system, such as key error, scoring error, or an item breach. Aberrant item behaviors will be examined based on item *p*-value, item discrimination, and item fit statistics. In addition, the report can be configured so that only items with statistics falling outside a specified range are flagged for reporting.

**Item *p*-Value.** For multiple-choice (MC) items, the proportion of students selecting each response option is computed; for constructed-response (CR) items, the proportion of student responses classified at each score point is computed. For MC items, if the keyed response is not selected by the largest proportion of students, the item is flagged as exhibiting a non-modal keyed response.

**Item Discrimination.** Biserial correlations for the keyed response for MC items and polyserial correlations for CR items are computed. Currently bi/polyserial correlations for all items are reported, but USOE can choose to identify only those items with values below a threshold.

**Item Fit.** In addition to the *p*-value 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.

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 multiple-choice items we use the 3PL IRT model to calculate the expected score on item *j* for student *i* with estimated ability \( \theta_i \) as

\[
E(z_{ij}) = \sum_{j=1}^{K_j} \frac{\exp(D_{aj} (\theta_i - b_{j,k}))}{\sum_{k=1}^{K_j} \exp(D_{aj} (\theta_i - b_{j,k}))} 
= 1 \sum_{m=1}^{K_j} \exp(D_{aj} m \sum_{k=1}^{m} (\theta_i - b_{j,k}))
\]

For each item *j*, the residual between observed and expected score for each student is defined as

\[
o_{ij} = z_{ij} - E(z_{ij})
\]

The statistic \( \delta_j \) is aggregated across students of different abilities for each item,

\[
\delta_j = \frac{1}{n} \sum_{i=1}^{n} \delta_{ij}
\]

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

\[
\frac{\delta_j}{\text{se}(\delta_j)} > 1.96, \text{ where } \text{se}(\delta_j) = \text{SD}(\delta_j) / \sqrt{n}.
\]
Two blueprint match reports will be generated for each assessment, one based on evaluation of match to content standards and a second for evaluation of affinity groups (e.g., item type, depth of knowledge). 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, and the number of administrations in which the blueprint requirements were not met. Administrations in which specifications were not met are reported by 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.

**Inconsistent Item Response Pattern (Person Fit)**

In IRT models, the 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 test-taker who responds to the items fairly and honestly.

If a test-taker has prior knowledge of some test items (or is given answers during the exam), he or she will respond correctly to these 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 this student. If a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index.

The person-fit index is based on all item responses. An unlikely response to a single test question may not result in flagging the person-fit index. However, unlikely patterns that do not indicate intentional testing irregularity may be detected for a lucky guesser 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 will be summarized for a testing session and a test administrator.

The person-fit index will be computed using the standardized log-likelihood statistic. Following Drasgow, Levine, Williams (1985), Sotaridona, Pornel, and Vallejo (2003) define aberrant response patterns as a deviation from the expected item score model. Snijders (2001) showed that the distribution of $l_2$ is asymptotically normal (i.e., with an increasing number of administered items, $I$). 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 $l_2$ for systematic flagging of aberrant response patterns. Aggregate units will be flagged when $t$ exceeds some threshold.

$$t = \frac{\text{Average } l_2 \text{ values}}{s_{l_2} \cdot \sqrt{n}},$$

where $s = \text{standard deviation of } l_2 \text{ values}$, and $n = \text{number of students in an aggregate unit (e.g., testing session, test administrator)}$. The QA report will include a list of the flagged aggregate units with the number of flagged stu-
dents in the aggregate unit (school, test administrator, test session).

Plan for Studies to Evaluate Evidence for Validity of Test Score Interpretations

AIR recognizes that peer review studies will be required for Utah’s new adaptive assessments to be fully approved by the U.S. Department of Education. AIR is committed to providing high-quality studies and clear documentation of those studies for Utah. Our experience with state testing programs, as well as a review of the federal critical elements, suggests that AIR can support the USOE to obtain validity evidence related to the following:

- Test Content
- Internal Structure
- Relations to Other Variables
- Consequences of Testing

Evidence Based on Test Content

Utah has adopted academic content standards that define what students should know and be able to do upon completion of instruction at each grade level and in the critical content areas. These expectations are explicated in the Common Core State Standards (CCSS) and the Utah Core Standards and will be assessed through the online adaptive assessment system. The Common Core Standards for Mathematics and Reading and the Utah Core Standards are both organized into multi-level frameworks. The CCSS include strands, anchor standards, and grade-level standards. Strands represent broad statements about the major content areas within each subject area and the learning goals for Utah students. Anchor standards denote more specifically what students should know and be able to do by the end of instruction in each subject area.

Grade-level standards are specific statements that identify exemplar knowledge and skills that indicate whether students have achieved an anchor standard.

The Utah statewide assessments measure student progress toward achievement of the CCSS and Utah Core Standards targeted for instruction in each grade level and course. The validity of Utah’s test score interpretations will be evaluated along several dimensions. As a criterion-referenced system of tests, the meaning of test scores will be, in part, appropriately evaluated by the degree to which test content is aligned with the CCSS and the Utah Core Standards.

Alignment of content standards is achieved through a rigorous test development process that stems from the CCSS and Utah Core Standards and refers back to those standards in a highly iterative test development process that includes USOE, AIR, and two committees composed of Utah citizens, including not only secondary educators representing each course but also other stakeholders. The Fairness Item Review Committee is charged with ensuring that test content does not unfairly disadvantage students because of group membership, and the Content Item Review Committee is specifically charged with ensuring that test content is aligned with academic content standards and is grade-appropriate. Committees review items twice during the item development process—once prior to the inclusion of items on field-test forms and again following field-testing and with item performance information, prior to the inclusion of items in the operational item bank.

This review-intensive item development process ensures that items in the operational test banks, and subsequently administered to students in either adaptive or fixed-form assessments, are tightly aligned to the CCSS and the Utah Core Standards and that test
scores provide valid feedback to parents and the educational community about student, classroom, school, district, and statewide progress toward achieving those standards.

In addition to this review-intensive item development process, it is necessary to formally investigate the alignment of Utah’s adaptive CRT test items to the CCSS and the Utah Core Standards. AIR will work with USOE to evaluate alignment to standards for Utah’s statewide assessments. Although we can conduct a standard alignment study as might be done with a fixed-form test, such a model is not particularly well suited for evaluating the alignment of an adaptive test. To apply standard procedures would require constructing multiple exemplar forms, forms that could have been administered to students following the adaptive blueprint, and evaluating these exemplar forms for alignment to the standards. USOE may want to consider a different, or additional, type of alignment study for the adaptive tests. The alignment of an adaptive test can be considered as two sequential questions:

1. Are the test items in the bank accurately aligned to the standards and benchmarks and depth-of-knowledge categories with which they are putatively associated?
2. Do the tests administered from the bank appropriately cover the content standards?

The first question can be answered using typical techniques, such as Webb’s alignment method (Webb, 2005). The second can be answered quickly and inexpensively through a statistical analysis of the range of content covered within tests and across tests. Alignment can be asserted if (a) the items are properly aligned, (b) each student sees an appropriate range of content, and (c) across all students the full set of testable standards is covered.

**Evidence Based on Internal Structure**

In addition to defining the content domain for academic achievement in each course, Utah’s statewide assessments represent a structural model of student achievement in grade level, as well as course-specific content areas. Following the first operational administration of each of Utah’s statewide assessments, AIR proposes to conduct a confirmatory factor analysis of the operational test items. The confirmatory factor analyses test the fit of the structural model posited in the CCSS and Utah Core Standards frameworks as implemented in Utah’s statewide assessments to student responses on the operational tests.

AIR will use commercially available software for structural equation modeling to perform the second-order confirmatory factor analyses on the structure of Utah’s adaptive CRT tests. Within each subject area, items measure a single standard. Standards 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 I.D.1-7. As the exhibit illustrates, each item is an indicator of a single primary factor or academic content standard. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each academic content standard serves as an indicator of achievement in a subject area. As at the item level, the content standards include an error term indicating that the content standards are not pure indicators of overall achievement in the subject area. The paths from the content standards to the items represent the first-order factor loadings, the degree to which items are correlated with the underlying academic content standard construct. Similarly, the paths from subject area achievement to the content standards represent the second-order factor loading, indicating the degree to which academic content standard constructs are correlated with the
underlying construct of subject area achievement.

For each operational test, we will report the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a $\chi^2$ statistic. Good model fit is indicated by a nonsignificant $\chi^2$ statistic. The $\chi^2$ statistic is sensitive to sample size, and even well-fitting models will demonstrate highly significant $\chi^2$ statistics given a very large number of students. Therefore, fit indices, such as the Bentler-Bonett Normed Fit Index (NFI; Bentler & Bonett, 1980) and the Comparative Fit Index (CFI; Bentler, 1990), which are more appropriate for very large samples, are also reported. Values above .9 on these indices indicate good model fit.

In addition to testing the overall fit of the second-order confirmatory factor analysis model, we will examine the degree to which the second-order model improves fit over the more general one-factor model of academic achievement in each subject area. Because the second-order model is nested within the one-factor general achievement model, a simple likelihood ratio test can be used to determine whether the added information provided by the structure of the CCSS and Utah Core Standards frameworks improves 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.

**Evidence Based on Relations to Other Variables**

Convergent and discriminant validity.

Validity evidence based on relations to other variables includes:

- **Exhibit I.D.1-7: Second-Order Confirmatory Factor Analysis Model**

![Diagram showing second-order confirmatory factor analysis model](image-url)
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.

To obtain evidence of convergent and discriminant validity for Utah’s statewide assessments, we propose to examine the patterns of correlations between test scores on Utah’s statewide assessments and teacher-assigned grades in academic course content. Teacher-assigned course grades provide us with an alternate indicator of student achievement of course objectives for students participating in Utah’s statewide assessments. Evidence for convergent and discriminant validity is determined by examining the patterns of correlations among Utah’s grade-level and course-specific statewide assessments and teacher-assigned grades.

Observed correlations between alternate indicators of student achievement of course objectives, such as teacher-assigned course grades and Utah’s statewide assessment scores, should be limited only by the unreliability of the measures. When both assessments measure student achievement in common subject areas, as with, for example, teacher-assigned grades and test scores based on statewide assessments in Algebra I, 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 Biology and Algebra 1 tests are, for example, quite different, Biology test scores should correlate less well with the teacher-assigned grades in Algebra 1 than do Algebra 1 test scores. 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 are 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. Previous studies indicate that we can expect cross-subject correlations about .6 in magnitude.

Evidence Based on Consequences of Testing

Messick (1995) argued that a unitary theory of validity should address not only the evidence supporting the meaning of test scores but also the consequences of test use. He argued that whether made explicit or not, testing applications likely include benefits and risks (e.g., assessment of developmental disability can lead to effective treatment but also carries social stigma). The justification for the test should therefore include an evaluation of whether a test’s use has the intended positive benefits, whether there are unintended negative consequences, and whether the benefits outweigh the costs.
Central to the notion of standards-based, accountability assessments is that implementation of centralized assessment systems ensures that schools hold students to a set of shared standards for learning outcomes that all students are expected to achieve and that schools are held accountable for attaining those outcomes. Assessment systems are explicitly designed to have consequences for the educational system. The question then is whether test scores are having the intended positive effect on instruction and student learning, which is the stated purpose of the test.

Because the focus of consequential validity is on the impacts that a testing system may have on the community, whether for individual students, parents, teachers, principals, superintendents, or aggregate units, including schools, complex areas, and statewide impacts, among others, evidence for consequential validity is less focused on the validity of inferences made from test scores and more focused on how test score use has an impact on various community members and organizations.

Although research suggests that the process of testing, independent of feedback, may have important implications for long-term retention of information (Roediger & Karpicke, 2006), the primary impact of testing, whether for students, educators, or parents, is based on the information value of the feedback provided. With Utah’s new adaptive statewide assessments, Utah is working to ensure that students statewide receive effective instruction in key grade-level and secondary school course content and that the achievement of instructional goals is evaluated along a common standard. The impact of these assessments and score reports, used to provide feedback about student, classroom, and school performance, should therefore be investigated.

We propose to conduct a series of focus groups with students, teachers, principals, superintendents, and parents to elicit the kinds of intended and unintended consequences that implementation of Utah’s adaptive statewide assessments has for students, classrooms, schools, and complex areas. Educators will be able to provide feedback concerning the kinds of impacts that Utah’s assessments are having on curricular decisions at all levels, the classroom and school instructional climate, and opportunities and needs for professional development in standards-based education and educational testing. Students and parents will be able to provide feedback about how the CCSS and the Utah Core Standards are communicated to them, how feedback on the adaptive assessments translates into actions in the home and influences student-teacher and parent-teacher communication, and whether Utah’s statewide assessment results are perceived as providing meaningful information about the academic progress of students.

**Technical Reporting**

**Annual Technical Report**

For Utah’s online adaptive statewide assessment system, AIR will produce a multi-volume technical report, providing comprehensive documentation of all aspects of Utah’s statewide assessment system—test development, technical quality, test administration procedures, and reporting of test results.

In Section IV.E, Technical Manuals, we discuss in detail the plan to obtain federal peer review approval for the Utah technical manual.
I.D.2 Plan for Enacting Complex Blueprints

As noted previously, because the adaptive system implements a set of algorithms to administer test forms to students, computations and algorithmic logic take the place of professional judgment in the item selection process. AIR’s adaptive algorithm is highly configurable, allowing USOE to vary the weighting of different blueprint constraints to control the balance between blueprint match and the precision of measurement. As described previously, the adaptive algorithm attempts to select items to meet three objectives:

1. Match the blueprint, which specifies items along multiple dimensions
2. Accurately classify as many students as possible with respect to the reporting categories or content domains
3. Minimize the measurement variance

Test blueprints in turn are designed to accomplish multiple objectives. Blueprints specify ranges across multiple levels of the academic standards framework to ensure that test scores reflect student performance on a representative sample of the content domain defined by the standards. Thus, blueprints specify ranges of items and score points at each level of the hierarchy, indicating the number of test items that contribute to each of the content standard domains or reporting categories, and in turn specify the distribution of items across standards within each reporting category.

Test blueprints guide the construction of test forms (administrations) to accomplish other important objectives as well. The CCSS and Utah Core Standards place greater emphasis on demonstrating subject area achievement through complex cognitive processes, requiring assessments aligned to the Utah Core Standards to measure knowledge and skills more deeply. To accomplish this, the adaptive blueprint allows users to specify a range of additional constraints that operate in conjunction with the academic content standards. We refer to these blueprint elements as affinity groups. Commonly employed affinity groups include depth of knowledge (DoK), item type, and genre. As with other blueprint elements, a range of items is specified for each affinity group (e.g., each level of DoK, or each of multiple item types such as multiple-choice, constructed-response, and simulation). The adaptive algorithm seeks items that best fulfill the multiple blueprint objectives, while also seeking to maximize the diagnostic utility of domain scores and precision of overall ability estimates.

While overall test scores are critical for accountability, educators are equally interested in more fine-grained evaluations of student achievement, such as domain scores, that they can use to tailor their instructional strategies. The diagnostic utility of domain scores is limited by unreliability of the typically very short subscale scores, and the very high intercorrelations among domains within a subject area. Identifying differences in student achievement across highly intercorrelated domains requires very precise measurement of domain performance. Otherwise, real differences in student achievement across domains will be masked by measurement error. AIR addresses this issue in two ways. First, we report domain performance levels, classifying student performance as above or below the proficient standard or as residing in a band of uncertainty defined by the standard error of measurement. Performance-level classifications are more reliable than domain scores. The adaptive algorithm then seeks to maximize the likelihood that a student’s performance in each domain is above or below the performance standard, increasing the precision of domain performance-level classifications. This provides educa-
tors with more ability to identify areas of strength and weakness.

Consistent with its description as an adaptive algorithm, maximizing the test information near students’ ability estimates is also an important objective of the assessment system. At each stage of the test administration, the algorithm uses the most current estimate of student ability to select the item with a difficulty parameter nearest to the student’s ability, resulting in greater precision of students’ scale scores, especially for students at the high and low ends of the ability distribution. Typically, a graph of the standard error of measurement in fixed-form tests is U-shaped, with test information greatest near the proficient standard or population mean and less information near the ends of the scale. In adaptive tests, the standard error curve is relatively flat across a wide range of ability, indicating greater measurement precision for a wider swath of the population.

**Blueprint Construction**

Proposed blueprints for each of the grade level and course-specific assessments are provided in Appendix A. Distribution of item ranges is designed to accomplish several important goals. First, as a criterion-referenced test, the Utah 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 CCSS and Utah Core Standard domain scores, each domain must be assessed with sufficient items to support reliable subscale performance-level reporting. The blueprints are designed to accomplish both goals, with items targeting the full breadth of the academic content standards but constrained to ensure enough 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 are provided the opportunity to demonstrate their knowledge and skills through the range of item types.

In addition to providing ranges of items for each content standard and affinity group, the adaptive algorithm provides additional mechanisms for controlling test administrations, including blueprint weights and the ability to designate that critical item ranges be strictly held. Because the Utah statewide assessments are criterion-referenced tests, test termination rules are properly determined by content coverage. The typical implementation of the adaptive algorithm therefore sets the total number of test items as fixed and specifies that the range of items for each content domain is to be strictly enforced. This ensures that all students are administered tests of the same length and that each student is administered sufficient items from each content domain to support student-level reporting of performance in each domain.

In some assessment contexts, there are content domains that must be assessed for purposes of content validity and alignment but that would require over-representation in the blueprint to support student-level reporting of test results. The adaptive algorithm, in conjunction with AIR’s real-time online reporting system, is uniquely capable of resolving this dilemma. In the event of conceptually important domains defined by few standards, AIR can construct a blueprint that ensures coverage of the domain but provides score reporting only at aggregate levels. In this case, while each student may be administered only a few items representing the content domain, because the test is administered adaptively across all test administrations in a classroom, students are likely to have been administered a large sample of items representing the domain, yielding both highly reliable and content-valid indices of perform-
ance at the aggregate level. AIR’s benchmark-level reporting—described in I.C.3.e, Score Reporting—also takes advantage of variable item selection in the adaptive algorithm to support our strengths and weaknesses reports. These reports provide educators with highly detailed summaries of classroom-level performance.

**Test Configuration**

Constructing forms for adaptive tests consists of two steps: establishing the item pool and then optimizing the item selection parameters for the selected pool. Our adaptive engine is robust to imbalances in the item pool as long as there are enough 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 noted above, 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 within the item pool be proportional to the test blueprint. Nevertheless, as we describe in II.B.2, Item Development, AIR test development and psychometric staff inventory the item pools prior to 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 within content domains to guide our item development strategy.

With respect to item pool adequacy, we note that the evaluation and the selection of items into the operational pool proceed somewhat differently between an adaptive and 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 within 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 USOE’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. USOE will receive a final report on all simulations containing the detailed statistical output.

**Paper Test Form Construction**

Construction of fixed forms begins with the identification of a target or reference form or bank. For Utah’s fixed-form tests, we will target test information to minimize standard errors near the proficient cut score. More generally, we will seek to distribute test information consistent with the item bank, with the caveat that because the item banks will be explicitly designed to measure achievement across the full range of achievement, the fixed-form assessment will not be capable of measuring very low- and high-achieving students with the same level of precision as the adaptive algorithm.

We propose using the adaptive engine to automate construction of initial paper forms. AIR will generate an initial set of forms using the same computer-adaptive engine used for all other students. AIR will input to the adaptive algorithm a hypothetical student with an ability estimate exactly at the proficient performance standard and allow the adaptive engine to select items for the hypothetical student. This process will yield a test form for which the test information function is maximized near the proficient cut-point and the items conform to the test specifications. A test development team will review the form and may replace some items after consulting with USOE. No field-test items will appear on the paper forms.

The auto-generation of the paper form brings an additional benefit with respect to equating. The auto-generation process draws items from the operational item bank. Because a pre-equating strategy is used, scores generated from these forms are on the same scale as the forms constructed for students online and are immediately comparable without subsequent need for any other linking strategy.

**I.D.3. Plan for Setting Performance Standards**

AIR’s plan for recommending performance standards ensures that cut scores for grade-level assessments are vertically articulated, so that any changes in percentage of students scoring at or above the proficient level can be meaningfully tied to maturational or instructional differences between grade levels, rather
than an idiosyncratic difference between recommended performance standards across grade-level panels. Moreover, AIR recommends that panelists be provided with national and international benchmark data to provide greater context to the meaning of statewide assessment scores and to encourage panelists to recommend performance standards that are nationally and internationally competitive. As a cost option, AIR will establish linkages between the Utah statewide assessments and PISA test scores, which, in addition to available NAEP results, can provide important benchmark information.

**Bookmark Procedure.** To ensure vertically articulated performance standards across grade-level assessments, AIR proposes to employ a modified Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001) to establish cut scores for the Utah statewide assessments. In the modified bookmark approach, AIR convenes standard-setting panels by grand-bands (e.g., grades 3–5, 6–8, and 9–11 for English Language Arts). Panels begin by recommending performance standards in anchor grades (e.g., 4, 7, and 10), using the standard bookmark method. Following recommendation of cut scores for the anchor grades, panelists convene to vertically articulate performance standards across the anchor grades. Once recommended cut scores for the anchor grade have been established, AIR then interpolates the location of performance standards in the remaining grades. With interpolated cut scores 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 (RP level) of just-barely-meet-the-standard students would be expected to successfully respond to the item. If yes, panelists are asked to endorse the interpolated cut score, but if no, panelists are asked whether a bookmark location can be identified in the vicinity of the interpolated cut score. The benefit of this approach is that panelists are able to explicitly take into consideration the important goal of recommending vertically coherent performance standards when making bookmark placements. AIR has successfully used this modified Bookmark method to recommend performance standards for other state assessments.

**Augmenting Ordered-Item Booklets.** Although we would ideally like to recommend performance standards on the bank of items comprising each statewide assessment, such an approach would be too cumbersome for standard setters. Recommending performance standards on a single operational test form, however, is often too restrictive, requiring panelists to recommend performance standards based on a highly constrained sample of items that cannot represent the range of content standards and benchmarks across the spectrum of student achievement. We therefore recommend using an augmented ordered-item booklet (OIB) to recommend performance standards. We propose to construct an OIB that is proportional to the blueprint but contains more items than any one operational test administration. Augmenting the OIB in this way helps ensure that standard setters are making judgments on a more representative range of the content standards assessed. Augmenting the OIB also helps ensure that there are sufficient items in the higher and lower ability ranges to support bookmark placements for cut scores at the higher and lower performance levels. For example, when the OIB is constructed from a single operational test form or administration, it is often the case that moving the bookmark placement by even one page can have an outsized impact, and panelists are reluctant to make bookmark placements based on content considerations when shifting the placement of a bookmark by even one page can result in 5% or more additional students not meeting the standard. Augmenting the OIB modulates the impact of bookmark placements in the higher and lower
ability regions, freeing panelists to focus on content considerations when placing bookmarks.

Workshop Panelists and Stakeholder Review Committees. While it is critically important to include a range of stakeholders in the standard-setting process, our experience has shown that it is essential for panelists to have direct knowledge of academic standards and student grade-level performance to participate meaningfully in the bookmarking process. For this reason, we propose to restrict panel participation to classroom teachers and curriculum specialists with expertise in the subject area content for each assessment. To ensure that the widest range of stakeholders has meaningful input into the standard-setting process, we propose to conduct a stakeholder review and moderation session with a more general group comprising table leaders from the bookmarking workshop and representatives from other important stakeholder groups, including members from the educational community such as building and district administrators, local school board members, and faculty from higher education. We also propose including members from outside the educational community, including parents and business and community representatives.

Performance-Level Descriptors. Developing a clear and meaningful description of each achievement level is central to establishing reliable performance standards and effectively communicating assessment results to parents, educators, and other stakeholders. AIR test development and score reporting staff collaborate to draft performance-level descriptors (PLDs), which will then be submitted to USOE for review and subsequent approval. All PLDs are reviewed for sensitivity and lack of bias and are professionally edited. The draft PLDs will be reviewed at a PLD review meeting conducted prior to the standard-setting workshop. Performance-level descriptors (PLDs) define the content area knowledge, skills, and processes that test-takers at a 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 Common Core State Standards and the Utah Core Standards to guide them in placing their bookmarks.

AIR will work with USOE to draft appropriate PLDs. Prior to standard setting, AIR will work with USOE to create PLDs that are clear, concrete, and reflect Utah’s expectations for proficiency. This process usually requires multiple rounds of review, revision, and collaboration.

AIR will then convene panels of Utah educators to review and revise the PLDs. The PLDs jointly authored by USOE and AIR will be used as input for these meetings, and we expect that the panels of educators will fine-tune them. We will capture all comments and moderate the meetings to achieve consensus on all revisions. The proposed revisions will be documented and delivered to USOE. AIR will work with USOE to resolve all comments and make the approved changes.

Standard-Setting Materials and Procedures

Panelist Recruitment. For grade-level assessments, AIR will convene elementary, middle, and, for ELA, high school panels. For course-based assessments that require specific content expertise, AIR will convene a separate panel for each course-based assessment. As indicated in Exhibit I.D.3-1, AIR will recruit approximately 15 panelists for each standard-setting panel to recommend performance standards. For grade-level assessments, this represents approximately five panelists per grade in each grade-band panel. To recruit panelists, we
will actively seek individuals from the diversity of backgrounds found in Utah. We will seek the representation of males and females found in the grade- and course-specific teacher population of Utah. The same principle will be applied to the geographical representation of panelists. In addition, we will strive for proportional representation of racial and ethnic backgrounds and members with special education students and English language learners.

**Training.** An essential element of a standard-setting workshop is thorough 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 test specifications, the Common Core State Standards and Utah Core Standards, and the PLDs for each performance standard.

Table leaders will have the additional responsibility of ensuring that table activities remain focused on the task at hand, ensuring that panelists understand the task, and alerting workshop leaders of any issues encountered by panelists as they engage in their workshop tasks. Table leaders are not expected to provide training to panelists; instead, they serve as liaisons between the panelists and workshop leaders to ensure that workshop activities are implemented correctly, alert workshop leaders of issues that may arise during the course of conducting workshop activities, and represent the table in the stakeholder deliberations. We will convene a table leader orientation meeting prior to the standard-setting workshop to familiarize table leaders with their role and responsibilities, including suggestions on how to provide leadership at the tables during the standard-setting process and how to manage the secure materials.

Central to their training, panelists will use the PLDs to develop a representation of students who just barely meet each of the performance-level descriptions. 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. Characterizing students who just barely meet expectations is not an intuitive judgment, and panelists are encouraged to spend time discussing with one another the minimum characteristics of student achievement for entry into each performance level.

Finally, panelists will receive training in the placement of bookmarks in the ordered-item booklets (OIBs). As part of this training, panelists learn to place a bookmark between the two items that best delineate two achievement levels (e.g., between items on which students must demonstrate mastery to meet the minimum requirements for the proficient level and those items on which demonstration of mastery is not necessary). In addition to showing panelists how to place bookmarks in the OIB

### Exhibit I.D.3-1: Standard-Setting Panels and Panelists

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade/Course Panel</th>
<th>Number of Panelists</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ELA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary (3–5)</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Middle (6–8)</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>High (9–11)</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td><strong>Math</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary (3–6)</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>Math 7</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Pre-Algebra</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Algebra 1</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Geometry</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Algebra 2</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td><strong>Science</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary (4–5)</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Middle (6–8)</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Earth</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Biology</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Chemistry</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Physics</td>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>
as part of the large-group training, AIR will provide sample OIBs to panelists so that they can practice placing bookmarks in the OIB.

**Bookmark Placement.** Following training in the Bookmark procedure, panelists will receive anchor grade OIBs. For each item in the OIB, panelists will be instructed to ask what a student must know and be able to do to answer each question and what makes each item in the OIB more difficult than the preceding item. This review of the OIB allows panelists to gain new perspectives on the knowledge and skill requirements of items and to share information regarding their thoughts on the location of the threshold region. During this discussion, the workshop leader will circulate through the room to monitor progress, assist panelists who are having trouble with the task, and answer any questions.

Using the PLDs as a guide, panelists will then be instructed to place a bookmark between the pages in the OIB that best delineate each of the performance levels, beginning with the proficient level performance standard. Once panelists feel comfortable about their proficient level cut, they will move on to place bookmarks for the other performance standards.

Bookmark placement is conducted in two rounds, allowing panelists to make independent judgments while still benefitting from discussion with their fellow panelists. The placement of the bookmark is illustrated in Exhibit I.D.3-2. Each panelist uses the PLDs and the content standards to decide which item represents the lower bound of each achievement level. In the example, a panelist concludes that students who are just barely at the proficient level would demonstrate mastery on the item on page 25, whereas students below the proficient standard would not. Therefore, the panelist decides that the proficient level performance standard begins on page 25. The panelist believes that students below the proficient performance level will not be able to demonstrate mastery of items beyond those on page 24 in the OIB.

**Feedback and Impact Data.** Prior to round two, AIR psychometricians will evaluate results from round one and provide feedback forms for each table and for the room as a whole. The forms for each table contain summary statistics showing the median, lowest,
and highest cut scores for that table. The room form contains summary statistics showing the median, lowest, and highest cut scores for each table. In addition, an impact data sheet containing scale scores and the percentage of students at or above the level for each possible cut score is provided to panelists for reference and discussion.

Round two begins with a discussion of two types of information:

- Feedback on agreement data, beginning with table-level feedback and discussion and progressing to room-level discussion
- Impact data (scale scores and percentage at or above each cut score) from the operational test administration

Each table spends time reviewing and discussing cut score placements, focusing on the lowest and highest recommended cut scores and the table’s median score. Panelists review the items between the lowest and highest cut scores for proficiency at their table, discussing the PLDs, as well as the newly introduced impact data. Discussion then expands to the room level, with each table reviewing the basis for its own recommendations for the group at large.

After completing these discussions, panelists again work through the OIB, placing their cut scores for all three performance levels.

**Vertical Articulation and Moderation.** Following the recommendation of performance standards for the initial grades, we propose to conduct a moderation session comprising table leaders from each of the panels. Although all panelists will be allowed to attend the moderation sessions as observers, participation will be limited to table leaders to effectively manage moderation session activities. The task of table leaders in the cross-grade moderation session will be to evaluate whether the recommended performance standards are reasonable across all grade levels assessed and to make modifications to the location of the standards if necessary.

To recommend performance standards in the intermediate grades, we modify the traditional bookmark procedures so that panelists are instructed to determine whether the PLDs support the placement of a specific bookmark on the interpolated page. If the PLDs do not support the placement of the bookmark on the interpolated page, then panelists are asked whether they can identify a bookmark placement near the interpolated page that would be supported by the PLDs. Panelists are instructed that their bookmark placements must be guided by content considerations, regardless of whether they recommend the interpolated page in the OIB or another bookmark placement.

**Stakeholder Review and Moderation.** To achieve a coherent set of standards across the system of statewide assessments that reflects the input of a wide range of stakeholders, both within and outside the education community, AIR will convene the workshop table leaders and stakeholders for the purpose of evaluating the performance standards recommended by the panel and smoothing the cut scores to achieve a coherent system of standards across grades. Stakeholders and table leaders will have the opportunity to offer additional recommendations on the proposed standards to USOE.

**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 record panelists’ decisions, provide feedback to the workshop participants, and draft the technical report. The room and all workshop materials will be kept secure at all times. The workshop participants will sign a
nondisclosure agreement, and a research assistant will make sure that no secure materials are taken from the room.

AIR staff will arrive one day early for the meeting to set up equipment and distribute materials. AIR staff will also be on-site for one day after the workshop to clean up the site, shred secure documents, and prepare statistical results for the standard-setting technical report.

Evaluation. Panelists will have the opportunity to provide feedback concerning the procedures and outcomes of the standard-setting workshop through group discussions and a series of standardized evaluation forms. Panelists will be asked to evaluate aspects of the standard-setting process throughout the workshop; these interim evaluations will inform workshop facilitators about the readiness of the panelists to recommend performance standards and allow workshop leaders to address issues that may arise, including, for example, the need for additional time or training.

Prior to placing recommended cuts in each round and for each performance level, panelists will be asked to complete a readiness form to indicate their preparedness to place their OIB bookmarks. This form will also ask panelists to identify any aspects of their task that are not clear. If any panelists indicate that they do not feel prepared, the workshop leader will provide additional training and opportunity for discussion. Panelists must indicate that they feel ready to move forward before they recommend a cut.

At the completion of the workshop, panelists will be asked to complete an evaluation form designed to elicit feedback on all aspects of the workshop, including clarity of training and tasks, appropriateness of the time spent on activities, and satisfaction with the outcome of the workshop. Appendix C presents a sample of a panelist evaluation form.

Results of these evaluations will be summarized and presented in the standard-setting technical report.

Technical Report. AIR will provide a comprehensive standard-setting technical report that will be available for review by USOE and the Technical Advisory Committee (TAC). It will detail the procedures, participants, and other relevant elements. The standard-setting technical report will be specifically oriented toward providing the type of comprehensive details required by NCLB peer review. The report will contain these sections:

1. Executive Summary: written for policymakers and can be used by USOE for workshops and provided to the Board of Education
2. Background and History of Utah’s statewide Assessments
3. Preparation for the Standard-Setting Workshop: covers the development of the PLDs, psychometric details of the field test, item calibrations and the creation of the vertical scale, estimation of impact data, explanation of the articulation process and international benchmarking procedures, and the composition and demographics of the standard-setting panels
4. Conduct of the Standard-Setting Workshop: includes agendas, training procedures, details of the round one and round two estimates of cut scores, stakeholder review and moderation sessions, security procedures, and evaluation
5. Results of the Standard-Setting Workshop: includes summaries of each round, reliability of the achievement-level standards, changes in cut scores resulting from the stakeholder review and moderation ses-
sions, final cut scores, and the percentage at and above each achievement level.

In addition, AIR will assist USOE with any materials or any other assistance needed for state Board of Education approval of the standards.

References


Cohen, J., & American Institutes for Research. (2002). *AM statistical software* (Beta Version 0.06.00) [Computer software].


II. Summative and Interim Test Content
II.B.1. Overall Approach

Overview

AIR proposes to develop the item banks for the adaptive delivery of Utah’s ELA, Mathematics, and Science tests. We recognize and applaud Utah’s strong history of involving educators in assessment development. Assessment literacy, and the deep understanding that comes from the active learning involved in developing items, makes teachers better at their craft.

AIR proposes that the majority of new items will be developed and reviewed by Utah teachers. Each item will also go through a series of additional editorial and content reviews at AIR and USOE and will be presented to content and fairness committees of educators and stakeholders. Finally, items will go to a committee of parents for a final review.

AIR’s Item Tracking System (ITS), which was described in detail in Section I.C.1 of this proposal, provides a powerful tool for involving teachers. ITS incorporates powerful tools for test developers to directly create technology-based items. The point-and-click approach enables test developers to bring their visions to life. Utah’s teachers will find themselves able to exploit these tools to develop next-generation items with ease and comfort.

Working with AIR’s item developers from well-defined development plans, we will collaborate with Utah’s teachers to create items that

- cover the full range of content;
- cover the full range of depth-of-knowledge; and
- utilize a broad and varied array of item types.

In this section we describe our plans for

- assembling an item pool from multiple sources;
- using teachers to create new items;
- ensuring the alignment of existing and new items; and
- ensuring that the items cover the breadth and depth of the content.

As part of the development and review process of all items, items will undergo a parental review, as required by House Bill 15.

Organization of This Section

AIR proposes to develop content in all three subject areas. Therefore, we integrate the description of our proposed item development process, item characteristics, and other requirements outlined in the RFP. We begin with a brief discussion of the content standards in each area and outline the distinguishing characteristics of the Utah Core Standards.

We offer sample items that demonstrate how our robust and flexible technologies support a wide variety of innovative items and item types and, specifically, how these technologies can be used to measure deep and meaningful content.

Finally, we address the specific requirements enumerated in the RFP. Subsequent sections describe the following:

- Our overall approach to ensuring that items, whether developed or obtained, are aligned to the Utah core and cover the full depth of knowledge. We present estimates of the number of items to be developed each year
- Our plans to obtain items from various sources and develop new items that are tightly aligned to the Utah core and cover the full depth-of-knowledge. (In this sec-
tion we address the requirements of Appendix A, which are incorporated by reference under this requirement.)

Our approach to creating an annual development plan each year to guide new item development and acquisition

Assurances that we will meet each technology requirement

Background

The Utah Core Standards provide the basis for instruction and assessment. When designing courses in language arts, math, social studies, and science, teachers are urged to follow the concepts delineated in the standards to promote consistency in instruction and assessment across the state. The standards provide the framework for major content understandings; moreover, they address clustered topics to be addressed during the school year.

In 2010, the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA) released the Common Core State Standards (CCSS). These standards were designed to provide students the focused, in-depth curriculum required for higher education and the workforce. The CCSS are

- research and evidence linked;
- aligned with college and work expectations;
- rigorous; and
- internationally benchmarked.

Input from educators, content specialists, parents, and other stakeholders contributed to the new, rigorous standards. On August 8, 2010, Utah officially adopted the CCSS. The CCSS builds on the foundation of Utah’s previous standards, continuing to provide the basis for instruction and assessment.

The CCSS provides clear, grade-level expectations for instruction and assessment. At the same time, the CCSS allows for flexibility within individual states, serving as the structure from which states and districts develop curriculum. The flexibility inherent in the standards allows educators to focus instruction and assessment on both the macro and micro levels. Four clear benefits of this approach exist: the curriculum developed (1) guarantees consistency across the state and the country, (2) addresses necessary knowledge and skills, (3) reflects local Utah values, and (4) allows instruction and assessment to target the individual needs of students. Additionally, administrators and educators can pinpoint professional development needs.

The CCSS does not replace the Utah Core Standards. Rather, the CCSS supplements the Utah Core and reflects an understanding of interdisciplinary connections. Recognizing the interconnection between subjects not only strengthens instruction and assessment but also provides another way to ensure that students are college and career ready. The resulting Utah Core Standards lead to quality instruction and assessment and ensure that all students have access to quality education.

The Utah Core Standards raise expectations for students. The quality and rigor of the content standards require students to apply higher-order thinking skills. The Utah Core delineates clear, grade-level expectations for all subject areas reflected in the standards. The standards are broken into smaller objectives and indicators that define the knowledge and skills needed. The standards from one year connect to ones in subsequent years, scaffolding for students and providing a clear scope and sequence for educators and parents. Each successive year works to expand and refine these skills.

Finally, the Utah Core also provides a stable target for assessment. Through formative and summative assessments, students must illus-
strate an understanding of subject material. In addition, students must apply the knowledge learned and skills gained to novel situations. Students must analyze, synthesize, and generalize concepts, increasing the cognitive demand. As a result, students are prepared for a wide range of postsecondary settings and the workforce.

Below, we provide background on the distinguishing characteristics of the Utah Core Standards in ELA, Mathematics, and Science.

**Overview of the Utah Core Standards in English Language Arts**

The Utah Core Standards for English Language Arts (ELA) are organized into two parts: College and Career Readiness (CCR) anchor standards and grade-specific standards. The CCR anchor standards work with the grade-specific standards to target necessary knowledge and skills to prepare students for postsecondary education and the workforce.

The CCR anchor standards provide broad targets, which guide the progression across grades. Focusing on the same key areas at each level ensures consistency of instruction and assessment and provides a clear map for educators, parents, and students.

The CCR anchor standards are broken down into grade-specific standards. The grade-specific standards focus on smaller objectives and indicators. While these are specific from one year to the next, the standards from each year connect to ones in subsequent grades. The structure allows for scaffolding for students and provides a clear scope and sequence for educators and parents.

Each successive year works to expand and refine skills learned the year before. The cognitive demand increases as students must show a deeper understanding of the knowledge and skills learned. For example, under the CCR anchor standard of Craft and Structure, students focus on point of view in literary texts. In the lower grades, students work to identify points of view and distinguish them from their own. The depth of knowledge needed to master the skills is relatively low. As the student progress, the cognitive complexity required also progresses. Students must compare and contrast different points of view used, and they must begin to interpret how the point of view affects the development of the plot. The depth of knowledge required is further increased as students move into middle school. Students must analyze how the point of view relates to the setting, character development, and resolution of the main conflict. In high school, the standard becomes even more specific. Students must extend and generalize their thinking to analyze literature from outside the United States.

The Utah Core Standards require a range of reading and an increase in text complexity. The consistency of CCR anchor standards between literary and informational texts illustrates that key elements exist between the two genres. The Utah Core Standards, however, recognize inherent differences, which, in turn, require different skills to comprehend and interpret various texts. Inclusion of a range of genre is essential. The Utah Core Standards also demand an increase in text complexity at all grades, thus increasing the rigor. The focus on these two elements helps ensure that students are prepared at the end of high school.

The Utah Core Standards in English Language Arts promote a belief in an “integrated model of literacy.” Standards not only address reading and writing skills but also focus on listening, speaking, and language. The standards incorporate interdisciplinary connections between ELA and history, social studies, science, and technical studies.
The writing standards also reflect ELA’s integral role throughout curricula. Students must complete writing assignments tied to research. They must work to interpret multiple sources of information, synthesize ideas, and formulate opinions. The students then must write about these ideas, providing valid reasoning and evidence to support their claims.

Overview of the Utah Core Standards in Mathematics

Three characteristics distinguish the Utah Core Standards in Mathematics:

- The standards focus on fewer topics in a more in-depth way for each grade.
- The topics align across grades to form coherent learning progressions.
- The standards identify and emphasize mathematical practices that cut across topics and are integral to mathematical thinking.

The Utah Core Standards focus more narrowly on critical areas at each grade level than was typical of state standards in the past. For example, in grade six, teachers are expected to focus on these four critical areas:

- Connecting ratio and rate to whole number multiplication and using concepts of ratio and rate to solve problems
- Completing understanding of division of fractions and extending the notion of number to the system of rational numbers, which includes negative numbers
- Writing, interpreting, and using expressions and equations
- Developing understanding of statistical thinking

The high school standards are grouped in conceptual categories: Number and Quantity, Algebra, Functions, Modeling, Geometry, and Statistics and Probability. Depending on the specific high school mathematics course, certain conceptual categories are more heavily emphasized than others. For example, geometry is more heavily emphasized in Mathematics II than it is in Mathematics I. Algebra and functions are fairly equally emphasized in Mathematics I and Mathematics II; however, the actual standards under these two conceptual categories vary depending upon the course.

These emphases align across grades to form coherent learning progressions in which the content from one grade supports the topics to be learned in subsequent grades. Elementary school standards (which have a heavy emphasis on numbers in base ten, fractions, and operations and algebraic thinking) are designed to prepare students for success in middle school mathematics. Middle school standards (which include ratio and proportional reasoning, equations and functions, etc.) emphasize the knowledge needed for high school mathematics. And the high school standards are designed to ensure college and career readiness. In particular, the high school standards emphasize mathematical modeling and allow for more advanced studies in mathematics.

Throughout the grades, the standards emphasize the eight Standards for Mathematical Practice identified in the Common Core State Standards:

- Make sense of problems and persevere in solving them
- Reason abstractly and quantitatively
- Construct viable arguments and critique the reasoning of others
- Model with mathematics
- Use appropriate tools strategically
Attend to precision
Look for and make use of structure
Look for and express regularity in repeated reasoning

These eight practices do not stand apart from the Standards for Mathematical Content (i.e., the actual mathematics content covered by the standards). Connections need to be made between the Standards for Mathematical Practice and the Standards for Mathematical Content, particularly within assessments. For example, when students are given tasks that demand understanding of a concept rather than pure recall, it is easier to see the connection between the Standards for Mathematical Practice and the Standards for Mathematical Content. Similarly, higher DOK tasks require a particular attention to the intersection between these two types of standards.

Overview of the Utah Core Standards in Science

Adopted in August 2010 by the Utah State Board of Education, the Utah Core Standards for Science (UCSS) emphasize equally understanding science content and using science skills. The standards recognize science as both a body of knowledge and a process for gaining insight into the natural world. Students are expected to demonstrate understanding of subject matter as well as proficiency in using scientific methods to explore, analyze, and make discoveries. The UCSS underwent a thorough review process by Utah science teachers, science educators, USOE staff, and others before adoption and is consistent with the vision of science education put forth by the American Association for the Advancement of Science Project 2061 and the National Academies of Science Framework for K-12 Science Education. This vision includes the following ideas:

- Standards will reflect the nature of science. The UCSS is designed around a set of Intended Learning Outcomes (ILOs, discussed in detail below) that require students to understand, perform, and communicate science concepts and skills.
- Standards will be coherent. The science concepts taught will have logical connections to each other and will progressively increase in complexity and cognitive demand at higher grades.
- Standards will be developmentally appropriate. The UCSS tailors the required science concepts and vocabulary to the developmental level of students at each grade, focusing on experiential learning at the early grades and more abstract ideas about the nature of science at later grades.
- Standards will encourage good teaching practices. The goals of the UCSS can be met only if students actively engage in the scientific process, individually and in
groups, and are encouraged to make connections between science and their everyday lives.

- **Standards will be comprehensive.** Instead of presenting science as a collection of unrelated facts and ideas, the UCSS focuses on a deep understanding of important scientific ideas and the nature of scientific thinking.

- **Standards will be feasible (specific to elementary science).** The UCSS includes only science concepts that can be easily demonstrated in a typical Utah classroom setting and provides resources for teachers that include UCSS-aligned lesson plans.

- **Standards will be useful and relevant.** Because science deals with the workings of the natural world, the skills that students attain in their science education are directly transferable and relevant to other aspects of their lives.

- **Standards will encourage good assessment practices.** The UCSS encourages performance-based assessments that require students to exhibit science skills as well as understanding of science concepts.

These guidelines ensure that the UCSS will be consistent with the aforementioned vision for science education and literacy. They also provide a framework for the logical progression of content with grade level in which the early grades focus on the exploratory aspect of science and later grades focus on increasing the sophistication of content and skills.

The UCSS presents standards for each grade from grade 3 to 8 and individual high school courses (Earth Systems Science, Biology, Chemistry, and Physics). Each grade is organized hierarchically, starting with an overall theme (e.g., “Structure,” “Cycles,” or “Utah Natural History”) and becoming more specific at each successive level. A description of each level follows.

- **Intended Learning Outcomes (ILOs)** are a broad description of the science skills and attitudes that students are expected to have by the end of the grade or course. The ILOs encourage students to develop curiosity about the natural world, the skills and understanding necessary to do science, the ability to communicate effectively, awareness of the social and historical implications of science, and an understanding of the nature of scientific methods and thought.

- **Science benchmarks** provide a broad description of a particular subset of science content that students are required to know. For example, an elementary benchmark concerns the nature of forces, motion, and gravity.

- **Standards** break down the benchmarks into more specific concepts, such as the relationship between the force applied to an object and the object’s resulting motion. Each benchmark will have at least one standard.

- **Objectives** focus the content even further, and all of a standard’s objectives must be mastered in order for a student to be considered proficient at that standard. An example of an objective is understanding that a greater force results in a greater change in motion.

- **Indicators** are specific and measurable actions that communicate a student’s understanding of an objective. For example, the actions can be “demonstrate,” “predict,” “observe and record,” and “identify.”

The UCSS addresses each grade individually, instead of by grade band. This allows item writers to better match the grade level appropriateness of items, and item alignment will be more focused. This, combined with the broad requirements of the ILOs, will lead to assessments that match the UCSS vision of progressive increase in the complexity of student tasks and content knowledge.
Sample Items and Item Types

AIR’s robust and flexible technology puts the ability to create a wide variety of item types in the hands of item developers. While it is typical to discuss items in terms of “item types,” our success has derived from a broader, more generalized understanding. Instead of thinking about item types, our systems are built around a broad and growing collection of “response mechanisms.” In addition to multiple-choice items, our system currently supports 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 within a passage
- Equation response, in which students respond by entering an equation
- Proposition response, in which students respond in one or more English language sentences, which may be scored by our proposition scoring engine, human-scored, or a mixture
- Essay response, in which the student response is a longer written response to be scored using an essay-scoring engine or a mix of machine and human scoring
- Simulator interactions, in which students interact with a simulator, and their sequence of choices and interactions can be scored according to an explicit rubric
- Spoken response, which captures student speech for human scoring

In addition, we are currently developing a flexible open-ended response mechanism that we call “scratch-pad response,” which will enable students to integrate text, free-hand drawing, and equations to respond to items. Currently, we plan to score scratch-pad responses by hand.

Our graphic response mechanism supports most of the typical “technology-enhanced” item types, including sorting, matching, hot-spot, drag-and-drop, etc. In addition it supports items where students actually draw a machine-scorable response; respond by constructing complex, open-ended diagrams; and many more possibilities. Uniformly deriving from a single response mechanism means that 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, 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 equations using a palette of symbols. Test developers can specify which symbols are available on an item-by-item basis, or USOE can choose to have the palette remain consistent across all of the items within a grade level.

The proposition response mechanism and essay response mechanism are very similar to short constructed-response and extended constructed-response items in which the students type their answers into a space provided. The response mechanisms differ by the length of response accepted (both configurable by item) but, more importantly, the scoring engine used to score the response.

Our system includes a very powerful and flexible mechanism for test developers to define very interactive “simulation” tasks. The test developers can choose from many standard web input widgets and can define output that
can (a) drive an animation and (b) populate a table. These interactive “simulations” can be offered as a mechanism for students to explore or experiment with the simulated phenomenon. If USOE chooses, student interactions can be captured and scored—for example, evaluating the quality of an experiment represented by a sequence of trials.

Our system is capable of capturing speech and routing it to our scoring system. We have not included such items in this proposal due to the logistical requirements, which may be prohibitive. Specifically, each student would need headphones with a microphone. Testing would have to be individual or small-group in order to minimize the noise and distraction in the room (imagine 30 students all speaking at the same time in the testing setting). However, we mention that we have this capability in operational use, and we are happy to negotiate the inclusion of items that use it if Utah finds the capability attractive.

As mentioned above, our forthcoming scratch-pad response provides students the tools necessary to express mathematic reasoning in much the same way that they might on a paper-and-pencil test. The scratch-pad response mechanism is scheduled for release in January 2013. The very flexible and open-ended nature of this response mechanism dictates that responses using it be human-scored, at least in the short term.

The availability of tools organized around response mechanisms creates a very flexible capability for test developers to create authentic, challenging tasks. The sample items below, organized by subject area, illustrate the flexibility that these tools offer test developers to create authentic, challenging tasks that probe the full range of depth of knowledge.

Readers are encouraged to review the samples from all subjects. Differences across the samples help illustrate the flexibility and range of capabilities of the system.
**Mathematics sample items**

Below, we present four sample tasks:

- A multiple-choice item targeting DoK 2
- A graphic response item, also targeting DoK 2
- A constructed-response item requiring the student to construct an equation to formalize a mathematical model from a verbal description
- A performance task that includes five items (including an extended-response item) that targets DoK 3 and 4 and requires some human scoring of the final response.
<table>
<thead>
<tr>
<th>Item Type</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment (i.e., Standards)</td>
<td>5.NBT.4. Use place value understanding to round decimals to any place.</td>
</tr>
<tr>
<td>DoK</td>
<td>2</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>Students apply their understanding of place value to interpret two given conditions and identify a number that satisfies both conditions. This is a DoK2 item because it’s not a straightforward “round this number.” Students must apply their knowledge of rounding and number sense to successfully solve this problem.</td>
</tr>
</tbody>
</table>
Felicia is thinking of a number.

- Rounded to the nearest tenth, this number rounds up.
- Rounded to the nearest hundredth, this number does not change.

Which could be Felicia’s number?

(A) 6.705
(B) 7.685
(C) 8.590
(D) 9.544
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Graphic Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Hot Spot</td>
</tr>
<tr>
<td>Alignment (i.e., Standards)</td>
<td>3.MD.3. Draw a scaled picture graph and a scaled bar graph to represent a data set with several categories. Solve one- and two-step &quot;how many more&quot; and &quot;how many less&quot; problems using information presented in scaled bar graphs. For example, draw a bar graph in which each square in the bar graph might represent 5 pets.</td>
</tr>
<tr>
<td>DoK</td>
<td>2</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>A typical DoK 1 item for this standard might ask students to simply identify the bar graph that matches a table of given data. A DoK 2 item like this requires more than basic recall and reproduction. In this case, students must interpret the context, establish relationships among given information, and construct their own bar graph. Further, there is no single solution. This open-ended item allows students to demonstrate their levels of understanding in ways that go beyond one particular solution.</td>
</tr>
</tbody>
</table>
Sam, Paul, and Renee each set up lemonade stands and recorded how much money they made in one day.

- Sam made $15 more than Paul.
- Renee made $5 less than Sam.

Click above each name to create a bar graph that shows how much money each student could have made in one day.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Equation Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Constructed Response</td>
</tr>
</tbody>
</table>
| Alignment (i.e., Standards) | 4.OA.1. Interpret a multiplication equation as a comparison, e.g., interpret $35 = 5 \times 7$ as a statement that 35 is 5 times as many as 7 and 7 times as many as 5. Represent verbal statements of multiplicative comparisons as multiplication equations.  
4.OA.2. Multiply or divide to solve word problems involving multiplicative comparison, e.g., by using drawings and equations with a symbol for the unknown number to represent the problem, distinguishing multiplicative comparison from additive comparison. |
| DoK                 | 2 |
| Rationale/Description | Students interpret situations involving multiplicative comparison, an often difficult concept for fourth graders. Students create an equation to represent the multiplicative relationship and offer one possible set of solutions. In a constructed-response item, students go beyond identifying the solution—they create their own. Further, this open-ended item allows for multiple equations and an infinite number of solutions. |
William has 3 times as many baseballs as Jim.

Use the equation that shows this relationship. Use \( w \) for William and \( j \) for Jim.

**Click the buttons to create your answer.**

Operations

Other Symbols

Fractions

Use the keypad to show how many baseballs William could NOT have.

```
0 1 2
3 4 5
6 7 8
```
Sample Math Performance Task

The task that follows is a sample mathematics performance task. We estimate that it will require approximately 20–40 minutes for a student to complete this task. Questions 1–3, the equation response items, require students to interpret data presented in the form of a graph. From this graph, students must recognize the function as a cosine function, as well as the various components of the cosine function modeled by the graph. Students go beyond basic recall of mathematical ideas. They construct their own function by interpreting and analyzing the key pieces of the graph, namely, the height of the axle, the diameter of the Ferris wheel, and the number of rotations per minute. These three key pieces of information are not explicitly stated by the graph; rather, students must determine them from the graph and then construct the function modeled by the graph. Thus, these are examples of DoK 3 items.

The final two items in the task represents DoK 4. Students apply previously learned concepts in a novel context, they analyze the given information, and they create their own profit model given the constraints in the task. Five of the Standards for Mathematical Practices are addressed in the performance task:

1. **Make sense of problems and persevere in solving them.** Students must make sense of the cosine function graph and interpret the information presented graphically and use this information to solve problems.

2. **Reason abstractly and quantitatively.** Students use the graph and information to construct their own functions (abstract) and determine key values (quantitative) provided by the stem.

3. **Model with mathematics.** By building a new function, students create a model for a different number of rotations per minute.

4. **Use appropriate tools strategically.** Students analyze graphs of functions and solutions generated by the simulator.

5. **Attend to precision.** Students must build their functions using exact values and conventional notation for the functions.
<table>
<thead>
<tr>
<th>Item Type</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td>9-12.F.BF.1.1 Write a function that describes a relationship between two quantities</td>
</tr>
<tr>
<td>(i.e., Standards)</td>
<td>9-12.F.TF.2.5 Choose trigonometric functions to model periodic phenomena with specified amplitude, frequency, and midline.</td>
</tr>
<tr>
<td>DoK</td>
<td>3</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>This item requires students to interpret data presented in the form of a graph. From this graph, students must recognize the function as a cosine function, as well as the various components of the cosine function modeled by the graph. Students go beyond basic recall of mathematical ideas. They construct their own function by interpreting and analyzing the key pieces of the graph, namely, the height of the axle, the diameter of the Ferris wheel, and the number of rotations per minute. These three key pieces of information are not explicitly stated by the graph; rather, students must determine them from the graph and then construct the function modeled by the graph. Thus, this is an example of a DOK3 item.</td>
</tr>
</tbody>
</table>
A local amusement park offers Ferris wheel rides to children and adults. A rider is shown on the Ferris wheel.

1. Write a function \( h(t) \) that models the height \( h \), of the rider at any point on the Ferris wheel in terms of \( t \), time in minutes.

Click the buttons to create your answer.

Functions \( h(t) \) Do I ( 

Operations \(+ - \times \div \) \( h(t) \) Do O

Signs \(< > \leq \geq \) \( h(t) \) Do O

Trigonometry \( \sin \cos \tan \arcsin \arccos \arctan \)

Press the Play button to begin the Ferris wheel ride. The height of the rider at any point in the ride will be shown on the graph.
<table>
<thead>
<tr>
<th>Item Type</th>
<th>Keyboard</th>
</tr>
</thead>
</table>
| Alignment (i.e., Standards) | 9-12.F.BF.1.1 Write a function that describes a relationship between two quantities  
                        9-12.F.TF.2.5 Choose trigonometric functions to model periodic phenomena with specified amplitude, frequency, and midline. |
| DoK               | 3                         |
| Rationale/Description | This item requires students to use the function they created in question 1 in order to answer a related question. Students assess the meaning of one key component of their function and use it to calculate the number of rotations per minute. In doing so, they use concepts to solve non-routine problems. Thus, this is an example of a DOK3 item. |
A local amusement park offers Ferris wheel rides to children and adults. A rider is shown on the Ferris wheel.

Press the Play button to begin the Ferris wheel ride. The height of the rider at any point in the ride will be shown on the graph.

1. Use the keypad to show the number of rotations per minute for the function you wrote in Question 1.

```
1 2 3 4
5 6 7 8
9 0 /
Delete
```

2. Each ride consists of 15 rotations of the Ferris wheel. Use the keypad to show the duration, in minutes, of each ride based on the function you wrote in Question 1.

```
1 2 3 4
5 6 7 8
9 0 /
Delete
```

3. The managers of the amusement park want to change the number of rotations per minute to have more rides per hour. They determine the maximum safe speed is 4 rotations per minute.

4. Write a function that models the height of the rider at any point on the Ferris wheel, based on a different number of rotations per minute.
### Sample Performance Task: Mathematics

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Keyboard</th>
</tr>
</thead>
</table>
| Alignment (i.e., Standards) | 9-12.F.BF.1.1 Write a function that describes a relationship between two quantities  
9-12.F.TF.2.5 Choose trigonometric functions to model periodic phenomena with specified amplitude, frequency, and midline. |
| DoK | 3 |
| Rationale/Description | This item builds on the responses to questions 1 and 2 in order to answer another related question. Students draw conclusions about the graph and their function, as well as the number of rotations per minute in order to determine the length of the ride. Thus, this is an example of a DOK3 item. |
A local amusement park offers Ferris wheel rides to children and adults. A rider is shown on the Ferris wheel.

1. Each ride consists of 15 rotations of the Ferris wheel. Use the keypad to show the duration, in minutes, of each ride based on the function you wrote in Question 1.

2. Calculate the diameter of the Ferris wheel. The diameter is the distance across the wheel at any point in the ride.

The managers of the amusement park want to change the number of rotations per minute to have more rides per hour. They determine the maximum safe speed is 4 rotations per minute.

3. Each ride consists of 15 rotations of the Ferris wheel. Use the keypad to show the duration, in minutes, of each ride based on the function you wrote in Question 1.

4. Write a function that models the height of the rider at any point on the Ferris wheel, based on a different number of rotations per minute. Click the buttons to create your answer.

5. The height of the rider at any point in the ride will be shown on the graph.

6. Use the keypad to show the duration, in minutes, of each ride based on the function you wrote in Question 1.
Sample Performance Task: Mathematics

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment (i.e., Standards)</td>
<td>9-12.F.BF.2.3 Identify the effect on the graph of replacing $f(x)$ by $f(x) + k$, $k f(x)$, $f(kx)$, and $f(x + k)$ for specific values of $k$ (both positive and negative); find the value of $k$, given the graphs. Experiment with cases and illustrate an explanation of the effects on the graph using technology.</td>
</tr>
<tr>
<td>DoK</td>
<td>3</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>After students have constructed a function which models the given animation and resulting graph, they construct a new function based on a different number of rotations per minute. In doing so, they investigate possible scenarios within the context of a safe Ferris wheel ride. Thus, this is an example of a DOK3 item.</td>
</tr>
</tbody>
</table>
A local amusement park offers Ferris wheel rides to children and adults. A rider is shown on the Ferris wheel.

The managers of the amusement park want to change the number of rotations per minute to have more rides per hour. They determine the maximum safe speed is 4 rotations per minute.

4. Write a function that models the height of the rider at any point on the Ferris wheel, based on a different number of rotations per minute.

Click the buttons to create your answer.

Functions: \(6 \ 0 \ 0\)
Operations: \(E \ J \ [Jt]\)
Signs: \([j@] \ [J] \ [EJ] \ [0] \ [i]\)
Trigonometry: \(\sin \ \cos \ \arcsin \ \arccos \ \arctan\)

1 2 3
4 5 6
7 8 9
0

43187

The Ferris wheel has 40 seats, and each seat yields a profit of $5.00.

5. If the number of rotations per minute increases, as completed in Question 4, how will the number of rides per hour change? Explain how you arrived at your solution.

Type your answer in the space provided.
<table>
<thead>
<tr>
<th>Item Type</th>
<th>Constructed Response</th>
</tr>
</thead>
</table>
| Alignment (i.e., Standards) | 9-12.F.BF.2.3 Identify the effect on the graph of replacing $f(x)$ by $f(x) + k$, $kf(x)$, and $f(kx)$ for specific values of $k$ (both positive and negative); find the value of $k$, given the graphs. Experiment with cases and illustrate an explanation of the effects on the graph using technology.  
9-12.A.CED.1.3 Represent constraints by equations or inequalities, and by systems of equations and/or inequalities, and interpret solutions as viable or nonviable options in a modeling context. |
| DoK                    | 4                    |
| Rationale/Description  | Students apply previously learned concepts in a novel context, they synthesize the given information, and they create their own profit model given the constraints in the task. In addition, they must justify their reasoning and provide evidence for their responses. Thus, this two-part question is an example of a DOK4 item. |
A local amusement park offers Ferris wheel rides to children and adults. A rider is shown on the Ferris wheel.

Press the Play button to begin the Ferris wheel ride. The height of the rider at any point in the ride will be shown on the graph.

The Ferris wheel has 40 seats, and each seat yields a profit of $5.00.

5. If the number of rotations per minute increases, as completed in Question 4, how will the number of rides per hour change? Explain how you arrived at your solution.

Type your answer in the space provided.

6. How does the profit per hour change? Explain how you arrived at your solution.

Type your answer in the space provided.
Science sample items

Below, we present four sample science tasks. The first three are stand-alone items, and the last one is a completely machine-scored performance task:

- A technology-enhanced, selected-response item designed to probe students’ knowledge of cell structures and functions and have them apply their knowledge by recognizing the structure in a photograph. We note that this could be accomplished with a traditional multiple-choice selected-response item, but the response format would increase the cognitive load unnecessarily. This item targets DoK 2.

- A technology-enhanced, machine-scored constructed-response item in which the student is asked to create an illustration of physical forces at work and predict their effect on a physical object. This item targets DoK 3.

- A technology-enhanced drag-and-drop item that asks the student to construct an experiment that will determine the conductive properties of unknown materials. This item targets DoK 3.

- A fully machine-scored performance task asks the student to conduct an investigation that leads to the design of a tool to detect the presence of a particular substance. This targets DoK 4.
Response mechanism | Graphic Response
---|---
Item Type | Technology-enhanced and Selected Response
Alignment (i.e., Standards) | Grade 7  
Grade Theme: Structure  
Intended Learning Outcome 3a: Know and explain science information specified for their grade level.  
Standard 3: Students will understand that the organs in an organism are made of cells that have structures and perform specific life functions.  
Objective 1: Observe and describe cellular structures and functions.  
Indicator e: Gather information to report on how the basic functions of organisms are carried out within cells (e.g., extract energy from food, remove waste, produce their own food).
DoK | 2
Rationale/Description | The student is required to recall that RNA is produced in the nucleus and be able to recognize the nucleus in a microscope photograph of an onion cell. This item is classified as DoK 2 because the student must integrate two pieces of knowledge (visual recognition and the function of different cell structures) to be able to answer the question. The student must be able to apply the academic understanding of the structure’s function at the same time as he or she recognizes the structure in the cell.

The interactions on this item show how technology can make even selected-response items clearer and reduce construct-irrelevant variance. As the student mouses over different structures in the cells, the structure to be selected is highlighted and the rest of the cell fades. Clicking selects the structure. Exhibit II.A.3-1 shows what the student sees when he or she mouses over the nucleus. A similar visual approach is used for the other visual structures and parts of the cells.
Exhibit II.A.3-1: Change in image when student mouses over nucleus.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Graphic Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Technology-enhanced Constructed Response</td>
</tr>
</tbody>
</table>
| Alignment (i.e., Standards) | **High School Physics**  
Theme: Forces acting on objects  
Intended Learning Outcomes: 3c — Apply principles and concepts of science to explain various phenomena.  
3d — Solve problems by applying science principles and procedures.  
Standard 3: Students will understand the factors determining the strength of gravitational and electric forces.  
Objective 2: Describe the factors that affect the electric force (i.e., Coulomb's law).  
Indicator a: Relate the types of charge to their effect on electric force (i.e., Coulomb's law).  
Indicator b: Describe how the amount of charge affects the force.  
Indicator c: Investigate the relationship of distance between charged objects and the strength of the electric force. |
| DoK | 3 |
| Rationale/Description | The student must perform a multi-step process to determine the direction of the force. First, the student must recall that the electric force is proportional to the amount of charge, that opposite charges attract and like charges repel, and that the strength of the force decreases with increasing distance. Second, he or she needs to use the properties of vector addition, along with the symmetry of the charge arrangement, to determine that the direction of the force on each charge will be toward the center of the grid. Finally, Part B requires the student to understand that only at the center of the grid will the combined forces of the charges be zero, which again requires understanding of the nature of electric force and vector addition. Because the student is engaged in mental processing beyond just recalling information, and he or she must apply knowledge of physical forces to predict where the red sphere would reach equilibrium and explain this by identifying the directional forces, this item reaches DoK 3. |
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.

B. Move the red charged sphere to the point on the table where there will be no net force from the other four spheres.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Graphic Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Technology-enhanced Drag-and-Drop</td>
</tr>
</tbody>
</table>
| Alignment (i.e., Standards) | Grade 5  
Theme: Change and Cause and Effect  
Intended Learning Outcomes: 1f – Plan and conduct simple experiments.  
6c – Science findings are based upon evidence.  
Standard 4: Students will understand features of static and current electricity.  
Objective 2: Analyze the behavior of current electricity  
Indicator e: Make a working model of a complete circuit using a power source, switch, bell or light, and a conductor for a pathway. |
| DoK                | 2 |
| Rationale/Description | The student must understand how to design a complete circuit with a power source and that the brightness of the bulb will depend on the conductivity of the unknown material in the circuit. Furthermore, he or she must properly control the other aspects of the circuit, such as the number of batteries and bulbs. Because the palette bar allows the student to drag in multiple copies of a particular object to the answer space, the student is not restricted in his or her circuit design. This forces the student to think carefully about how to proceed, and reduces the chances of receiving credit for blindly guessing.  
This item is assigned a DoK 3 because the student must formulate a plan about how to design a controlled experiment in an unrestricted space that has many correct answers.  
This item mimics the cognitive demand of performing an actual experiment in a laboratory setting and assesses a deep understanding of science principles. |
Your teacher gives you two unknown materials and tells you that one material conducts electricity very well and the other does not.

Using the two circuits, design an experiment that will allow you to see which material is the better conductor.

Place an object in each box in the circuit diagram to complete the circuit.
**Sample Science Performance Task**

In sample task that follows, the student must design a capsule that, if dropped in a solution, can detect the presence of a certain molecule. He or she is provided with a simulated laboratory environment in which to perform tests on sample capsule membranes. These tests will provide information about the properties of the membranes and the unknown molecules. However, because there are many possible input combinations, the student must systematically plan an experimental procedure—blindly choosing inputs will rarely provide useful information. He or she must carefully plan which experiments to run, while avoiding experiments that are redundant or yield no information and while keeping the final goal in mind.

A correct solution requires the synthesis of knowledge about diffusion across a semipermeable membrane, an understanding of experimental methods, and the higher-order thinking involved in formulating a strategy.

The simulation item type allows for scoring based on data that appear in the output table. In Exhibit II.A.3-2, the student could be scored on the efficiency of his or her experimental method or on the number of trials that produce useful information. This functionality enables item writers to assess the problem-solving process as well as content knowledge. The Utah Core Science Standards emphasize that science is not only a body of knowledge but also a way of thinking and a set of skills, and AIR’s simulation technology provides a way to directly measure science process skills.

This performance task requires the student to design a solution to a novel problem. It provides a simulation environment for conducting a strategic multi-step experiment, and the student must combine content understanding with proficiency in experimental methods. Therefore, it has been assigned a DoK of 4.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Graphic Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Technology-enhanced Hotspot</td>
</tr>
<tr>
<td>Alignment (i.e., Standards)</td>
<td>Theme: Structure</td>
</tr>
<tr>
<td></td>
<td>Intended Learning Outcomes:</td>
</tr>
<tr>
<td></td>
<td>1a – Observe objects and events for patterns and record both qualitative and quantitative information.</td>
</tr>
<tr>
<td></td>
<td>1e – When given a problem, plan and conduct experiments.</td>
</tr>
<tr>
<td></td>
<td>3d – Solve problems appropriate to grade level by applying scientific principles and procedures.</td>
</tr>
<tr>
<td></td>
<td>3a – Know and explain science information specified for their grade level.</td>
</tr>
<tr>
<td>DoK</td>
<td>4</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>This item requires the student to design a solution to a novel problem. It provides a simulation environment for conducting a strategic multi-step experiment, and the student must combine content understanding with proficiency in experimental methods. Therefore, it has been assigned a Depth of Knowledge of 4.</td>
</tr>
</tbody>
</table>
You are working for a spy agency and must design a capsule that can detect Molecule X in a solution. The capsule will have a semipermeable membrane and will be filled with a solution.

You are given the following information:
1. If Molecules X and Y are both in a solution, it will turn blue.
2. If Molecule Z is present along with X and Y, it will neutralize the reaction and the solution will not turn blue.
3. Molecules X, Y, and Z all dissolve in water.

You must find the right membrane and fill it with the appropriate solution. A laboratory is provided with three different capsule membranes and samples of Molecules X, Y, and Z. You can perform experiments to determine the properties of the capsule membranes and the molecules.

Select a capsule membrane, capsule solution, and beaker solution. The capsule will be submerged in the beaker solution, and the results will be recorded in the table.

Choose Capsule Membrane
- Membrane A
- Membrane B
- Membrane C

Choose Capsule Solution
- 10% Molecule X
- 10% Molecule Y
- 10% Molecule Z

Choose Beaker Solution
- 10% Molecule X
- 10% Molecule Y
- 10% Molecule Z

Which two conditions are required for a molecule to diffuse across a semipermeable membrane?

Type your answer in the space provided.
### Sample Performance Task: Science

<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Graphic Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Natural Language</td>
</tr>
</tbody>
</table>
| Alignment (i.e., Standards) | Grade 7  
Theme: Structure  
Intended Learning Outcomes:  
3a – Know and explain science information specified for their grade level.  
Standard 3: Students will understand that the organs in an organism are made of cells that have structures and perform specific life functions.  
Objective 1: Observe and describe cellular structures and functions.  
Indicator d: Model the cell processes of diffusion and osmosis and relate this motion to the motion of particles. |
| DoK                | 1 |
| Rationale/Description | This item requires the student to understand the conditions under which diffusion occurs. Because this only requires recalling information, it is assigned a DoK of 1. |
You are working for a spy agency and must design a capsule that can detect Molecule X in a solution. The capsule will have a semipermeable membrane and will be filled with a solution.

You are given the following information:

- If Molecules X and Y are both in a solution, it will turn blue.
- If Molecule Z is present along with X and Y, it will neutralize the reaction and the solution will not turn blue.
- Molecules X, Y, and Z all dissolve in water.

You must find the right membrane and fill it with the appropriate solution. A laboratory is provided with three different capsule membranes and samples of Molecules X, Y, and Z. You can perform experiments to determine the properties of the capsule membranes and the molecules.

Select a capsule membrane, capsule solution, and beaker solution. The capsule will be submerged in the beaker solution, and the results will be recorded in the table.

Choose Capsule Membrane
- Membrane A
- Membrane B
- Membrane C

Choose Capsule Solution
- 10% Molecule X
- 10% Molecule Y
- 10% Molecule Z

Choose Beaker Solution
- 10% Molecule X
- 10% Molecule Y
- 10% Molecule Z

Which two conditions are required for a molecule to diffuse across a semipermeable membrane?

Type your answer in the space provided.

Arrange the molecules in order of size.

Click and drag the molecules into the answer boxes.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Graphic Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Technology-enhanced Drag-and-Drop</td>
</tr>
</tbody>
</table>
| Alignment (i.e., Standards) | Grade 7  
Theme: Structure  
Intended Learning Outcomes:  
1a – Observe objects and events for patterns and record both qualitative and quantitative information  
3a – Know and explain science information specified for their grade level.  
3d – Solve problems appropriate to grade level by applying scientific principles and procedures.  
Standard 1: Students will understand the structure of matter.  
Objective 3: Investigate the motion of particles.  
Indicator c: Design and conduct an experiment investigating the diffusion of particles. |
| DoK                | 3 |
| Rationale/Description | This item requires the student to use information gathered from the simulation to determine the relative sizes of the unknown molecules. The student must understand the conditions under which diffusion occurs and must have carefully set up the individual trials of the experiment to gather useful information. |
You are working for a spy agency and must design a capsule that can detect Molecule X in a solution. The capsule will have a semipermeable membrane and will be filled with a solution.

You are given the following information:
- If Molecules X and Y are both in a solution, it will turn blue.
- If Molecule Z is present along with X and Y, it will neutralize the reaction and the
- Molecules X, Y, and Z all dissolve in water.

You must find the right membrane and fill it with the appropriate solution. A laboratory is provided with three different capsule membranes and samples of Molecules X, Y, and Z. You can perform experiments to determine the properties of the capsule membranes and the molecules.

Arrange the molecules in order of size. Click and drag the molecules into the answer boxes.
ELA Sample Items and Performance Task

Below, we present sample items in the context of a sample performance task. This sample performance task includes a variety of item types and serves to illustrate both the task and the variety of item types.

In this performance task, the student is provided with three documents:

- An online article on compulsory voting
- A blog post about a court case involving compulsory voting in Australia
- A simulated website offering statistical information about voting rates by demographic groups

In the final extended-response item, the student is asked to analyze a quote and relate it to the arguments presented in each document. In the process of working through the performance task, the student responds to

- a machine-scored constructed-response item (proposition response), which asks the student to critique a statement from the perspective of one of the authors. This item targets DoK 3.
- a technology-enhanced “hot text” item in which the student selects sentences within the passage as part of a critique of the author’s argument. This item targets DoK 3.
- a selected-response item in which the student is asked to draw a conclusion based on data presented in a text. This item targets DoK 2.
- an extended-response item in which the student must relate a particular quote to the arguments offered by each author.

We note that the scoring rubric of the extended-response item is critical. This is designed to measure reading standards, rather than writing standards. Therefore, this item’s rubric is based exclusively on whether the student fulfilled the Reading standard by

- explaining the meaning of the quotation in the context of the passages;
- applying the quotation to each passage;
- evaluating how each passage supports, or refutes, the quotation; and
- connecting the ideas across passages to make an overall judgment.

Recall from the section on scoring (Section 1.C.3.d) that we plan to score all responses to these extended-response items using our essay scoring engine. However, we will incorporate a diagnostic procedure that detects potential validity issues and routes suspect papers to human scorers. We expect that approximately 20–25% of papers may be flagged for a human review. The full reading for this performance task is provided in Appendix D.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Hot Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Hot Text</td>
</tr>
<tr>
<td>Alignment to Utah Core Standards for ELA</td>
<td>Reading Standard 5. Analyze and evaluate the effectiveness of the structure an author uses in his or her exposition or argument, including whether the structure makes points clear, convincing, and engaging.</td>
</tr>
<tr>
<td>DoK</td>
<td>3</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>This item requires students to evaluate an author’s argument and critique his rhetorical approach. Students must both understand the author’s overall argument as well as identify fallacies in the logic.</td>
</tr>
</tbody>
</table>
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.

The leading argument against compulsory voting is that it is not consistent with the freedom associated with democracy. Voting is not an intrinsic obligation. The enforcement of the law would undermine 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 citizenship—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.

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 citizenship—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.

**Passage #2** states: “Advocates of compulsory voting argue that decisions made by democratically elected governments are more legitimate when higher proportions of the population participate.”

Explain how the author of Passage #2 would counter the claim that election results are “more legitimate.” Type your response in the space below.

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 citizenship—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.

Explain why the sentence you highlighted undermines the author's own argument.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Multiple Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Selected Response</td>
</tr>
<tr>
<td>Alignment to Utah Core Standards for ELA</td>
<td>Reading Standard 7. Integrate and evaluate multiple sources of information presented in different media or formats (e.g., visually, quantitatively) as well as in words in order to address a question or solve a problem.</td>
</tr>
<tr>
<td>DoK</td>
<td>2</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>This item requires students to make an inference based on a text. In this item students analyze data presented in a text and use the information to draw a conclusion.</td>
</tr>
</tbody>
</table>
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 generally 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.

The leading argument against compulsory voting is that it is not consistent with the freedom associated with democracy. Voting is not an intrinsic obligation and the enforcement of the law would be an infringement of the right to freedom of action in an autonomous state.
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Multiple Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Selected Response</td>
</tr>
<tr>
<td>Alignment to Utah Core Standards for ELA</td>
<td>Reading Standard 7. Integrate and evaluate multiple sources of information presented in different media or formats (e.g., visually, quantitatively) as well as in words in order to address a question or solve a problem.</td>
</tr>
<tr>
<td>DoK</td>
<td>2</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>This item requires students to make an inference based on a text. In this item students analyze data presented in a text and use the information to draw a conclusion.</td>
</tr>
</tbody>
</table>
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.

The leading argument against compulsory voting is that it is not consistent with the freedom associated with democracy. Voting is not an intrinsic obligation and the enforcement of the law would be...

---

**43175**

Explain why the sentence you highlighted undermines the author's own argument.

Type your answer in the space provided.

**43178**

Which graph could you use to demonstrate that voters have become more engaged in the electoral process?

(i) "Trends in Voting"

(ii) "Age"

(iii) "Sex"

(iv) "Education"

**43180**

Read the following quote by political analyst Larry J. Sabato: "Every election is determined by the people who show up."

Analyze the meaning of this quote and explain how it applies to each passage provided. Connect the ideas and information presented in the passages to show how they either support, or refute, this statement.

---
<table>
<thead>
<tr>
<th>Response mechanism</th>
<th>Essay Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type</td>
<td>Extended Response</td>
</tr>
<tr>
<td>Alignment to Utah Core Standards for ELA</td>
<td>Reading Standard 7. Integrate and evaluate multiple sources of information presented in different media or formats (e.g., visually, quantitatively) as well as in words in order to address a question or solve a problem.</td>
</tr>
<tr>
<td>DoK</td>
<td>4</td>
</tr>
<tr>
<td>Rationale/Description</td>
<td>This item requires students to gather, analyze, organize, and synthesize information from multiple sources to draft a reasoned response. Students must go beyond a single analysis and consider multiple perspectives on an issue. Students are required to apply higher-order thinking to analyze a stance within the context of their reading and evaluate how that stance can be applied across multiple texts.</td>
</tr>
</tbody>
</table>
DO

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.

The leading argument against compulsory voting is that it is not consistent with the freedom associated with democracy. Voting is not an intrinsic obligation and the enforcement of the law would be an infringement of the right to freedom of association.
Conclusion

The foregoing examples provide a glimpse of some of the types of items that our technology supports. Other examples will become available on the Smarter Balanced website (SmarterBalanced.org) on October 8. Those items are stored in our item bank and presented and scored using AIR technology. We encourage readers to sample those items on a supported tablet as well.

Readers are invited to interact with the sample items offered here and other example items at https://air.tps.airast/utah.

username:  password:

The samples will be available on October 8. Given the ongoing competition, we ask that this URL, username, and password be treated with the same confidentiality as our proposal.
II.B.2. Item Development/Procurement

Overview

AIR proposes that the interim and summative assessments be virtually identical—using the same system, blueprints, and items drawn from the same bank. We propose to model the system on the system we are implementing for. In, the full item pool has been partitioned into an interim and a summative pool, and identical adaptive tests (with different items) are administered in separate windows.

We recommend that USOE offer two opportunities for the interim assessments and one opportunity for the summative assessment. This configuration is a recommendation, but is not the only one supported by this proposal. For example, offers three opportunities. They can be given at any time and track progress over time. In this regard they are interim. The highest score a student achieves, regardless of test date, is used for accountability.

AIR will collaborate with Utah educators and USOE to craft a model that works well for Utah.

We do not recommend offering more than three interim/summative opportunities. In general, a good test, even a good adaptive test, has a standard error of measurement of between one-half and one-full year’s growth. When tests are administered more than once about every 8–10 weeks, the measurement error swamps the individual changes.

We remind readers that the Utah system will also include a powerful and flexible formative assessment system that will help educators track learning over time. The system, described in the final section of this technical proposal, is designed to encourage educator-student interaction, as well as the leveraging of data to make instructional choices.

Below, we describe the item development and procurement process, including the procedures that we will use to ensure alignment of the items to the standards.

Item Development and Procurement

This section details our plans for creating or obtaining items for the adaptive assessments. As we describe below, we will obtain items from multiple sources and custom-develop many more items. We describe in detail our process for ensuring alignment to the Utah Core Standards. The final step in our proposed alignment process will yield an independent alignment review, which will meet the USOE standards for peer review.

The RFP includes Appendix A in this section, and we address the requirements from that appendix here.

Sources of Items

The RFP identifies four potential sources of items:

1. Existing vendor-owned items
2. Existing items from Utah’s current CRT
3. Open-source items from the ARRA-funded assessment consortia
4. Items newly developed for this contract

AIR does not currently offer off-the-shelf assessments, and therefore we have no existing vendor-owned items. However, we have been providing technology-enhanced adaptive tests in four states, three of which have entered into a series of bilateral item sharing
arrangements: , , and . This agreement allows each state to select up to 40 items per year from each other state to use securely. Two of these states, and , have invited Utah to enter into a similar agreement (’s leadership is currently in transition). Both and are developing math and ELA items for the common core, and therefore items are being directly developed to measure the same core standards that form the basis of the Utah Core Standards.

AIR expects that approximately 50 percent of Utah’s existing items will align to the common core. We recognize that this proportion will vary by grade and subject and will probably range from about 30 percent to about 70 percent. We also recognize that this 50 percent of existing items will over-represent traditional multiple-choice items and be heavily weighted toward DoK 1 and 2. We propose to identify the aligned items and import them into our Item Tracking System (ITS).

In producing this proposal, AIR contacted both consortia to solicit open-source items, or to understand any licensing approach that they may take in the future. The consortium grants require that the grant-funded items become freely available, but allow for restrictions based on security concerns. Neither consortium offered use of its items during the first two years of this contract. Apparently both consortia are still formulating their plan to comply with this provision of the grant. In essence, the availability of such items is uncertain.

With regard to items available from the consortia, we will be able to import and deliver these items if they are, in fact, in QTI/APIP format. If and when these items become available, we will negotiate a price to import, prepare, and deliver these items for Utah. We note that the 10,000 Smarter Balanced items used in its pilot test will exist in our Item Tracking System and will be delivered using our test delivery system. If licenses can be negotiated, the cost of incorporating these items in Utah’s test will be minimal.

Finally, AIR will develop custom items for the Utah assessment. In sections that follow we describe how we will evaluate the alignment of all items—those that are procured and those that are developed. We describe in detail our item development process and the methods that we will use to ensure that every item meets the standards outlined in Appendix A of the RFP.

For the initial item pool, we propose to deploy 400 items per test and expect tests that are between 45 and 55 items each (including the performance tasks). As discussed in the section on Psychometrics, this should be sufficient for a robust, adaptive interim and summative assessment system offering a total of three opportunities per test. A student using all three opportunities would see approximately 135–165 items across all opportunities, leaving plenty of room for adaptation. We propose to refresh the pool with 50 items per test each year, which will leave Utah on the path to continuous improvement in the measurement characteristics of the test and provide an opportunity to release more items over time.

Exhibit II.B.2-1 As described in our more detailed development plans, new items will be largely constructed-response and technology-enhanced items. The new items will include performance tasks and emphasize greater depth of knowledge.

Please note that the numbers in Exhibit II.B.2-1 are counts of items that will make it to the field-test pool. We typically lose very few items at field test. The development process, however, sees higher attrition. We are prepared to develop up to 50 percent more items than the numbers that appear in Exhibit II.B.2-1 in order to reach those numbers on the field-test forms.
Process for Evaluating the Alignment of Items

As described in the introduction to this section, the new Utah Core Standards, like the Common Core Standards from which ELA and math derive, require a new approach to understanding alignment. The test should measure the standards in a way that drives instruction in the right direction. Dividing skills into small discrete pieces helps design curriculum and communicate what students should know and know how to do; however, most real application requires the application of skills and knowledge in combination. Utah envisions the Utah core being taught this way, and it should be measured this way as well.

We note that the requirement that some measurement take place across standards, and even across clusters of standards, does not eliminate the need for rigorous alignment. Quite the contrary, the alignment approach must not only evaluate the fidelity of the items to the (possibly multiple) standards that they measure; it must also attend to the propriety of the specific combinations found in the items.

For this reason, we recommend that we begin the alignment process by establishing an alignment specification. This specification will detail which standards are best assessed in combination with specific other standards. Items will possibly be aligned to multiple standards, and those collections of alignments should align to a sensible, natural, and explicitly specified combination of standards.

With the alignment specification in hand, we will begin the process of aligning existing items to the Utah Core Standards. We will collaborate with experts within USOE to develop training materials to teach subject matter experts (SME) who will evaluate alignment about the standards and the alignment specifications.

We propose to use AIR and USOE SMEs to evaluate the initial alignment of existing items, whether those items come from existing Utah banks or from item sharing arrangements. In this section, we discuss the process AIR proposes to use to conduct the item alignment to the Utah Core Standards and to the DoKs.

As a final step, we propose to have an independent committee of teachers validate the alignment of a sample of all items, including newly developed ones. In each grade and subject, we propose to select a random sample of
approximately 60 items to be independently aligned by a committee of teachers. In order for this review to be fully independent, AIR cannot direct the work. We have set aside funds ($150,000) to fund an independent alignment study. Independence requires that AIR not oversee the work, but we are prepared to help Utah arrange for it.

**SME Training**

After all the items have been imported into AIR’s Item Tracking System along with their current alignment to the 2007 Core Curriculum, AIR’s senior test development specialists (TDS) will train the SMEs assigned to performed the alignment. AIR’s senior TDSs have extensive experience with the Common Core Standards and Webb’s DOKs. They have participated in the alignment of existing items to the Common Core Standards and Webb’s DOKs for , and , and because they have collaborated with our clients to verify the alignment, they are familiar with common mistakes aligners make and can help eliminate them before they arise.

**Internal Expert Review**

AIR proposes to use a two-step internal alignment process. First, SMEs will review the items online using the Item Tracking System and assign the appropriate Utah Core Standards and DOK levels. Then, senior TDSs will review and approve the alignments.

**External Expert Review**

After AIR completes the internal item alignment, an independent evaluator will run the committees of educators described above to evaluate the alignment of the pool. After USOE approves the alignment, the items will become eligible for inclusion in either the summative or interim assessment pools.

---

**Alignment of New Items to the UT Core Standards**

All new item alignment to the Utah Core Standards and Webb’s DOKs will arise from our standard development process. When developing new items, we start with assessment targets and item writers craft items specifically to measure them. The entire process, including our internal reviews, USOE reviews, and reviews by content alignment committees review the items for alignment.

In the next section we describe how we will run the item writer workshops in which Utah teachers will participate. During these trainings, participants will work extensively with the Utah Style Guides, Utah Item Specifications, Utah Core Standards, Webb’s DOKs, and AIR’s best item development practices. The item writer trainings are discussed in detail below.

After the trainings, item writers will be given assignments for specific content targets from the Utah Core Standards (Domain, Cluster, Standard) and DOK levels. The writers will use the Utah Item Specifications to draft items that measure specific skills, knowledge and abilities from a domain/cluster/standard. The items then follow a sequential review process automatically enforced by the Item Tracking System.

At each review level, TDSs verify the item alignment to the UT Core Standards and Webb’s DOKs, in addition to verifying adherence to style guides, item specifications, and best item development practices. After the item has gone through all internal reviews, the items will be reviewed by the USOE staff. After USOE review, AIR’s test development specialists will have a conference with the USOE staff to resolve any issues or concerns about the items. After all issues have been resolved, the
items will be ready for review by a Content Advisory Committee (CAC).

Like all the other reviewers, the CAC is tasked, among other things, with verifying item alignment to the UT Core Standards and Webb’s DoKs.

After items have been approved by the CAC and USOE, the items will be eligible for field-testing. As described above, all items, including the newly developed ones, will be part of the external alignment study.

**Item Writing Workshops**

The writing of high-quality items begins with the screening and training of item writers. Our item writer screening is a layered process in which a candidate submits a resume to AIR, an assessment staff member reviews it, and, if the candidate has potential, sends him or her an item writing instrument that is used to gauge item-writing ability. The candidate completes the instrument and returns it to AIR, where it is assessed in terms of the candidate’s writing, editing, and revising aptitude as well as the ability to align or identify items to the proper Domain, Cluster, and Standard of the Utah Core Standards. This rigorous process ensures that items reach the highest quality from the outset. This section describes AIR’s approach to effectively train item writers to write items for the Utah summative and interim assessments.

We note that the requirement in the RFP mentioned alignment, fairness, data review, and other aspects of item development. We consider these committee meetings to be distinct from item writer workshops. These meetings are described in detail below.

**Writer Recruitment and Selection**

AIR will recruit external item writers from our current pool of experienced writers, as well as Utah educators. AIR’s Item Tracking System (ITS) is accessible from any place with an Internet connection and enables us to provide training from locations in Utah and allows Utah-based item writers to work securely from their homes, schools, or offices. AIR will recruit and train item writers who are adept at following item specifications, adhere to tenets of good measurement practice, and are able to devise creative ways to assess standards. Our objective is to recruit Utah educators because of the breadth of their core knowledge of the Utah Core Standards and their genuine interest in the Utah summative and interim assessments.

Regardless of their education and employment history, all item writers will receive training in both item construction and item review. The training will include the fundamentals of item construction, a thorough explanation of Utah Core Standards and how and what they measure across content and grade level, an explanation of Webb’s Depth of Knowledge levels, and an introduction to the Principles of Universal Design and the principles of fairness and sensitivity through Language Accessibility, Bias and Sensitivity (LABS) guidelines. In addition, item writers will receive training on the use of AIR’s ITS and our security and administrative procedures.

All potential item writers will submit their resumes. AIR’s item development leaders will review all resumes and will determine which applicants meet or exceed both USOE’s and AIR’s minimum item writer prerequisites. Once identified, each candidate will be mailed an item writing instrument that measures how well the candidate writes and edits items and aligns them to Utah Core Standards. The instrument is a useful tool because it gives the
candidate a realistic expectation of what item writing entails and it provides AIR with an indication of how well a candidate performs on the most rudimentary tasks of item writing. Next, AIR will contact potential participants by phone or e-mail to confirm their selection as outside item writers (OIW). Letters of confirmation will be mailed that explain the item development procedures, project schedule, and details of the mandatory item writing workshops.

AIR will document the qualifications and experience of all item writers. AIR will submit to USOE all item writer resumes, contracts, and confidentiality agreements for documentation and reference.

**Item Writer Training**

Item writer training begins with a two-day workshop during which item writers are taught the basics of item development and review and given specific assignments. Items are then subjected to a multi-tiered review process to ensure clarity, alignment, appropriateness, and fairness in both individual and group settings.

AIR’s senior content staff lead group reviews of draft items, and OIWs are invited back to participate in item review seminars. The seminars are led by senior content staff and generally consist of a group review of items by the participating item writers, as well as those drafted by more experienced item writers. The item review seminars are part of AIR’s ongoing test development program, and these group reviews occur even when we are not training new item writers. Item writers continue to attend review seminars periodically, with their frequency of attendance diminishing as their skill increases.

We will recruit from among Utah teachers and will conduct workshops and seminars across several Utah cities.

**Workshops**

AIR will conduct item writing training workshops for all item writers according to content area. The workshops will be held in convenient central locations. AIR will coordinate the workshop arrangements, including securing a facility, preparing the necessary materials, setting up the meeting room, and providing food and refreshments during the workshops.

The item writing workshop will begin with an overview of the following three kernels:

1. Item writing and the Utah Core Standards
2. ITS
3. Item writing and deliverables

The AIR item writing trainers, who have significant experience with the appropriate content area, will educate item writers about the purpose and demands of the Utah summative and interim assessments and review sound item writing practices. Item writer training will provide the following guidelines:

- Writing multiple-choice and constructed-response items
- Writing scoring rubrics for constructed-response items
- Selecting or writing stimulus material when necessary
- Writing to the Domain, Cluster, and Standard of the Utah Core Standards
- Measuring item complexity
- AIR’s LABS guidelines
- USOE item specifications
- AIR’s ITS

AIR trainers will facilitate reviews of sample items and provide an opportunity for item writers to critique and practice reviews with items. Item writer participants will study
Utah’s item specifications and Utah Core Standards until the facilitators are confident that these writers understand the standards and what they measure.

Upon completion of the first kernel of the item writing training, the OIWs receive item writing assignments, based on development needs, for the number and types of items and for the content and requisite standards and benchmark. The OIW spend time collaborating on, discussing, and generating new item ideas, which will serve as a springboard for their assignments. Writers will review, edit, critique, and revise the work of fellow OIW to ensure that they acquaint themselves with the basic processes involved in item development.

For multiple-choice items, the focal points are as follows:

- Stem
- Options
- Key
- Distractors
- Standard
- Benchmark
- Indicator
- Item complexity/depth of knowledge
- Rationales (general and option-specific)

Next, item writers will receive training on ITS. The OIW learn how to upload item files using the multiple-choice and constructed-response templates, associate the proper attributes, and submit items for internal review. ITS allows remote and secured Internet access by item writers, enabling efficient delivery of items and standardizing format while ensuring item security.

When the OIW have demonstrated competency with item writing and maneuvering through ITS, they draft items under the guidance and expertise of experienced AIR staff. Authoring authentic items with the benefit of experienced AIR tutelage ensures that each OIW has received ample instruction to develop high-quality items. These workshop exercises contribute significantly to our item pool.

Once trained, OIW will be given specific content assignments to complete within the established timeline. The capabilities and user-friendliness of ITS allow AIR to monitor the quality of items that each item writer submits and to mark his or her areas of strength and weakness. The review components built into ITS enable AIR to provide individualized feedback to each item writer. With that feedback, the OIW hone their skills and are able to review their own items with a more candid self-assessment. This same feature enables AIR to identify the most promising item writers for use in later development efforts.

All item writers receive an Item Writer’s Manual, customized for the particular content area and the Utah context, which reflects and supports their workshop training. This document includes

- the Utah Core Standards, which emphasizes the task of targeting items to assess specific Domains, Clusters, and Standards;
- the procedures for constructing the different types of items (including appropriate stimulus materials, if applicable);
- the guidelines for ensuring fair and accessible language;
- samples of appropriate and inappropriate items, accompanied by relevant critiques for item writers’ reference;
- instructions for authoring scoring rubrics; and
- detailed instructions for item writers’ remote use of ITS.
Group Reviews

Workshop participants leave the workshops with item writing assignments, which they complete over the course of the following week. Promising item writers are asked back to participate in our review seminars, in which groups of individuals work together to review items.

The review seminars are part of AIR’s ongoing test development program, and these group reviews occur even when we are not training new item writers. The review seminars help calibrate our item development to look for the same characteristics during item reviews and provide a forum to explore new item formats and other ideas for improvement.

Apprentice item writers continue to attend review seminars periodically, with their frequency of attendance diminishing as their skill increases. Even experienced item writers are required to participate periodically to ensure that they get feedback on the preferences of our clients and learn from the comments received in client and committee reviews.

Conclusion

AIR’s recruitment and training procedures result in the harvest of a renowned item writing team that offers experience, diversity, and expertise across all content areas. Coupled with the talents of the OIW, AIR brings remarkable experience and mastery, as well as content and pedagogical expertise, to the item writing undertaking. Item writers, including educators from Utah’s own system and other experienced educators, produce a blend of professional perspectives, and our inclusive training workshops ensure that these writers are well versed in the context of the Utah assessments.

Item Preparation

AIR’s system of item development enforces rigorous training and review processes, while allowing the customization required to meet the individual needs of each of our clients. Here, we present the process in more detail by describing our Item Tracking System (ITS) and explaining how it will be used to manage the item development process effectively. Our item development process is predicated on a high level of interaction with our clients, and below we describe how ITS manages the workflow and maintains records about item development that meet USOE’s requirements.

For the following discussion, please refer to Exhibit II.B.2-2, which shows the item development process. In this exhibit, the AIR internal processes are in blue boxes, the internal USOE processes are in green boxes, and the external or committee review processes are in orange boxes.

Managing the Item Development Workflow

Described in detail in Section I.C.1, ITS is an electronic item management system that facilitates item revision, catalogs changes and comments, provides accurate documents (items), and keeps final versions of items from encountering version control problems after they have received final approval. This powerful tool supports every phase of our item development process, allowing us to effectively manage the process of multiple reviews. The system captures the outcomes and rationales at each review and maintains previous drafts of each item. ITS allows remote Internet access by item writers and reviewers and by our clients while ensuring complete security with individualized passwords for all users, limited access for external users, and strong encryption of all information.
Item Writing

When AIR begins a new project, we train item writers to make sure that they understand the new content standards, the allowable item formats, and the content limits within which they must operate to produce high-quality items. In keeping with Utah tradition, we propose to use Utah teachers as item writers and to conduct an interim level of review. The item writing workshops are described above.

Item writers produce their best items when they understand the cognitive capabilities of the students in the grade level for which they are writing. Therefore, when giving item writers assignments, we are careful to match the expertise of item writers, as defined by their teaching experience, with the grade level(s) with which they are most familiar.

Two logistical issues arise when working with teachers. First, when creating technology-enhanced items, there is a learning curve as teachers and writers learn the tools for creating machine rubrics. Second, producing graphics and other media (such as animations) requires the skills of our production staff.

We are willing and able to train teachers to use the rubric creation tools. The training typically requires about three days. For the training to be retained, item writers must continue to use the tools after training has been completed. Novice users often tend to favor items for which the rubrics are simpler, and these are often more selected-response (such as hotspot) items. While such items can be quite useful, they rarely rise to the level of DoK 3. Therefore, at least initially, we propose to teach the teachers about the capabilities of the scoring engines but have them describe the rubrics for machine-scored items, rather
than actually create the machine-readable versions of them. For this reason we have inserted a step in the process called “machine rubric review,” during which AIR test developers will review the human-readable descriptions of the rubrics and ensure that they are sufficiently precise for our staff to create the machine-scorable rubrics. In a subsequent phase, AIR test developers will implement the machine rubrics.

Graphics, animations, and other media will be treated similarly. ITS allows item writers to upload descriptions, scanned sketches, clip art, and other rough drafts of graphics to be created by graphic artists on our production team. Item writers will be trained to create and upload rough drafts in a format with which they are comfortable and route the item to our media team for creation.

**AIR Internal and Teacher Review Process**

AIR reviews items through a structured sequence of reviews. Although the exact sequence of reviews varies across programs, all our review processes include content reviews; language accessibility, bias, and sensitivity reviews; editorial reviews; and lead developer approval. ITS ensures that each review is complete before an item can make the transition to the next level of review. ITS captures every comment and revision to an item by requiring reviewers to enter annotations and preserving every version.

During the content review process, reviewers ensure that the attributes associated with an item accurately reflect the item content. Reviewers focus on each item’s match to its item specifications as well as on each item’s content and structure. Reviewers also check every item against its alignment to the academic standards to evaluate the correctness of the alignment. They make sure that the item and related materials comply with AIR’s guidelines for clarity, style, accuracy, and appropriateness for the population being assessed. For a constructed-response or technology-enhanced item, the specialists study the item’s presentation to ensure that it yields a clear response and check the scoring guide to make sure that it adequately accounts for all possible responses. For a multiple-choice item, the reviewers study the clarity and importance of each question, check that the correct response answers the question posed, and ensure that the incorrect options are indeed incorrect. In all cases, the reviewers pay special attention to the quality, equity, and accuracy of the item, as well as to any manipulatives that form an integral part of items in some content areas and to tasks and items intended for students who take alternate assessments.

The reviewers then accept the item and classification as written, make revisions, or recommend that the item be discarded. Whatever their recommendation, ITS stores the rationale and, if an item is revised, archives the older version(s).

Before USOE’s content staff and committees review an item, AIR’s lead item developer for the content area must indicate approval. The lead item developer will consider all aspects of the item and related materials for content, language accessibility, bias, sensitivity, and editorial issues, referring to the item’s history through ITS as appropriate.

We discuss the specifics of our internal and teacher reviews below.

**Machine Rubric Review**

During this review phase, our test developers will review the human-readable rubrics created by the teachers (item writers). To facilitate machine scoring of items that extend beyond
technology-enhanced selected-response mechanisms such as hotspots and matching items, rubrics must be much more detailed and specific than rubrics to which traditional item writers are accustomed. During this review, AIR content experts will review the rubrics for completeness and consistency. These reviewers will be selected from those among our staff who are skilled at developing machine rubrics for more open-ended technology-enhanced and machine-scored constructed-response item types and a variety of such item types.

Teacher Content Review

Teachers who have proven to be skilled item developers will be asked to serve as item reviewers. These teachers, working independently over the Internet, will perform the first level of content review and revision of the teacher-created items. For online reviews, the participants will view the items online and recommend acceptance, acceptance with specific revisions, or rejection of these items.

Machine Rubric Implementation

At this stage machine-scorable items that pass the teacher content review will have their machine-scoring rubric implemented by AIR test developers. If teachers are willing and able to implement the rubrics, those rubrics will be tested and evaluated at this stage.

AIR Internal Review

Our internal content reviews ensure that each item measures its intended construct. This initial review can be achieved as an individual preliminary review or a group preliminary review, where an assembly of test developers reviews items. The latter process is beneficial because the exchange of ideas and the cooperative nature of the review allow the reviewers to better acclimate themselves to the Utah Core Standards. It exposes reviewers to different approaches that may make an item meet the standard or challenge the student from a different angle. During the review phase and with a degree of knowledge beyond that of an item writer, the reviewers focus on item content and structure, scrutinize items for bias and sensitivity and language accessibility, and revise according to the client’s established style.

The test developers make sure that the items and related materials comply with AIR’s guidelines for clarity, style, accuracy, and appropriateness for the population being assessed. For technology-enhanced constructed-response items and performance tasks, the test developers review these items and verify the specific skills being assessed. Additionally, test developers ensure each item directs the student to a clear understanding of the task, allows the student to provide a clear response, and contains scoring guidelines that adequately account for all logical responses.

For multiple-choice items, the reviewers study the clarity and significance of each stem, check that the correct option answers the question posed, and ensure that the incorrect options are attractive and plausible but still incorrect.

For all items, the reviewers verify that the item is properly aligned to the intended standard(s) and DoK and ensure that a given task is of a type appropriate to the intended specifications. Based on the meticulous review of each item, the reviewers accept the item and classification as written or revise the item, attributes, or classification or all three. Another review alternative is to reject the item because it is too problematic in content or does not align to any standard, or both. Whatever the recommendation, the review comments are noted in ITS and, if an item is revised, its previous version is automatically archived.
During the review process, test developers check for the specific features of multiple-choice and constructed-response items that tend to facilitate (or impede) understanding.

For multiple-choice items, the following item components are confirmed:

- The items test the knowledge and skills required by the test specifications.
- The items represent an appropriate level of cognitive complexity for the intended examinees and are grade-level appropriate.
- The items do not introduce content errors or promote misconceptions.
- The items require a level of reading skill appropriate to the measured construct.
- The items meet the standards for fairness.
- The items within a larger set based on the same stimulus material do not provide clues to the answer for any other item in that set.
- The items within a larger set do not depend on any other items in the set for the correct answer.
- The items adequately present the problem addressed.
- The items are clear, concise, and free of construct-irrelevant detail. Excess text and extraneous graphics, for example, make it impossible to know whether the examinee’s response results from his or her knowledge or from interference from construct-irrelevant information, thus compromising validity.
- The options contain only one correct answer.
- The options are grammatically consistent with the stem and reasonably parallel in grammatical structure and length.
- The options are free from overlap.
- The options are attractive and plausible but incorrect and are based on a feasible misunderstanding about the content.
- The options are logically ordered (numerically, chronologically).
- The options are free of cues that would indicate the correct answer, such as the repetition in the key of wording from the stem.
- The options are accompanied by a rationale for why they are wrong (in the case of distractors) or right (in the case of the key).
- The options are free of language such as “all of the above” and “none of the above.”
- The options are generally free of absolute wording such as “always” and “never,” because these absolutes actually apply to very few concepts.
- The stimulus materials provide all necessary information and exclude irrelevant information.
- The stimulus and/or item has been appropriately fact-checked, if necessary.

For constructed-response items, the test developers will confirm the following:

- The items/tasks test the knowledge and skills required by the test specifications.
- The items/tasks represent an appropriate level of cognitive complexity for the intended examinees and are grade-level appropriate.
- The items/tasks require a level of reading skill appropriate to the measured construct.
- The item does not introduce content errors or promote misconceptions.
- The items/tasks within a larger set based on the same stimulus material do not provide clues to the answer for any other item/task in that set.
The items/tasks adequately present the problem addressed.

The items are clear, concise, and free of construct-irrelevant detail. Excess text and extraneous graphics, for example, make it impossible to know whether the examinee’s response results from his or her knowledge or from interference from construct-irrelevant information, thus compromising validity.

The constructed-response format of items/tasks is more useful than the selected-response format of multiple-choice items for making judgments about an examinee’s understanding of the measured construct.

The item makes no construct-irrelevant demands that would interfere with a scorer’s, or an automated scoring system’s, ability to identify and correctly credit correct elements of the response.

The items/tasks meet the standards for fairness.

For scoring rubrics, the lead test developers will confirm the following:

- The rubric applies to all logical and potential responses.
- The item/task will elicit responses from candidates at all score points of the rubric.
- The score criteria in the rubric are consistent with the question asked/task set, as well as any information conveyed by an item or a task stimulus.
- The levels of the rubric are clearly differentiated.

Machine-scoring rubrics are tested with as broad an array of possible responses as the test developers can imagine to exercise the machine rubrics.

Language Accessibility, Bias, and Sensitivity Reviews

AIR standards require that tests exclude items that disadvantage examinees of different races, genders, ethnicities, regions, handicapping conditions, or language background. Further, the standards require documentation that such disadvantage does not occur. (Note that this does not imply that all groups perform equally well on all items, but rather that any systematic differences are due to differences in the proficiency levels of the groups on the targeted trait.)

AIR pairs the language accessibility with the bias and sensitivity review because language accessibility is an aspect of fairness. This review accentuates AIR’s commitment to ensuring that items and any associated stimulus materials are without 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.

AIR Editorial Review

AIR editors will review every item for clarity, correctness, and appropriateness of language for the grade level assessed, and conformity with acceptable item-writing practices. Editors ensure clearly worded, understandable, and fair presentation of items, instructions, and administrative documents. The editorial process is related to, but separate from, the item development process so that editors can look at the items objectively.

Editors have numerous tasks. First, editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items. AIR editors have adopted standard reference resources, among which
are Merriam Webster’s Collegiate Dictionary (11th edition) and Webster’s Third International Dictionary, Unabridged for spelling and capitalization; Associated Press Stylebook and Briefing on Media Law for usage; Words into Type for grammar; and the Publication Manual of the American Psychological Association for reference citations. Subject-specific reference sources include Webster’s New Biographical Dictionary and Webster’s American Biographies; The Harvard Dictionary of Music; History of Art (Jansen); Webster’s New Geographical Dictionary; Dorland’s Illustrated Medical Dictionary; Mathematics Dictionary (James and James); Scientific Style and Format: The CBE Manual for Authors, Editors, and Publishers (Council of Biology Editors); and The Macmillan Dictionary of Measurement.

AIR will use Utah’s style guides that synthesize frequently used and assessment-relevant information that addresses such test-specific issues as numerals, abbreviations, symbols, and terminology. AIR editors become very familiar with USOE’s preferred format and style.

Second, editors ensure that all items are accurate in content and will query the item developer when questions arise. The editors compare reading passages against the original publications and make sure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. The editors ensure that the keys are correct and that all information in the item is acceptable and correct. For example, editors verify common facts (e.g., the diameter of Earth, the scientific name of the fruit fly, the correct spelling of Far from the Madding Crowd, the formula for calculating the volume of a cube, the proper use of the semicolon, and the dates of the Civil War). Mathematics assessments present a specific challenge because of the potential for transposing numerals and the difficulty of identifying computation errors by sight. Therefore, our editors perform all calculations to ensure accuracy.

Third, AIR’s editors review all material for fairness and language accessibility issues. Although external committees and the lead item developers look at all material, editors raise questions before the material reaches the committees and then recheck all material that undergoes any changes.

Finally, our editors confirm that items reflect the accepted guidelines for good item construction. For example, in mathematics items, they ensure that options given with calculator-active items reflect errors in thinking, not errors in calculator use. In all items, they look for language that is simple, direct, and free of ambiguity with minimal verbal difficulty. Editors confirm that a problem or task and its stem are clearly defined and concisely worded with no unnecessary information. For multiple-choice items, editors check that options are parallel in structure and fit logically and grammatically with the stem and that the key accurately and correctly answers the question as posed, is not inappropriately obvious, and is the only correct answer to an item among the distractors. For constructed-response items, editors review the rubrics for appropriate style and grammar.

**AIR Senior Review**

We seek to deliver complete, finished, and very high-quality items to USOE the first time. A subject area content lead must sign off on the overall quality, including content, fairness, language accessibility, and style before it is deemed ready for USOE review. That happens in this review level.


**USOE Review and Resolution**

USOE review is a critical step in the overall quality and adequacy of the assessment. AIR will deliver newly developed items in batches to allow our test development specialists to calibrate these items to the expectations of the USOE reviewers. The initial batches will contain items that target a representative sample of the standards, item types, and depth of knowledge. After the initial batch delivery, the item distribution across the content and DoK in subsequent batches will vary according to the item development plan.

After we receive feedback for each batch, our test development specialists and USOE staff will hold a conference call to discuss USOE feedback and finalize revisions.

**Content Committee and Fairness and Sensitivity Committee Review**

AIR will work with USOE to develop a description of the role and responsibilities of the Content Advisory Committee (CAC) for each assessment and provide a toll-free number for questions from interested Utah residents. The most fundamental criterion for selecting the advisory committee members will be that the member is a teacher in a Utah school district.

AIR will employ the follow procedures to establish the committees:

- AIR will coordinate with USOE and contact Utah teachers by e-mail to determine their willingness to serve as committee members. AIR will describe the nature of the work, the number of meetings, and any honoraria. If the prospective committee member meets the requirements, AIR will secure contact information and a description of his or her background, including current teaching assignment in the State of Utah and/or the grade level(s).

- AIR will provide Utah with information on potential committee members for approval and recommendations. Utah will send its approved list to AIR, along with the appointment of a committee chairperson(s) if deemed advisable.

- AIR will contact the approved committee members to invite them to the committee meetings in Salt Lake City, Utah, or any other city Utah deems necessary, and to make travel and hotel arrangements.

- AIR will send the following to each committee member well in advance of the meeting:
  - Meeting agenda
  - CAC member contract
  - Copy of ‘s Grade-Level Expectations (GLEs)
  - Ancillary background information the member will need to fully participate in the discussions at the meeting
  - Travel arrangements, if needed
  - Hotel reservations, if needed

The CAC, which is composed of Utah teachers, administrators, and content experts, will be responsible for reviewing the items and scoring rubrics from a state educator’s perspective. Content reviewers will focus on item and scoring-rubric quality, the appropriateness of the items for the grade level, the representative nature of the item set, and the match of the items to Utah’s standards and test specifications.

For the CAC face-to-face meetings, we propose having 12 groups of three or four educators review different sets of items in each of ELA and math, and four similar groups for science, where there will be fewer items. The groups will recommend that each item be either accepted or rejected. Each group will designate a member as recorder; that member will record group decisions in a review log. For
items recommended for rejection, reviewers may suggest fixes in the “Comments” column of the review log. AIR and Utah staff will float between groups to answer questions and provide guidance as needed.

After all the groups have concluded their review, AIR and USOE will have a resolution/reconciliation meeting to go over the committee’s recommendations. At this meeting, AIR and USOE will review the items recommended for rejection and discuss possible fixes. Any items that can be fixed through collaboration between USOE and AIR will be revised directly in ITS. USOE can then decide whether the items can move on to the Fairness and Sensitivity Committee or be flagged for Content Committee review the following year.

This process is similar to one that is used by one of our current clients. It allows for the review of hundreds of items each day. Exhibit II.B.2-3 shows the number of days the Content Committee will take to review items during face-to-face meetings.

Exhibit II.B.2-3: Items and Committee Days

<table>
<thead>
<tr>
<th>Subject</th>
<th>Items</th>
<th>Subcommittees of 3–4</th>
<th>Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>2450</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>Math</td>
<td>2570</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>Science</td>
<td>580</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

**Fairness Committee Review**

An integral practice embedded in the review process is to have all items and item-related materials (stimuli and manipulatives) reviewed by an outside panel of Utah educational stakeholders. This Fairness and Sensitivity Committee is made up of Utah state representatives, citizens, and educators from across the state, so that there will be a diverse representation of the Utah population.

Training will cover these guidelines established for fairness and sensitivity training:

- Sensitivity
- Diversity
- Stereotyping
- Inflammatory or controversial subjects
- Ethnocentrism
- Differentially familiar subjects
- Appropriate language and grammatical constructions
- Appropriately accessible graphics

AIR will distribute training materials that encompass the pertinent fairness guidelines to each committee member. The materials will also include documentation to cover the following:

- Purpose of fairness review
- Guidelines
- Examples for guidelines
- Exceptions to guidelines

The AIR facilitator ensures that each item is discussed based on the fairness guidelines. The possible outcomes for items and stimulus materials that are reviewed at the Fairness and Sensitivity Committee meeting are accept as appears (AA), accept as revised (AR), and reject (RJ). AIR will make notes and record the committee’s decisions on the review log.

AIR will employ the following procedures to establish the committees:

- AIR will coordinate with USOE to recruit Utah citizens to participate on the Fairness and Sensitivity Committee. AIR will send information relevant to the nature of the
work, the number of meetings, and any honoraria. If the prospective committee member meets the requirements, AIR will secure contact information and a description of his or her background.

- AIR will provide USOE with information on potential committee members for approval and recommendations. USOE will send its approved list to AIR, along with the appointment of a committee chairperson(s) if this is deemed advisable.

- AIR will invite the approved committee members to the committee meetings in Salt Lake City, Utah, and make travel and hotel arrangements.

AIR will send the following to each committee member well in advance of the meeting:

- Meeting agenda
- FSC member contract
- Copy of fairness guidelines
- Ancillary background information the committee member will need to fully participate in the discussions at the meeting
- Travel arrangements, if needed
- Hotel reservations, if needed

We propose to hold the fairness committees using the same structure, participant counts, and meeting days as the content committees.

**Committee Review Resolution**

After all the groups have concluded their review, AIR and USOE will have a resolution/reconciliation meeting to go over the committee’s recommendations. At this meeting, AIR and USOE will review the items recommended for rejection and discuss possible fixes. Any items that can be fixed through collaboration between USOE and AIR will be revised directly in ITS. USOE can then decide whether the items can move on to the Fairness and Sensitivity Committee or be flagged for Content Committee review the following year.

**Annotations and Translations**

Under this contract, AIR includes the annotation of items for text-to-speech (TTS) and annotation of a fixed-form of each test in Braille. We offer, as a separately priced option, to annotate and prepare for Braille delivery enough items to allow for adaptive delivery of the tests to students who use Braille.

**Text-to-Speech**

TTS is an accommodation offered to remove barriers that interfere with students’ ability to demonstrate what they know and can do. TTS differs from audio CDs and material read aloud by a proctor, which are test-form specific, because the entire item bank can be field-tested with TTS. Thus, operational tests for students needing the TTS accommodation will be selected from the entire operational pool—not from a subset, as is the case with audio CDs and read-aloud accommodations.

AIR uses a structured TTS tagging process that enforces a predetermined review sequence. First, the items/passages are tagged using state-specific tagging guidelines. Second, the tags are reviewed and revised as needed by a different TTS reviewer. Third, the tags are reviewed and approved for review by AIR’s test development specialists. Fourth, test development specialists preview the items/passages as students will see them, verify that the formatting is correct, verify that the TTS tags meet the state guidelines, and approve the item/passages for the USOE’s review/approval.
**Text-to-Braille**

Text-to-Braille (TTB) is an accommodation offered to remove barriers that interfere with the ability of blind or limited-vision students to demonstrate what they know and can do. TTB requires extensive tagging; therefore, we do this tagging during the final phase of development. It should also be noted that items tagged for TTB are not included in field-testing. Because only a limited number of students use the TTB accommodation, the data gathered in a field test would not add meaningfully to the statistical properties of the items. In addition, Braille tests take longer to administer, and it would be unfair to the students who need this accommodation to require them to take the assessment when they will not receive a score.

AIR will use a structured TTB tagging process that enforces a predetermined review sequence. First, the items/passages will be tagged using state-specific tagging guidelines by a reviewer who is well versed in Braille. Then the tags will be reviewed and revised as needed by a different TTB reviewer.

**Parent Review**

As stated in the State of Utah’s House Bill 15, the items must be reviewed by a panel of 15 parents before moving forward to final review. This annual review will take place after the items have passed through the preliminary stages of item development and content/bias review by the staff at USOE and AIR. All edits must be made and the item must be approved by USOE before being evaluated at the Parent Review.

We propose an in-person committee meeting. Parents will not be asked to modify items, only to identify items that they believe are inappropriate. If we assume that parents can review approximately 25 items per hour, and we assume a seven-hour work day, the final pool will require approximately 85 committee days to complete. We propose to assemble up to 20 committees of parents to review the items. We propose to ask each committee to review items over two sequential weekends, and we propose running five committees at a time. The parental review of the total initial pool will require approximately 8 weeks.

AIR will provide the space and local ground transportation for participants, along with meals and snacks during the committee meetings. We do not plan to offer a stipend for participation. At USOE discretion, we can hold the committee meetings in a variety of population centers in Utah.

**Field Test and Data Review**

Following our analysis of field-test items, AIR will provide to USOE a Master Data Sheet (MDS) and access to the ITS item cards. The MDS is a Microsoft Excel data file that contains a comprehensive summary of the statistics on each of the field-test items. The item cards are specially formatted files that provide a high-level summary of each field-test item. The item cards are accessed electronically via the ITS system and can generate a PDF file for each field-test item. Both files provide information on the items that USOE can use to determine whether the items will remain in the pool or whether specific items should be removed based on some of the field-test statistics.

AIR proposes to use in-person data review committees made up of approximately 10 members per subject who represent expertise in the diversity of students in Utah’s schools. We will prepare item cards and have the committee review only the items that are flagged for a high risk of bias. In general this is a relatively small proportion of the items field-tested.
USOE Reviews, External Content Expert Reviews, and Committee Reviews

The USOE review process provides the next layer of quality assurance in the development of valid, effective items. USOE may review items either online or on paper. In the latter case, our staff will enter USOE’s comments and review outcomes into ITS. Similarly, we will enter the comments and outcomes in real time during the various committee reviews. We expect that the outside content experts will use ITS directly, thereby ensuring an efficient workflow.

During these reviews, USOE and your committees will view the items along with the passages and graphics, using the same software that displays these items for students. Thus, both USOE and your committees will see exactly what the students will see as they take the assessment.

Final Review

After going through AIR, USOE, and external reviews, each item must go through a final resolution by one of our senior test developers, typically the content area lead. This final sign-off indicates that the item is ready for the field-test pool and can be placed on forms. It certifies that all previous revisions have been appropriately made and that the revision process did not introduce new, unforeseen problems in the items. Test developers assemble the items into a computer-delivered field-test pool. The senior test developer reviews the items during web review—so that they are seen just as the student will see them—and signs off if the items are approved. Once approved, the pools will go to USOE for web approval. Again during this phase, USOE will view the items using the same software the students will use and therefore you will see the items as they will appear on their assessments.

Rubric Validation and Rangefinding

Following field testing, the items must be scored and the data aggregated across students. Before the data for constructed-response items and computer-scored constructed-response items can be aggregated, the items must be scored. Written constructed-response items are typically scored after a rangefinding process takes place. During rangefinding, the rubric for the items is used by a committee of teachers and the project’s scoring director to determine the scores for several sets of papers. Papers for each set are selected because they are representative of each point in the range of scoring points. These papers are used to train the scorers, assess their ability to use the rubric accurately, and validate their scoring during the entire scoring process.

Machine-scored constructed-response items go through a process called rubric validation. In this process, the machine-scored papers are validated as correct by a committee of teachers. If errors are noted in the scoring, the scoring rubric is tweaked to take unanticipated responses into account. Therefore, all items that leave open the possibility of innovative and unexpected student responses or that are being scored by artificial intelligence (AI) approaches go through rubric validation.

Our rubric validation process is described in detail in Section I.C.3.d, Scoring.

Moving Items from the Field-Test Pool to the Operational Pool

Items transition from the field-test pool to the data review pool after field testing. Items are reviewed by the data review committees and the USOE content staff to evaluate the implications of any statistical flags. Approved items are then moved to the operational pool, from which they may be assembled onto operational
forms. ITS maintains information about each use of an item (either field-test or operational) and the item’s release status.

**USOE’s Role**

The foregoing discussion of our item development process described the points at which USOE is involved in the review process. In summary:

- USOE is invited, but not required, to attend our item writing workshops.
- After items are deemed ready by a teacher reviewer and AIR internal review processes, USOE will review and provide comments on the items.
- USOE will attend CAC and Fairness and Sensitivity Committee reviews and participate in resolutions following these reviews.
- USOE is invited, but not required, to attend some or all of the Parent Review committee meetings, which we will schedule on weekends to make them accessible to parents.
- USOE will participate in the data review following field-testing.

**Item Style Guide**

AIR will review the current Utah style guides and help USOE update them. If Utah plans to use items from one or more consortia, bringing the style guides into alignment may prove helpful.

We expect that as Utah offers an increasing array of item types and accommodations, USOE will need to revise the style guide to capture requirements for these as well.

Much of the work of Accessibility by Design must be done through specifications such as style guides. These requirements are subtle and indirect, and we believe that our experience delivering accessible, online adaptive tests will prove valuable to USOE as it moves into this arena.

We offer a very simple example. One common accommodation offered is different background and foreground colors. Often this is done to reduce contrast and help students with visual/perceptual issues or increase student focus. To support this, graphics must be on a transparent background. Furthermore, when a graphic contains the background color, the image can become confusing, so graphics that appear on different backgrounds must avoid all potential background colors. These design elements must be built into the content to avoid conflicts with the accessibility features. Accessibility by Design requires this attention.

AIR will review USOE’s current style guide and recommend enhancements to support accessibility features. We are currently helping Smarter Balanced revise its style guide with the same objectives. Some features we find that need revision in Smarter Balanced’s style guide are constraints on background colors; rules for which words are annotated with translations, glossaries, or context-sensitive dictionaries; technical construction of graphics to support flexible font changes; and media formats and target file sizes. We will review the USOE style guides with equal care.

New items will be constructed in conformance with the revised style guide.

**Item Characteristics**

AIR values innovative technology and has extensive experience creating and delivering engaging technology-enhanced items that assess the breadth and depth of our clients’ state curriculum standards to the fullest extent possible. Currently, AIR successfully applies these innovative approaches in , and the
In fact, AIR is taking the lead on most innovative items for the Smarter Balanced pilot test, and the sample items that Smarter Balanced is releasing in early October were all developed using AIR’s technology and are hosted using some of the same item rendering and scoring engines we are proposing for Utah.

AIR understands that detailed, comprehensive test blueprints are at the core of all successful assessment programs as they inform the overall structure and composition of the item pools. AIR proposes to use the following blueprints for the Utah summative assessments in reading, math, and science. AIR will work with USOE to finalize the blueprints during the initial planning meetings.

**Reading Blueprints**

For reading, AIR proposes a blueprint with a computer-adaptive component (CAT) and a research task (RT). The CAT component will have 45 items and 4 to 6 stimuli; the RT component will have 4 to 6 items and 2 stimuli. Overall, each test will include 26–43 multiple-choice items, 2–12 technology-enhanced items, 0 or 1 simulation, and 1 research task. Additionally, each test will have the following distribution of items across the DoK levels: DoK 1, 20%–30%; DoK 2, 20%–70%; DoK 3, 10%–20%; and DoK 4, 1%–15%. The percent ranges across the DoK levels were derived from averages of other state tests. We propose to use machine scoring for 100% of the items and human scoring for up to 6 items per test if student scores on research task items are found to be significantly different from the scores on the CAT component of the test. Please see Appendix A, Blueprints, for detailed blueprints at each grade.

**Math Blueprints**

For mathematics, AIR proposes a blueprint with a computer-adaptive component (CAT) and a performance task (PT). The CAT component will have 40 items at grade 3 and 45 items for the other grades; the PT component will have 4 to 6 items. Overall, each test will include 28–39 multiple-choice items, 3–12 technology-enhanced items, 0 or 1 simulation, and 1 performance task. Additionally, each test will have the following distribution of items across the DoK levels: DoK 1, 20%–50%; DoK 2, 35%–70%; DoK 3, 15%–60%; and DoK 4, 1%–15%. The percent ranges across the DoK levels were derived from averages of other state tests. We propose to use machine scoring for 100% of the items and human scoring for up to 6 items per test if student scores on performance task items are found to be significantly different from the scores on the CAT component of the test. Please see Appendix A, Blueprints, for detailed blueprints at each grade.

**Science Blueprints**

For science, AIR proposes a blueprint with a computer-adaptive component (CAT) and a performance task (PT). The CAT component will have 48 items at grades 4, 5, 7, 8, Biology, and Physics and 53 items at grade 6, Earth Systems Science, and Chemistry. The PT component will have 4–6 items. Overall, each test will include 35–45 multiple-choice items, 5–11 technology-enhanced items, 0 or 1 simulation, and 1 performance task. Additionally, each test will have the following distribution of items across the DoK levels: DoK 1, 25%–55%; DoK 2, 40%–70%; DOK 3, 15%–30%; and DoK 4, 1%–15%. The percent ranges across the DoK levels were derived from averages of other state tests. The distribution of items across the Intended Learning Outcomes (ILOs) will be the same as those on the current Utah CRT.
blueprints. We propose to use machine scoring for 100% of the items and human scoring for up to 6 items per test if student scores on performance task items are found to be significantly different from the scores on the CAT component of the test. Please see Appendix A, Blueprints, for detailed blueprints at each grade.

**Recommended Pool Structure**

For all subjects, AIR proposes to create item pools that contain 8 to 10 times the number of items needed for a single administration. This is sufficient to support at least 3 opportunities per test (two interim and one summative). The pools will seek to have items distributed proportionally to the blueprint across the Utah Core Standards, item types, DoK levels, ILOs, and Mathematical Practices, as appropriate to the subject.

**Item Distribution Across Full Range of Achievement Spectrum**

AIR will develop a pool that covers the full range of the achievement spectrum. This pool will evolve over time, as it is not possible to predict item difficulty with any precision. Experienced test developers can usually identify items as easy, average, or difficult, but precision beyond that requires field-testing.

Two key features characterize our approach to ensuring that the tests cover the full range of the achievement spectrum:

- We will ensure that each standard is measured by items that cover the fullest range of achievement.
- Our pre-deployment simulation procedures will enable us to plan the item pool in the fall in time to begin the development cycle for items to be field-tested in the school year.

**Measuring the full range of content**

It is a myth that items measuring a single content standard cluster around particular difficulty targets. Exhibit II.B.2-4 from the item pool illustrates the difficulty range of items measuring a single middle school standard. As this chart clearly illustrates, items aligned to this single standard range in difficulty from far below basic to the highest reaches above advanced. Covering the common core for every student means ensuring that every student has a fair opportunity to show what he or she knows and can do across the range of difficulty and cognitive complexity. It is neither necessary nor acceptable to skip measuring some grade-level standards because of a student’s performance on some items measuring other content. Exhibit II.B.2-4 makes this very clear.

AIR will endeavor to obtain or develop items across the range of difficulty for all standards.

**Percentage of items tagged with accessibility information**

The word “tag” is often used to mean one of two things. First, item metadata is often referred to as “tags,” so an item might be “tagged” to identify it (for example) as accessible for students with visual disabilities. We will tag all items with accessibility tags of this sort, to be decided on collaboratively with USOE.

We also use the word “tag” to refer to the process of annotating items for delivery using text-to-speech or text-to-Braille.

The RFP requires only a single fixed-form Braille version of each test, and our base price includes the annotation of that form, but we propose to deliver that form using the online delivery system and embossers and refreshable Braille displays. A preferred solution would be
Exhibit II.B.2-4: Example of Range of Item Difficulty for a Single Standard

**Items Measuring the Same Skills Are Not All Similarly Difficult**

Difficulty of items measuring Minnesota Math Standards 8.2.3.1: Evaluate algebraic expressions, including expressions containing radicals and absolute values, at specified values of their variables.

<table>
<thead>
<tr>
<th>Below</th>
<th>Approach</th>
<th>Proficient</th>
<th>Above</th>
</tr>
</thead>
</table>

Some models of adaptive testing models start from the incorrect belief that items measuring similar content are similarly difficult. This is simply not true.

Therefore, we include the price of text-to-speech annotation of the full item bank as a separate, optional line item. We encourage USOE to include this in the base price if these annotations are included as mandatory requirements of the RFP or are included in the base prices of other offerers.

**Sample Items**

At the beginning of this component of the proposal we describe our innovative technology, and provide a gallery of sample items and performance tasks.

**Item Exposure Metrics**

Our system captures exposure data on every item, and we will make these data available to USOE. We will work with USOE to identify exposure targets and item rest and retirement...
policies. We recommend that items overexposed on the summative assessment transition to the summative assessment pool or be released as part of the formative pool.

We note that our pre-deployment simulation process predicts item exposure rates with great accuracy, and this provides an early opportunity to modify pools and administration parameters to manage item exposure during the testing window.

**Item Alignment**

Our plans for evaluating and reporting on item alignment are described earlier in this section under Process for Evaluating the Alignment of Items.

**Universal Design/Accessibility**

AIR has a long history of educational research and support for students with disabilities and English language learners. We have been committed to accessibility for all students ever since we conceptualized and designed our test delivery system and its items. We continue to make accessibility our top priority as we build more innovations in student testing and learning.

AIR’s test delivery system offers a robust array of accessibility features, as illustrated in Exhibit II.B.2-6. Before discussing these features, however, it is worth mentioning that the adaptive engine itself is an accessibility feature. The engine ensures that each student sees the full range of content and the full range of depth of knowledge, regardless of his or her performance, though these goals can be accomplished while still varying item difficulties.

AIR seeks universal usability through

- applying universal design principles wherever possible, in both our item development and our design of user interfaces; and
- supporting **Accessibility by Design** (accessible design), which engineers items and their delivery to accommodate and support variations in delivery for individuals with special needs that cannot be accommodated through universal design.

We discuss these design principles below.

**Universal Design and Accessibility by Design**

Universal, accessible, and usable designs describe similar but distinct approaches to assessment development that can result in an assessment that is easier for everyone to use. The concept of universal design began with Ron Mace, an architect who used a wheelchair. Mace insisted that special purpose designs were not necessary to meet compliance codes (and, furthermore, these designs were often stigmatizing to the individuals involved) and promoted a *universal design* that considered the needs of most people early in the design process (Center for Universal Design, 1997). Officially defined by the Center for Universal Design in 1997, universal design addresses the larger issues of usability by making accessibility easier for everyone. It is distinct from the other two approaches in that the needs of individuals are planned for from the beginning as opposed to retrofitting a solution after the fact.

As described by Thompson, Johnstone, and Thurlow (2002), the principles of universal design can be applied to large-scale assessments to create assessments that enable the participation of the widest possible range of students. By eliminating many access barriers associated with the test itself, the need for accommodations may be reduced (Thompson et al.). Even the use of universal design as the overarching approach for assessment design, though, cannot meet the needs of all assessment users. For this reason, the accessibility and usability of the assessment must also be considered by assessment developers.
Accessible design focuses on making the existing assessment usable for particular disability groups by retrofitting the assessment or by using assistive technologies to meet the specific needs of a particular disability group (e.g., closed captioning for individuals with hearing impairments). Some disability areas that should be considered when developing assessments include color blindness; repetitive stress; hearing loss; and visual, other hearing, speech, motor, and cognitive impairments.

Accessible design often requires planning at the item development phase. Annotations, as are used for text-to-speech or text-to-Braille, offer an obvious example. More subtly, if a system is going to offer different background colors, graphics must generally avoid these background colors. Often, entirely different graphics are required. Some of these preparations require only careful attention to style guides and specifications. Others, such as producing multiple versions of graphics, increase the labor required, and therefore increase costs.

AIR seeks to serve all students. We will work diligently with USOE to define the accommodations that will be offered immediately, or may eventually be offered. Where accessible design does not increase costs, but only requires careful planning and specifications, we will work to incorporate these details into the specifications and adhere to them. Where offering the features increases costs beyond the requirements of this RFP, we will work with USOE to find affordable solutions.

**Delivery Choices**

We 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 conundrum: How will students know what features are available? Students test once or a few times a year. Therefore, students will generally not become expert users of the software. The appropriate tools are best made *meaningfully available* by offering each student only the set of tools that he or she needs without the distraction created by unneeded tools.

Each tool or embedded support provided by AIR’s system, by design, may be configured to be

- available to all students;
- made available to students by one or more adult roles in the system (e.g., state, district, or school test coordinator, test administrator, etc.); and
- offered to all students, but configured for each student by adult users.

Exhibit II.B.2-5 illustrates some of the configurations other clients have chosen for some of our offered accommodations.

Some accessibility features require special attention and preparation to be effective and meaningful for the item and the student taking the test. These features were designed to serve their primary purpose in as meaningful a way as possible by pulling in the appropriate resources. In addition, each feature is effective only if it interacts seamlessly with the items it’s presented with as well as other features. Exhibit II.B.2-6 lists these features and their respective requirements for preparation.

**Supported Features**

Exhibit II.B.2-7 lists the accessibility features available on the system. With few exceptions, our system currently supports all the accessibility features stated in the RFP.

Each feature can be one of four different configurations: (a) available to all students; (b) assigned to students in advance by the desig-
nated state, district, or school administrators or through data upload; (c) assigned to students at testing time by the proctor; or (d) configurable statewide as a choice for USOE.

We note that the availability of a system feature does not imply content annotated or prepared for delivery using that feature. Please see the Test Development section of this proposal for a discussion of annotation and other aspects of content.

**Other Features That Are Available on AIR’s System but Not Indicated in the RFP**

Accessibility features, when built appropriately, can serve in various ways; (a) to help students visually; (b) to help focus students’ attention; (c) to help make students more agile; and (d) to provide references for students (still without changing the difficulty of the items).

From the conceptual and design phases of the test delivery system, AIR has included these accessibility features and built a scalable system so it can accommodate many more accessibility innovations without the need for system rewrites. Exhibit II.B.2-8 lists additional features that are not in the RFP but are in AIR’s system.

---

**Exhibit II.B.2-5: Examples of Access Configurations for a Sample of Embedded Supports and Accommodations**

<table>
<thead>
<tr>
<th>Accessibility Feature</th>
<th>Example of Configuration Choice</th>
<th>State’s Rationale behind Configuration Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color choice</td>
<td>Assigned to students at testing time by the proctor</td>
<td>Most of our clients recognize that the proctor is the best person to determine the appropriate color because the student is with him or her in the room. If this tool were made available to all students, it would provide a distraction.</td>
</tr>
<tr>
<td>Increased font size</td>
<td>Default font size (assigned to students in advance)</td>
<td>If a student needs this feature, this fact will be part of his or her school record and so can be determined in advance.</td>
</tr>
<tr>
<td>Text-to-speech: directions, passages, items</td>
<td>Assigned to students in advance OR Assigned to students at testing time by the proctor</td>
<td>In some states, a student’s need for this feature is part of his school record and so it can be determined in advance. In other states, the feature is available to all students but a Proctor is required to turn it on, so that he or she can ensure that the student has a headset and will not disrupt anyone else.</td>
</tr>
<tr>
<td>Magnification</td>
<td>Available to all students</td>
<td>This feature does not provide a distraction, and the student may need it on different screens.</td>
</tr>
<tr>
<td>Answer eliminator</td>
<td>Available to all students</td>
<td>This feature does not provide a distraction, and the student may need it to answer items.</td>
</tr>
<tr>
<td>Refreshable Braille</td>
<td>Assigned to students in advance</td>
<td>A student’s need for this feature is part of his school record and so can be determined in advance.</td>
</tr>
</tbody>
</table>
### Exhibit II.B.2-6: Accessibility Features and Preparation

<table>
<thead>
<tr>
<th>Accessibility Feature</th>
<th>Preparation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Color choice</strong></td>
<td>Colors that will be available on the system must 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 revision if color conflicts arise.</td>
</tr>
<tr>
<td><strong>Text-to-Braille and text-to-speech</strong></td>
<td>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 USOE.</td>
</tr>
<tr>
<td><strong>Large print and magnification</strong></td>
<td>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 given specific zoom levels.</td>
</tr>
<tr>
<td><strong>Sign language</strong></td>
<td>Recorded sign translations need to be created in advance and undergo review levels.</td>
</tr>
<tr>
<td><strong>Audio files</strong> (not included in RFP, but AIR has capability)**</td>
<td>Audio files need to be created or recorded in advance.</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>Content needs to be properly translated in supported languages other than English.</td>
</tr>
</tbody>
</table>

### Compliance with Accessibility Standards and Policies

AIR is committed to having our test delivery system and items comply with industry-wide standards and policies. We recognize that standards (especially the ones that are widely accepted and mentioned in the RFP) provide the foundation for many of the technologies and innovative accessibility features, as well as a consistent, sharing platform across education providers. This section describes our current level of compliance with each standard mentioned in the RFP.

#### APIP Standards Compliance

Accessible Portable Item Protocol (APIP) provides a standardized data model for exchanging file formats for test items.

AIR is currently developing the capability to import and export APIP-compatible items. APIP builds on the QTI specification. We can import and export most item types in QTI. Furthermore, our system supports a sophisticated set of “tags” or annotations to support text-to-speech, translations, and other content augmentations, including sign-language videos. The ability to create and store these data ensures that we will be able to export them in APIP format.

#### PNP Standards Compliance

We are proud that we offer the most individually customizable assessment platform in the industry. We do not currently use the Personal Needs and Preferences (PNP) representation. We are in the process of evaluating the stand-
## Exhibit II.B.2-7: Available Accessibility Features on AIR's System

<table>
<thead>
<tr>
<th>Accessibility Feature</th>
<th>Description</th>
<th>AIR System Capability and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased font size</td>
<td>The number of levels (generally, three) and rate of increase (generally, 1.25x the previous level) are configurable.</td>
<td>Currently available</td>
</tr>
<tr>
<td>Fore and background color</td>
<td>Any foreground color and background color can be supported.</td>
<td>Foreground color (currently available)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Background color (under development; scheduled to be available fall 2013)</td>
</tr>
<tr>
<td>Calculation devices</td>
<td>Three types of online calculators are currently available: four-function or basic, graphing, and scientific. The calculator can be dragged around the screen so that the student can always make the affected item visible.</td>
<td>Currently available</td>
</tr>
<tr>
<td>Protractor, ruler, periodic table</td>
<td>Protractors and rulers are available at the item level so that scaling is always synchronized with the item content.</td>
<td>Currently available</td>
</tr>
<tr>
<td></td>
<td>A periodic table is available at the subject level. The periodic table can be dragged around the screen so that the student can always make the affected item visible.</td>
<td></td>
</tr>
<tr>
<td>Additional time</td>
<td>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.</td>
<td>Currently available</td>
</tr>
<tr>
<td></td>
<td>However, if an unforeseen event such as a fire alarm triggers additional testing time, AIR's system can enable a grace period extension for a single test opportunity or multiple opportunities.</td>
<td></td>
</tr>
<tr>
<td>Breaks</td>
<td>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. An example is the English Language Proficiency Assessment in , which has a non-speaking and speaking segment.</td>
<td></td>
</tr>
</tbody>
</table>
Exhibit II.B.2-7: Available Accessibility Features on AIR’s System (continued)

<table>
<thead>
<tr>
<th>Accessibility Feature</th>
<th>Description</th>
<th>AIR System Capability and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text-to-speech:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>directions, passages, items</td>
<td>Computer reads text and graphics aloud on directions, passages, and items.</td>
<td>Currently available</td>
</tr>
<tr>
<td><strong>Text-to-speech:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>graphic description</td>
<td>Computer reads graphics and tables aloud.</td>
<td>Currently available</td>
</tr>
<tr>
<td><strong>Color overlay</strong></td>
<td>Any color can be laid on the screen. This persists throughout the test.</td>
<td>Currently available</td>
</tr>
<tr>
<td><strong>Reverse contrast</strong></td>
<td>The background turns to black while text turns to white.</td>
<td>Currently available</td>
</tr>
<tr>
<td><strong>Line reader</strong></td>
<td>This feature will allow a student to track the line he or she is reading.</td>
<td>Under development; scheduled to be available winter 2013</td>
</tr>
<tr>
<td><strong>Highlighter</strong></td>
<td>The student can select any text to highlight. This persists throughout the test.</td>
<td>Currently available</td>
</tr>
<tr>
<td><strong>Answer eliminator</strong></td>
<td>The student can eliminate any multiple-choice option, whether in text or in a graphic. This persists throughout the test.</td>
<td>Currently available</td>
</tr>
<tr>
<td><strong>Refreshable Braille/tactile with external embosser printer</strong></td>
<td>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.</td>
<td>Currently available</td>
</tr>
<tr>
<td><strong>Magnification</strong></td>
<td>The student can zoom and un-zoom the entire page. This persists throughout the test.</td>
<td>Currently available (An enhancement under development is scheduled to be available fall 2013.)</td>
</tr>
<tr>
<td>Accessibility Feature</td>
<td>Description</td>
<td>AIR System Capability and Comments</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Speech-to-text</td>
<td>Speech will be converted to text, and then saved in the database.</td>
<td>Under research</td>
</tr>
<tr>
<td>Fixed form paper: regular print</td>
<td>These are fixed paper tests on regular print.</td>
<td>Currently available</td>
</tr>
<tr>
<td>Fixed form large print</td>
<td>These are fixed paper tests on large print.</td>
<td>Currently available</td>
</tr>
<tr>
<td>Fixed form Braille/ tactile graphics</td>
<td>These are fixed Braille paper tests.</td>
<td>Currently available</td>
</tr>
<tr>
<td>Sign language: directions, passages, items</td>
<td>These consist of recorded videos in sign language. Experts in hearing impairment do not recommend avatars because they do not translate well into American Sign Language.</td>
<td>Currently available</td>
</tr>
<tr>
<td>Translations (all but ELA content)</td>
<td>Versions in alternate languages</td>
<td>Currently available</td>
</tr>
</tbody>
</table>

**Rehabilitation Act Section 508 Compliance**

Section 508 focuses on removing barriers in information technology and providing the same opportunities to general education as to students with disabilities. Specifically, the section’s technical standards recognize that software applications should be usable by people who are visually impaired. AIR’s test delivery system conforms to Section 508 through the following accessibility features:

- 100% keyboard navigability
- Alternative input devices, such as switch arrays
- Interface elements that move focus as the student navigates the test

PNP has not been implemented on any statewide platform, and we do foresee challenges. Specifically, some accommodations are not compatible with each other. PNP currently contains no standard for cross-validation of feature compatibility. Also, given the access design requirements of some features, some of the PNP specifications may have to become more constrained to be universally supportable.

Over the next 12–24 months we will likely move toward this standard. Our initial move will probably implement packages of features, rather than individual features. We will work with USOE as we move toward this standard.
### Exhibit II.B.2-8: Additional Accessibility Features Available in AIR’s System

<table>
<thead>
<tr>
<th>Accessibility Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio files for instructions, items, and other user support materials</td>
<td>This feature includes audio files for elements of the test such as items and other support materials.</td>
</tr>
<tr>
<td>Secure passage print facility</td>
<td>A visual accessibility feature, this 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; 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.</td>
</tr>
<tr>
<td>Test pauses and restarts</td>
<td>An attention accessibility feature, the test can be paused at any time and restarted and taken over many days. So security is not compromised, visibility on past items is not allowed when the test has been paused more than a specified period of time.</td>
</tr>
<tr>
<td>Mark-and-return can be used for item review</td>
<td>This is an attention accessibility feature; students can flag an item so they can review it.</td>
</tr>
<tr>
<td>Item notes</td>
<td>This is an attention accessibility feature; it allows students to jot down ideas about items or passages.</td>
</tr>
<tr>
<td>Writing checklists</td>
<td>An attention accessibility feature generally for essay items, the writing checklist enables a student to check off writing guidelines from list.</td>
</tr>
<tr>
<td>Review test</td>
<td>The test can be reviewed before ending it.</td>
</tr>
<tr>
<td>Area boundaries</td>
<td>This is an agility accessibility feature; area boundaries for mouse-clicking multiple-choice options mean students can click anywhere on the selected response text or button.</td>
</tr>
<tr>
<td>Language</td>
<td>Any language that is necessary can be supported.</td>
</tr>
<tr>
<td>Help section</td>
<td>This is a reference feature; the Help section explains how the system and its tools work.</td>
</tr>
<tr>
<td>Practice tests and tutorials</td>
<td>This is a reference feature; practice tests and tutorials familiarize students with the online testing system.</td>
</tr>
<tr>
<td>Adaptive test large print</td>
<td>An adaptive test can be administered for large print using the secure passage print facility.</td>
</tr>
<tr>
<td>Adaptive test Braille/tactile graphics</td>
<td>An adaptive test can be administered for Braille using the secure passage print facility. Because the adaptive test is not an RFP requirement, AIR has priced it separately; this information can be found in Section V.D.</td>
</tr>
</tbody>
</table>
Consistent use and meaning of icons and colors throughout the system (for example, the color red is used to indicate a Stop Session—when the proctor stops a session—and an End Test for the student)

A variety of foreground and background colors can be made available, provided they do not conflict with test content

Also, our system includes text-to-Braille fixed tests or adaptive tests (with an additional price option) that can be made available to students with visual impairments, and our screen-reader-optimized Accessibility Interface.

**Web Content Accessibility Guidelines 2.0**

Web Content Accessibility Guidelines (WCAG) focus on making content accessible to a wider range of audiences, primarily students with disabilities including visual impairments. AIR’s system conforms to many of these standards by having

- text alternatives such as large print, text-to-Braille, and text-to-speech;
- adaptable item layouts that best fit each individual item;
- 100% keyboard navigability;
- navigable interfaces by focusing on active elements of the page and having simple, easy-to-use [Back] and [Next] buttons; and
- a common code base compatible with current and future user agents.

**Utah’s Current Accessibility Requirements**

Utah’s accessibility requirements recognize accommodations in the areas of presentation, response, setting, and timing/scheduling that provide equitable access to students with disabilities. AIR conforms to these areas by having the following features present in the system:

- Presentation: text-to-speech and text-to-Braille for items, directions, and passages; large print; magnification tool
- Response: calculation devices; recording device at the item level; graphic organizers
- Setting: only appropriate tools are accessible for students to reduce distractions; “active” elements on the page have a focus indicator
- Timing/Scheduling: tests can be paused and restarted within a specified period of time

In addition, the requirements state that “…providing students with accommodations that are not truly needed may have a negative impact on performance.” Thus, AIR has designed the system so that each tool or embedded support can be in one of four different configurations: (a) available to all students; (b) assigned to students in advance by the designated state, district, or school administrators or through data upload; (c) assigned to students at testing time by the proctor; or (d) configurable as a statewide choice for USOE. We will work closely with USOE to determine the appropriate configuration for each accessibility feature.

**Access by Design, Accessibility for All Students, and Accessibility for Some Students**

We discussed access by design in the introduction to this section. We identified the ways in which our item development will take place with an eye toward supporting accessibility features offered in the system. We also discussed our approach to universal design and customization of tests for individual students. As we mentioned above, each of our features can be configured for universal or restricted access.
This year, AIR introduced *individualized item pools* that take each student’s access restrictions and produces a custom filter on the pool in real time when selecting items.

**II.B.3. Annual Analysis, Review, and Revisions**

Each year AIR will review the item pools for the summative and interim assessments and recommend targets for new item development and acquisition. This annual analysis is part of a continuous improvement process and part of a process to replace items that have been removed temporarily or permanently from the pool. Through continuous planning and item development and acquisition we can improve

- the precision of measurement for students at the tails of the distribution;
- the depth of knowledge assessed;
- item exposure rates; and
- the richness and engagement of students during the testing period.

To achieve these goals, we must know which types of items would improve the pool in terms of content, depth of knowledge, and item types. To know this, we must know

- which items are overexposed;
- which standards tend to be tested with items that are too difficult or too easy for some students; and
- which standards are tested with a different depth of knowledge across the difficulty range.

Perhaps surprisingly, we are able to know this *before the testing season begins*. In Section I.C.3 we discuss our simulator. Prior to deploying the assessments each year, we run simulations evaluating the interaction between the item pool and the adaptive algorithm. One purpose of these simulations is to fine-tune the algorithm. A second purpose is to gather data on the functioning of the item pool. From these simulations we can gather the information needed to plan the year’s development and acquisition.

Analysis and planning for the year’s development can occur in the summer when the pre-deployment simulations are run. This coincides with the beginning of a typical development cycle, in which internal development and review occur over the school year. Post-workshop item writing assignments can go out in late summer, and items can pass through review processes over the course of the school year.

**II.B.4. Technology Requirements**

We propose to develop all our items in ITS or import them into ITS, which can import and export QTI 2.1-compliant XML. The items will be importable from, or exportable to, any system that adheres to this standard. We note that we are bidding the full scope of this contract (all five parts), and can warrant the compatibility of items with our delivery system.
III. Formative Assessments
III.B.1 Overall Approach

Background

AIR’s online instructional support system, Learning Point Navigator (Navigator), offers resources to support teachers and students as they strive for success. Navigator provides a complete formative system, including lesson plans and instructional resources, as well as formative assessments. The system incorporates feedback to students and their teachers, as well as the ability to compile a portfolio of work for students over time. Navigator

- serves as an instructional system where teachers and students can find standards-aligned resources such as assignments, activities, and lessons linked with various learning modalities to enhance student learning;

- guides instructional decisions by providing teachers access to roster and individual student score reports with detailed information about student performance;

- supports differentiated instruction by providing teachers access to materials for individual students based on performance data on each benchmark;

- empowers students to manage their own progress based on individual score reports and feedback, and guide their own learning by providing access to instructional resources based on areas of strength and weakness; and

- serves as a formative assessment system for both teachers and students with access to score reports and feedback.

Learning Point Navigator will serve as USOE’s formative assessment system and will present items in a manner that matches the interim and summative assessment system. AIR’s test delivery system and Navigator share a renderer that displays items and manages student interactions. The common renderer ensures that the formative, interim, and summative assessments all share a consistent look and feel.

Navigator is even more closely integrated with the interim and summative system than simply sharing a look and feel. As students complete interim and summative assessments, the data flow to the reporting system. This system can deliver information about how classrooms or other groups of students do on particular topics (standards or benchmarks), and enables teachers to navigate directly from those reports to relevant resources for their class in Navigator. The real-time interaction between our online reporting system and Navigator turns test results into actionable results to differentiate and improve teaching and learning.

Navigator Content Model

Learning Point Navigator’s content is designed to support both teachers and students. From its very inception, the system has recognized that students learn not only at different rates but through different mechanisms. Each activity in Navigator is tagged to indicate the learning style, or modality, that it is designed to support. We are currently enhancing the system to gather information about each student’s learning style to recommend activities that will be best matched to individual students.

The system seeks to empower students to take responsibility for their own learning. When working on assignments, students can access information about precursor topics, empowering them to catch up in a natural way in the course of completing grade-level assignments. As we illustrate below, students who get stuck when working on an assignment can navigate to precursor topics that may help them fill gaps needed to complete the current assignment.
Each piece of content is classified along three dimensions:

1. Use type
2. Content category
3. Learning modalities

Each classification helps teachers find the tools they need to differentiate instruction for their students.

**Use Types**

Three use types are defined in Navigator:

1. **Meta-instruction/professional development materials** are used by teachers to increase pedagogical knowledge, content knowledge, or pedagogical content knowledge.

2. **Lesson plans** are used by teachers with students or groups of students. Lesson plans encompass a variety of components for use in the classroom and may include references to additional student activities to be completed online.

3. **Activities and assignments** are used by students and may provide instruction, practice, or additional assessment.

**Content Category**

Content category, broadly defined, is simply the subject area(s) into which a particular Navigator resource falls. The system is currently populated with resources in the fields of ELA, mathematics, and science. Most resources are classified by only one content category, others by more than one. For instance, a science activity might also be tagged as mathematics because it addresses measurement standards. Another resource might be classified as ELA and science because of the technical topic addressed. Content classification is further delineated through the use of standard maps that organize the resources by appropriate grade levels, standards, and benchmarks.

**Learning Modalities**

As mentioned above, AIR classifies all resources in Navigator according to learning modalities. Learning modalities represent a combination of presentation, processing, and production styles. Each resource may receive multiple classifications in a taxonomy that represents a composite of Gardner’s work on multiple intelligences (Gardner, 1999), Fleming’s work on learning styles (Fleming, 1992; Fleming & Baume, 2006), and Higbee’s work on perceptual preferences (Higbee, 1991). The taxonomy classifies each resource as matching one or more of the following modalities:

- **Visual**—for learners who best absorb information presented visually, processing with their eyes. A visual learner takes in information more readily in picture or diagram form and may appreciate the use of color to help distinguish categories. Visual learners might convey information more efficiently and effectively through creative or artistic presentations.

- **Aural/Auditory**—for learners who best absorb information presented aurally. Preferences for receiving information might include lecture or audio presentations.

- **Reading**—for learners who best absorb information presented in print. While reading and writing are often categorized together, processing issues (learning disabilities) might arise in one and not the other so they are separated here.

- **Writing**—for learners who process information through active writing, whether taking notes from reading or lectures or summarizing information in written form.

- **Kinesthetic/Proprioceptive**—Proprioception is body sense and orienta-
tion, with kinesthesia placing emphasis on motion, and is for learners who best absorb information when active participation of some sort is involved. Activities that involve movement and exercise as well as body awareness and orientation appeal to these learners.

- **Haptic**—for learners who best absorb information when tactile stimulation is included. These are the students who doodle, needing touch to deeply incorporate concepts. Inclusion of manipulatives can help these learners.

- **Interpersonal**—for learners who absorb information when they are actively engaged with other learners. Interpersonal learners do best when speaking is involved as in group or cooperative work or oral processing and presentations; these are interactive learners.

- **Intrapersonal**—for learners who best absorb information when they are involved in personal reflection, sometimes in writing. Intrapersonal learners often engage in metacognition to deepen/internalize their learning.

- **Musical**—for learners who best absorb information when it is accompanied by or set to music. These learners make connections to underlying prosodic structure (stress and intonation in speech patterns) to enhance their learning. Musical learners might learn or create a song to internalize concepts.

- **Naturalistic**—for learners who best absorb information when it is tied to natural cycles and the environment around them.

- **Creative**—for learners who respond to problems in new ways and are most responsive in environments where open-ended questions and opportunities for individual expression are included.

Note that these learning modalities are not mutually exclusive. Consideration of learning modalities will offer teachers the opportunity to broaden their repertoires to reach more students by engaging a wider array of learning modalities. Over time, student portfolios may reveal the modalities that prove most effective for individual students.

## Navigator Content Libraries and Teacher-Created Resources

Navigator’s atomic units of content are called *activities*. An activity may be an assessment item, an instructional document, a video, a game, or any other unit of content. Activities are assembled into *resources*—for example, an instructional activity followed by a quiz. Educators may assemble activities or resources into *assignments*. Each formative assessment, for example, will be an assignment (the assessment) made up of activities (the items).

Navigator supports the same range of automated scoring options as our test delivery system. We can score

- graphic response items, including all the typical selected-response TE item types but also more open constructed-response drawn responses;
- propositional responses, English language responses in which the correct answer or answers can be represented as a logical set of propositions;
- equation responses, in which the correct response is a mathematical expression or equation;
- essay responses, in cases in which computer-readable training sets are available; and
- simulation interactions.

Learning Point Navigator is designed to hold multiple content libraries. For Utah, we propose to include two libraries:

- AIR’s open-source library of core content in ELA, mathematics, and science
A library of Utah-specific content imported from the UTIPS system

The AIR core content is a growing pool of resources, currently including more than 3,900 unique activities compiled into approximately 1,900 resources across grades and subjects, and projected to reach 14,000 resources by January 1, 2014. These resources are provided to all Navigator clients at no charge.

Navigator allows activities to be grouped as a single resource (at publication time) or as assignments (by teacher at assignment time). These capabilities provide a powerful ability to group activities (such as items) under common stimuli, under common instruction sets, or in any other grouping desired.

We also plan to import the high-quality, aligned content from the current UTIPS system and deliver it in a format consistent with the interim and summative assessments. We plan to augment this library throughout the life of the contract with items contributed by Utah teachers. We discuss our proposed process for teacher-contributed resources below.

We propose to have Utah educators develop formative activities in our content management system (ITS, which is discussed at length above). Exhibit III.B.1-1 illustrates our proposed process for supporting the extension of the Utah library with teacher-created formative resources. Teachers will draft items in our ITS. ITS incorporates a word processing program (Open Office, which is similar to Microsoft Word but is open source). Educators will enter resources into Open Office templates. ITS can be configured to support a robust review process to ensure that each activity benefits from input from other educators. ITS will support Utah educators as each activity works its way through a Utah-defined review process in which Utah educators review and approve activities.

Teachers will be able to upload media files to ITS, including graphics, audio, and spreadsheets. ITS will allow teachers to enter test items and upload graphics and files individually. In conjunction with this feature, Navigator

Creating Teacher-Contributed Items

Navigator is designed to deliver high-quality, finished resources to educators and students. We recognize the experience of Utah’s teachers in developing such high-quality resources. Here, we describe our plan to continue and support that tradition.

We propose to have Utah educators develop formative activities in our content management system (ITS, which is discussed at length above). Exhibit III.B.1-1 illustrates our
will render these file formats to be accessible on the teacher and student interfaces.

AIR will publish approved activities to Navigator’s Utah library twice per year. Once this content is published, students and teachers will be able to access it in the Utah Core library in the same way they access content from the other libraries in Navigator. AIR proposes that resources added to Navigator fall under the Creative Commons Attribution Only License.

Teachers will create formative assessments using items in the Utah Core library and the other libraries available through Navigator. This process is described in the Creating Assignments and Assessments section.

### Feedback for the State

To gauge the usage of Navigator among students and teachers, we propose to provide USOE with statistics outlining the frequency and number of teachers and students accessing Navigator, either to create, manage, or complete assignments. Reports can be delivered monthly, at the state level, and further, by school and district, with cumulative YTD data for the school year.

### Professional Development Workshops

AIR proposes to offer two training workshops on writing and reviewing items. These one-day workshops, which will host up to 40 educators each, will be designed to train educator-leaders in a train-the-trainer model. During the workshops educators will learn to operate ITS, review the Utah Core Standards, and receive training on content development.

### III.B.2. Educator and User Functionality

#### Navigator, from the Teacher’s Perspective

Learning Point Navigator gives teachers a powerful tool to manage and enhance classroom instruction and student learning. The teacher’s interface enables easy access to instructional materials; facilitates assignment delivery and management, including for formative assessments; and ensures instant feedback on student progress.

#### Finding Material

Teachers will be able to access Navigator directly from USOE’s portal and browse through an easy-to-navigate content library sorted by grade, subject, strand, and topic area. As teachers find relevant materials, they can flag as a “favorite” each item for immediate or future use. Favorite items will be added to a personalized library of resources, which teachers can access and organize at any time. As teachers identify favorite materials, the system will automatically generate a list of associated precursors and follow-ons as suggestions for resources that might be used before and after the selected resource is used.

#### Creating Assignments and Assessments

Teachers can create, manage, and view all assignments through the assignment page in Navigator. To build and customize an assignment, teachers can either access this page or simply click the “create assignment” icon while browsing for resources.
Through Navigator, Utah teachers can create formative assessments in the same manner as creating assignments with learning resources from content libraries. They can administer assessments at any point during the year. By searching items by grade, subject, content standard, and content library, teachers can align assessments to Utah Core Standards and incorporate multiple assessment items into assignments. Navigator supports multiple-choice and constructed-response items, including graphic response, natural language, and written-response items, from which teachers can create assignments to give to students.

Navigator will prompt teachers to create a unique name for the assignment, designate a start and end date for students, and enter any special comments or instructions for students (see Exhibit III.B.2-1). Any comments or instructions that teachers enter will be displayed for students before they begin the assignment. After an assignment is created, the system allows teachers to add resources or items easily and choose the sequence in which they will appear to students. Once teachers are satisfied with the content of the assignment, the system will automatically populate the screen with the names of students in each class and allow teachers to select individual students, an entire class, or a combination of class(es) and individual students to complete the assignment (see Exhibit III.B.2-2). Teachers can view and progress through the assignment.
Exhibit III.B.2-2: Assigning Resources to Students by Selecting an Entire Class, Individual Students, or a Combination of Class(es) and Individual Students: Teacher’s View

One roster has been selected, with the option to assign to additional rosters, individual students, or a combination of rosters and individual students to complete the assignment.

exactly as it will appear for students before sending it out. At any point before or after the assignment is sent to students, teachers can return to edit any of its features.

**Monitoring Progress on Assignments and Assessments**

Learning Point Navigator presents teachers with an easy method to monitor students as they complete assignments, including assessments. Teachers can access the assignment page to monitor progress on active and completed assignments. For active assignments, the system will display the percentage of the assignment that has been completed for each student and the average percentage completed for the entire class. For completed assignments, the average performance is displayed. Teachers receive immediate real-time results at the class or student level as students complete their work, as shown in Exhibit III.B.2-3.

**Reviewing Student Work**

After assignments are completed, teachers can view student work to determine next steps for each student. The results of each assignment will indicate the extent to which the student has mastered the concept and the need for further instruction.

Navigator presents teachers with access to real-time assignment score reports both for individual students and entire class rosters. More specifically, teachers can create and
Teachers can monitor students’ progress on active assignments, including assessments. This screen indicates the student has not yet started the assignment.

Teachers can jump directly to a view of a student’s specific responses on each activity in an assignment.

manage online rosters, view individual student and roster reports in table format, and filter score reports by student subgroup based on how they created their rosters.

Score reports will include a roster with student names, scores for each assignment, and an indication the assignment was completed or is pending completion. As displayed in Exhibit III.B.2-4, a summary of students’ performance, including each student’s raw score, overall percent correct, and a breakdown of performance for each resource or item, will be presented. In conjunction with the interim and summative assessment system, teachers will be able to find students who are struggling with certain standards and topic areas. A standards mastery report with a breakdown of performance on each benchmark within the assignment will be presented as well, for both entire roster and individual student score reports (see Exhibit III.B.2-5).

Teachers will have access to student portfolio reports that display individual student responses to multiple-choice items, the percent of students in the class who chose each option (with the ability to view those students), and the percent of students in the state who chose each option (see Exhibit III.B.2-6). Student portfolio reports will also display individual student responses to constructed-response items, including graphic-response, natural-language, and written-response items.

Navigator will provide reports with benchmark information at aggregate levels (e.g., class, group of students). This information can be presented in terms of relative strengths and weaknesses to convey to teachers that students in a class are performing better, worse, or the same on a benchmark compared to the assignment as a whole.

Teachers can track the progress of individual students and view an outline of their perform-
Exhibit III.B.2-4: Score Report Summary: Teacher’s View

Teachers can monitor students’ progress on active assignments, including assessments. This screen indicates the student has not yet started the assignment.

Teachers can jump directly to a view of a student’s specific responses on each activity in an assignment.

Exhibit III.B.2-4: Score Report Summary: Teacher’s View

Teachers can monitor students’ progress on active assignments, including assessments. This screen indicates the student has not yet started the assignment. Teachers can jump directly to a view of a student’s specific responses on each activity in an assignment.

Accessing Lesson Plans and Meta-Instruction

Teachers will have access to lesson plans and meta-instruction/professional development materials to enhance their pedagogical skills. Some of the meta-instruction/professional development materials will address research-based best practices of teaching in general, but the majority of these resources and all the lesson plan materials will be aligned to specific content standards, strands, and benchmarks. This will be organized through the content standard maps. In addition, meta-instruction and lesson plans can be filtered by learning modalities should teachers desire to incorporate more of a specific learning modality into instruction.

Student and Parent Interface

Learning Point Navigator gives students easy access to resources and activities for formative assessments, extra practice, and instruction.

As mentioned in the introduction, Navigator grows from the belief that students can, should, and will take charge of their own learning. As students are working through assignments, they may find themselves struggling. They can easily see a graph of precursor skills and, with a click, navigate to instructional material and practice activities to help them learn the precursor material they need to complete the current assignment. Exhibit III.B.2-7 shows a screen through which students navigate among related topics.

Likewise, students might be successful with a particular topic and be interested in further
Exhibit III.B.2-5: Standards Mastery Report with a Breakdown of Performance on Each Benchmark in the Assignment: Teacher’s View

Assignments > Report: Roster 2

<table>
<thead>
<tr>
<th>Score</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>29/72</td>
<td>40%</td>
</tr>
</tbody>
</table>

Mathematics

- Grade 1
  - 2. Algebra
    - 1.2.1: Recognize and create patterns; use rules to describe patterns.
    - Score: 9/24 (38%)
  - Score: 9/24 (38%)

- Grade 3
  - 2. Algebra
    - 3.2.1: Use simple operations input-output rules to represent patterns and relationships and to solve real-world and mathematical problems.
    - 3.2.1.1: Create, describe, and apply simple operations input-output rules involving addition, subtraction, and multiplication to solve problems in various contexts.
      - Score: 9/24 (38%)

- 4. Data Analysis
  - 3.4.1: Collect, organize, display, and interpret data. Use labels and a variety of graphs and units in displays.
    - Score: 7/24 (29%)

Utah Core Standards would appear here.

Teachers have the option to include scores for unanswered items when viewing benchmark-level data.

Teachers can view a breakdown of performance on each benchmark within the assignment.

Completing Assignments and Assessments

Just as teachers can create and monitor assessments in the same manner as other assignments, students can access and complete assessments in the same manner as other assignments. To begin work on an assignment designated by their teacher, students log on to Navigator. They can immediately view assignments waiting to be started and completed. Students can easily organize and sort all assignments according to start date, subject, teacher, and completion status. To work on an assignment, students need only to select either “start” or “continue.”

Student Work Portfolio

Through Navigator, students and parents can not only complete assignments but also manage and access their own assignments (including assessments) and learning resources. The Student Work Portfolio allows students to review past assignments, view a collective record of their score reports, and explore supplemental resources.
After completing an assignment, students can view their own score report with a summary of their performance on each resource included in the assignment, as displayed in Exhibit III.B.2-8. Similarly, with assessments, students can view their own data and reports summarizing their performance. Furthermore, students can search for and view standards-aligned resources for supplemental instruction and practice. They can search for resources by benchmark and view related content for each benchmark with associated precursors and follow-ons. Instructional resources may also contain embedded assessments to further gauge student learning and progress. By having access to their own score reports and the ability to search for supplemental resources, students can track their progress over time and guide their own learning.

Involving parents/guardians as participants in student learning is an important part of improving education. Navigator provides the opportunity to share information with these stakeholders, provided that their children have been active in the system. Through this view, parents/guardians can access progress on assignment completion as well as performance on completed assignments. Just as teachers can drill down to more specific detail for a particular student through assignments or benchmarks, so, too, can parents/guardians. Parents/guardians can have substantive discussions with their children concerning progress in their learning. Navigator also facilitates discussion between teachers and parents/guardians because both groups have access to the same data about student progress and performance.
Exhibit III.B.2-7: Searching for Resources by Benchmark with Precursors and Follow-ons: Student's View

Students can navigate to precursor skills to help them complete their current assignment.

Students can navigate to subsequent topics if they would like to further enhance their skills on a particular benchmark.

Exhibit III.B.2-8: Searching for Resources by Benchmark with Precursors and Follow-ons: Student's View

My Assignments (0) Completed Assignments (1)

Completed Assignments

This page shows the assignments you have finished:
- Click [View Assignment] to see the assignment.
- Click [Show Scores] to see how you did.

Overall percent correct on this assignment.

Students can view their completed assignment and see feedback for the items they answered correctly or incorrectly.

<table>
<thead>
<tr>
<th>Start Date</th>
<th>Due Date</th>
<th>Assignment Title</th>
<th>Score</th>
<th>Assigned by</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/18/2011</td>
<td>12/28/2012</td>
<td>Geometry_1</td>
<td>31%</td>
<td>Teacher_1</td>
<td>Mathematics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activity</th>
<th>My Score</th>
<th>Maximum Possible Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical Angles</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Unit Conversions</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Volume Word Problems</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Area of Rectangles</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>
### III.B.3 and III.B.4 System Functionality and Additional Requirements

<table>
<thead>
<tr>
<th>Educator Functionality</th>
<th>USOE Requirements</th>
<th>AIR Proposal to Meet Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construct assessments which can be administered at any point during the year</strong></td>
<td>• Teachers will have access to assessment items through Navigator, which they can use to create assessments at any point during the year.</td>
<td>• Teachers will assign item resources by selecting a start date and due date and specifying individual students and/or entire rosters to which to assign items.</td>
</tr>
<tr>
<td><strong>Construct assessments which can be</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Be aligned by the teacher to the Utah Core Standards for any subject and content area K-12, including subjects not assessed via the summative or interim systems (e.g., CTE, health)</td>
<td>• Teachers will be able to</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• include links to websites</td>
<td>• align assessments to the Utah Core Standards by searching items by grade, subject, content standard, and content library. Teachers can incorporate multiple assessment items into assignments.</td>
</tr>
<tr>
<td></td>
<td>• include an upload of media (e.g., graphics, audio, spreadsheets)</td>
<td>• include links to websites when uploading content in ITS. A more robust model is to include the media themselves.</td>
</tr>
<tr>
<td></td>
<td>• Be shared with educators from their grade, department, school, or statewide as appropriate, with the ability for other users to administer, copy, and change items</td>
<td>• upload media files to ITS, including graphics, audio, and spreadsheets. Navigator will render these file formats to be accessible on the teacher and student interfaces.</td>
</tr>
<tr>
<td></td>
<td>• Be grouped together under a common stimulus (e.g., passage, graphic)</td>
<td>• access assessment activities following the review process. After teachers draft items in ITS, other educators will review and approve the activities, which will then be published to Navigator twice per year for teacher use.</td>
</tr>
<tr>
<td><strong>Construct assessments using a combination of USOE/vendor items and educator’s own items</strong></td>
<td>• Teachers will develop formative assessment items in ITS.</td>
<td>• access activities in Navigator that are grouped as a single resource (at publication time) or as assignments (by the teacher at assignment time). Activities (such as items) can be grouped under common stimuli, common instruction sets, or in any other grouping desired.</td>
</tr>
<tr>
<td><strong>Allow educators to load tests into the software from a word processing format and enter test items and upload graphics/files individually</strong></td>
<td>• Teachers will create formative assessments using items in the Utah Core library and the other content libraries available through Navigator. These libraries include:</td>
<td>• ITS will allow teachers to enter test items and upload graphics and files individually.</td>
</tr>
<tr>
<td></td>
<td>• AIR’s open-source library of core content in ELA, mathematics, and science;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• resources from CTB/McGraw-Hill’s Acuity collection in ELA and mathematics; and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• a library of Utah-specific content imported from the UTIPS.</td>
<td></td>
</tr>
</tbody>
</table>
## System Functionality

<table>
<thead>
<tr>
<th>USOE Requirements</th>
<th>AIR Proposal to Meet USOE Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Calculate scores for selected response and short answer questions</td>
<td>• Navigator calculates scores for graphic-response items; propositional responses; equation responses; essay responses; and simulation interactions.</td>
</tr>
<tr>
<td>• Provide real-time reports for each assessment, filterable by class period, teacher, school, and LEA as appropriate</td>
<td>• Navigator allows educators to access real-time assessment score reports both for individual students and entire class rosters; and create and manage online rosters, view individual student and roster reports in table format, and filter score reports by student subgroup based on how the rosters were created.</td>
</tr>
</tbody>
</table>
| • Provide reports that aggregate and disaggregate student results according to specified demographic data and include:  
  - A roster to include student names, scores, time stamps, and the indication that extended response items need to be scored  
  - A distribution report outlining the number of students within specified benchmarks, with the ability to view those students  
  - An item analysis report showing the percentage and names of students who chose each option and individual student responses to extended response items  
  - A standards mastery report indicating student understanding of each content standard | • Navigator will provide score reports that include a roster with student names, scores for each assignment, and an indication the assignment was completed or is pending completion; score reports that allow teachers to find students who are struggling with certain standards and topic areas; student portfolio reports displaying individual student responses to multiple-choice items, the percent of students in the class who chose each option (with the ability to view those students), and the percent of students in the state who chose each option; and individual student responses on constructed-response items. standards mastery reports with a breakdown of performance on each benchmark within the assignment, for both entire roster and individual student score reports; and the ability to export roster and student-level summary reports to a CSV file. |
| • Specify the sources of items (e.g., released or retired summative, teacher created, vendor provided) | • Sources of items will be specified through content libraries set up for USOE on Navigator. Educators will have the ability to search items by selecting the appropriate content library. |
| • Specify the types of items supported (e.g., selected response, constructed response) | • Navigator supports multiple-choice and constructed-response items, including graphic response, natural language, and written-response items, from which teachers can create assignments to give to students. |
### System Functionality

<table>
<thead>
<tr>
<th>USOE Requirements</th>
<th>AIR Proposal to Meet USOE Requirements</th>
</tr>
</thead>
</table>
| • Describe the degree to which automated scoring is supported                     | • Navigator supports scoring for  
  • graphic-response items, including all the typical selected-response TE item types, but also including more open constructed-response drawn responses;  
  • propositional responses, English language responses in which the correct answer or answers can be represented as a logical set of propositions;  
  • equation responses, in which the correct response is a mathematical expression or equation;  
  • essay responses, in cases in which computer-readable training sets are available; and  
  • simulation interactions.                                                                                                             |
| • Mirror defined summative blueprints                                             | • Navigator will serve as USOE’s formative assessment system and will present items in a manner that matches the interim and summative assessment system.  
  • AIR’s test delivery system and Navigator share a renderer that displays items and manages student interactions.  
  • AIR’s reporting system  
    • receives data from interim and summative assessments;  
    • delivers information about how classrooms or other groups of students are doing on particular topics (standards or benchmarks); and  
    • enables teachers to navigate directly from those reports to relevant resources for their class in Navigator.  
  • AIR plans to import high-quality, aligned content from the current UTIPS system and deliver that content in a format consistent with the interim and summative assessments. |
| • Provide students with their own data and reports, over the course of the school year | • Navigator’s Student Work Portfolio enables students to  
  • complete assignments;  
  • manage and access their own assignments (including assessments);  
  • review past assignments;  
  • view a collective record of their score reports; and  
  • explore supplemental resources.  
  • Score reports will allow students to view a summary of their performance on resources included in an assignment.  
  • Students can search for and view standards-aligned resources by benchmark and view related content with associated precursors and follow-ons for supplemental practice and instruction. |
<table>
<thead>
<tr>
<th>System Functionality</th>
<th>AIR Proposal to Meet USOE Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>USOE Requirements</strong></td>
<td><strong>AIR Proposal to Meet USOE Requirements</strong></td>
</tr>
<tr>
<td>• Provide students with their own data and reports, over</td>
<td>• Navigator’s Student Work Portfolio enables students to</td>
</tr>
<tr>
<td>the course of the school year</td>
<td>• complete assignments;</td>
</tr>
<tr>
<td></td>
<td>• manage and access their own assignments (including assessments);</td>
</tr>
<tr>
<td></td>
<td>• review past assignments;</td>
</tr>
<tr>
<td></td>
<td>• view a collective record of their score reports; and</td>
</tr>
<tr>
<td></td>
<td>• explore supplemental resources.</td>
</tr>
<tr>
<td></td>
<td>• Score reports will allow students to view a summary of their performance on resources included in an</td>
</tr>
<tr>
<td></td>
<td>assignment.</td>
</tr>
<tr>
<td></td>
<td>• Students can search for and view standards-aligned resources by benchmark and view related content</td>
</tr>
<tr>
<td></td>
<td>with associated precursors and follow-ons for supplemental practice and instruction.</td>
</tr>
<tr>
<td>• Return usage statistics to the USOE outlining how many</td>
<td>• AIR proposes to provide USOE with usage statistics outlining the frequency and number of teachers and</td>
</tr>
<tr>
<td>tests are taken by school and LEA, monthly and annually,</td>
<td>students accessing Navigator, either to create, manage, or complete assignments.</td>
</tr>
<tr>
<td>as well as other reports that may be requested by the USOE</td>
<td>• Reports can be delivered monthly, at the state level, and further, by school and district, with</td>
</tr>
<tr>
<td></td>
<td>cumulative YTD data for the school year.</td>
</tr>
<tr>
<td>• Generate immediate reports available in a format that</td>
<td>• Navigator</td>
</tr>
<tr>
<td>can be exported by educators, based on role</td>
<td>• generates immediate score reports that are accessible on the teacher interface; and</td>
</tr>
<tr>
<td></td>
<td>• allows teachers to export roster and student-level summary reports to a CSV file.</td>
</tr>
</tbody>
</table>
III.B.5 Technical Requirements

Application Installation and Maintenance

Learning Point Navigator is a 100% web-based solution and requires no client software other than a browser. Navigator will be hosted on servers at Rackspace, our preferred hosting company, as described under Hosting below. Navigator will be available and accessible 24/7, except during pre-negotiated maintenance windows.

All of the production control and documentation policies that apply to our Test Delivery System apply to Navigator, and readers are referred to the corresponding subsection there for details on how we will remediate problems and inform USOE. Similarly, we will negotiate availability service-level agreements with USOE.

Currently, Navigator does not have an interface for states or districts to monitor usage. If these requirements are necessary, AIR will implement them and provide these services to USOE at no additional charge.

Security and Integrity

Navigator is password-protected and operates over secure connections. It is protected by the same set of physical and network security policies, practices, and equipment as the rest of our systems. These security features are detailed in Section I.C.4.b.

Supported Devices and Operating Systems

We meet this requirement fully for the Navigator system. Please refer to Section 1.C.4.c for more detail, with one adjustment: the secure browser is not required for students to access Navigator.

Hosting

We meet this requirement fully for the Navigator system. Please refer to Section 1.C.4.d for more detail.
References


IV. Other Contract Requirements
IV.A. Return and Destruction of Items and Data

AIR acknowledges that any test items developed under this contract or transferred to us or existing in Utah’s test item banks are and will remain the exclusive property of the State of Utah and may not be used for any other purposes (except by express written permission from USOE). We will deliver the full item bank in the agreed-upon format at the conclusion of the contract and annually at USOE request.

AIR will purge our systems of any Utah-owned items, including those developed under this contract, at the conclusion of the contract.

As discussed in detail under Test Delivery, Security and Integrity, we take our responsibilities as stewards of confidential student data seriously. We have physical, network, and software controls in place to ensure that the data remain confidential and secure. We will always maintain the data in ways that are fully consistent with federal and state law, and at no point will we disclose such data without prior written approval.

We will purge our systems of any student data at the conclusion of the contract or at any time USOE instructs us in writing to do so.
IV.B. Timetables

Timelines

AIR will deliver the testing system as specified on the desired timelines. We understand the legislative mandate to administer computer-adaptive tests aligned with Utah Core Standards no later than the 2014–2015 school year, and our proposal is based on a plan to achieve that goal. Our timelines also meet all of the other proposal objectives. Specifically, in this proposal AIR demonstrates both our plan and our ability to have the formative assessment system operational in 2013 and the new Utah system of interim and summative assessments field-tested and operational for the 2014–2015 school year with a bank of new items aligned to the Utah Core.

AIR would prefer to propose a more aggressive timeline and deliver the new Utah interim and summative assessments in just one year, as we did in and . In those cases, we went from contract award to field test in 6 months and launched operational interim and summative tests in less than 12 months. The barrier to this aggressive schedule in Utah is the availability of sufficient items aligned to the Utah Core in every grade and subject. To this end, we have proposed developing for Utah thousands of new items (which Utah will own) and item sharing of Common Core items from other states. We believe this is the strongest set of items available to Utah from any vendor but will require the time between award and through the 2013–2014 year to get the new items through USOE and Utah educator and parent committees.

We propose a plan that allows Utah to shed its current CRT after the spring 2013 administration. In spring 2014, we propose an operational field test—an operational assessment designed to support the testing and calibration of a large number of items, and once calibrated, supports the use of that data for accountability purposes. We have used this approach with great success in other assessment programs.

Below, we answer three key questions about our proposed timeline:

- What are the key milestones?
- What is AIR’s plan to meet these key milestones?
- Why should Utah believe AIR can do this?

Key Milestones

We built our timelines using the following key milestones, which of course are integrated with many interim and succeeding activities:

**Phase I – Contract Award to August 2014**

**Year 1-Award to August 2013**

- Import and align all existing USOE items by April 2013
- Identify and incorporate all items shared by other states under the proposed item sharing agreement by May 2013
- Develop approximately 4,000 new items by July 2013
- Ensure all new items are reviewed and approved by the USOE and committees during summer 2013, when teachers are available

**Year 2-September 2013 to August 2014**

- Launch the Learning Point Navigator Formative Assessment system with UTIPS aligned items and AIR resources by September 1, 2013
Deploy the interim/summative testing system by February 2014, including delivering training

- Conduct the operational interim/summative field test in spring 2014, with the final dates to be mutually agreed by USOE and AIR to meet reporting requirements
- Complete all scoring, standard setting, and reporting activities to allow for accountability reporting based on spring 2014 operational field test.
- Incorporate approximately 1,350 new and shared items for embedded field-testing and ensure all items are reviewed and approved by USOE and committees by August 30

**Phase II**

**Year 3-September 2014 to August 2015**

- Operational formative, interim, and summative testing, with immediate reporting
- Complete all scoring, standard setting, and reporting activities to allow for accountability reporting based on spring 2014 operational field test
- Incorporate approximately 1,350 new and shared items for embedded field-testing and ensure all items are reviewed and approved by USOE and committees by August 30

**Year 4-September 2015 to August 2016**

- Operational formative, interim, and summative testing, with immediate reporting
- Complete all scoring, standard setting, and reporting activities to allow for accountability reporting based on spring 2014 operational field test
- Incorporate approximately 1,350 new and shared items for embedded field-testing and ensure all items are reviewed and approved by USOE and committees by August 30

**Year 5-September 2016 to August 2017**

- Operational formative, interim, and summative testing, with immediate reporting
- Complete all scoring, standard setting, and reporting activities to allow for accountability reporting based on spring 2014 Operational field test.
- Incorporate approximately 1,350 new and shared items for embedded field-testing and ensure all items are reviewed and approved by USOE and committees by August 30

**AIR’s Plan**

In the Project Management Plan and the attached Gantt chart (Appendix F), AIR demonstrates our plan for each task leading up to the operational deployment. We need certain things from USOE. First, and most important to the timeline, we need the current item banks, and we need to work together to develop and finalize test specifications.

With these turned over to AIR, we can immediately begin item development, selection of shared items, and USOE review of items. We will develop a schedule to flow items to USOE to confirm AIR’s alignment to Utah Core. We will work with USOE to schedule the Content, Fairness, Parent, and other reviews prior to the field-testing.

Our schedule assumes that AIR will be responsible for the entire project. We have noted deliverables on the Gantt chart, showing for example that while Content, Fairness, Parent, Data, and other review meetings are included in Part II Summative and Interim Test Content, the standard setting meeting falls under Part I, Summative and Interim Technology.

Our plan for the initial field test includes well more than the minimum number of items needed for the adaptive interim and summative assessments. We did this to provide robust
item banks. If needed, we have the contingency to lower our Phase 1 targets and thus reduce the workload before the initial field-testing. We would shift this work to a future year, and insert the extra items for embedded field-testing once the system is operational. This is not our plan, but an example of mitigation plans we can employ.

Our software systems are operational now. We can configure and deploy our item management system aligned to Utah standards within a week. We have allowed sufficient time to configure the hardware and software to support the Utah testing and reporting environment. The software systems will be ready long before the item development and review processes have been completed. Once ready for the operational field test, we will provide help desk support and monitor system performance during the field-testing.

At the conclusion of the field-testing, we will need to move rapidly to conduct range-finding and rubric validation, complete handscoring, conduct data review, convene standard setting, and load cut scores for final score reporting. We will then prepare technical reports and other documentation for peer review, and prepare for opening the operational testing window in fall 2014. The Utah interim and summative assessment system will be operational.

Utah’s schedule represents some of the most advanced planning we have encountered in state testing programs. As described in our Capabilities section later in this project, AIR has accomplished similar transitions from paper testing programs to field-testing to new operational online adaptive testing programs in less than 12 months. This schedule allows significantly more time.

AIR is proud of our track record and references for meeting deadlines and satisfying contract requirements. Beyond just meeting requirements and deadlines, AIR has a record of extraordinary support, particularly in the transition of new testing programs. Delivering fast transitions with quality results is not unusual in AIR’s experience. During the past seven years, AIR has successfully transitioned more than a dozen separate testing programs in, and from other contractors, and AIR met all key deliverables. As these and many other projects demonstrate, AIR meets our timelines and deliverable requirements, including situations when aggressive timelines for project start-up/transition involve large item development efforts, creating new item banks from other sources, transitioning to entirely new standards or new assessment designs, meeting field-test windows, and delivering online testing systems. We offer this track record as evidence to support our commitment and our plan to meet Utah’s timelines and requirements.
IV.C. Quality Assurance

AIR’s quality assurance process is predicated on the understanding that people make mistakes. Therefore, we exploit every opportunity to automate processes and reduce opportunities for human error. Where automation is not possible or sufficient, we use replication—multiple independent analysts working from detailed specifications each produce the product (e.g., data file, analysis) and compare the results. Under the direction of project management, the analysts work together to resolve discrepancies and note any inconsistencies or missing information from the specifications documents. Where replication is not possible, we use multiple layers of independent review to prevent errors from affecting the program.

Quality assurance processes are integrated into everything that we do, and many of them are described in detail in the description of our plan to complete the work above. Here, we summarize the quality assurance processes that will ensure error-free operation for Utah.

All Steps of the CAT process

Quality assurance steps in the CAT process include the following:

- Quality assurance of the keys and scoring rubrics for the items
- Quality assurance of the display of items
- Pre-deployment testing of the performance of the adaptive engine
- Quality assurance checks during configuration and deployment
- Continuous monitoring of the performance of the adaptive engine
- Quality assurance of human scoring

Once tests are completed, these quality processes give way to the quality processes governing data exchanges, which we discuss in the subsequent section.

Keys and Rubrics

As we describe in Section II.B.2, several content experts review items, their keys, and machine-scored rubrics before publication. This is discussed at length under Item Development for the Summative and Interim Assessments.

Machine-scored rubrics are refined and perfected during our rubric validation process. This process is analogous to rangefinding and is designed to automatically identify samples of anomalous student responses for review. During this review, rubrics are refined to capture correct student responses that may have been overlooked during initial development.

Item Display

Two primary quality assurance processes protect the display of items:

1. ITS displays items using exactly the same code base as the test delivery system. The final review before deployment includes web approval, which requires that the item be viewed as it will appear to the student. Web approval locks the item display and creates the actual displayable files that will be published to the test.

2. During the UAT period we conduct platform review, in which a quality assurance team works together in a computer lab to simultaneously review each item on supported platforms.

Pre-Deployment Testing of the Adaptive Algorithm

As we discuss throughout this proposal, we engage in a robust simulation process prior to the deployment of any test. During that
process, we tune the parameters of the adaptive algorithm to optimize the performance of the engine for the particular item bank and blueprint.

The simulation process gives us very accurate insight into such quality indicators as

- item exposure;
- blueprint match;
- precision of measurement;
- degree of adaptation; and
- scoring accuracy.

**Configuration and Deployment**

ITS conducts numerous integrity checks on the data describing the item pool and blueprint and flags any issues, which are resolved by our test developers. The developers’ configuration tool can also detect and highlight any differences between the configuration that they are building and any past configuration—a capability that helps us ensure that all the necessary revisions are made and no accidental revisions are made.

The publication module compares the item display file to the item in ITS and flags any versions that are not the most current version.

The following steps provide quality assurance during deployment of the test:

- All software deployments or changes must have a release plan that must be approved by our Production Control Board.
- Any changes to any production system must be conducted in pilot/copilot mode, in which one engineer makes the changes while the other engineer inspects the changes and confirms them against the release plan.

- All deployments go through a formal testing cycle, including testing on our test servers, staging and UAT testing, and final shake-out testing when the system is moved to the operational servers.

**Continuous Monitoring during Testing**

Tests administered online are monitored in real time by our Quality Monitor (QM) system. Each completed test runs through QM as soon as the test is complete, where a variety of quality checks is conducted. For example, QM checks attemptedness calculations, raw scores, and other indicators. It periodically aggregates these data to check for items with extreme p-values and identify any items for which the most common response is not also the correct response, looks for item drift, and implements a suite of analyses designed to detect potential cheating or irregularities.

**Quality Assurance of Human Scoring**

Our quality assurance procedures for hand-scoring appear under Section I.C.3.d. These procedures begin with selection and training of quality staff and include continuous monitoring of scoring consistency and accuracy and final summary of key statistics.

The entire scoring process is managed by DRC’s electronic scoring system, which implements many programmatic controls to ensure that each item is double scored and discrepancies are appropriately resolved. The system enables team leaders to call up individual responses, monitor a variety of indicators, and designate items for rescoring. Throughout the scoring, three processes will monitor the validity and reliability of the scores assigned:
Backreading, in which team leaders continuously review the work of the scorers on their teams. DRC will backread 10 percent of the papers scored.

Double scoring, in which two scorers each independently rate each paper. A team leader resolves any discrepancies.

Validity testing, in which at least five validity papers are scored by each scorer each day. A validity paper is a paper that has been scored by expert scorers, who all agree on the score, so the right score is known in advance.

The data generated by these processes will be presented in a set of at least nine reports, as described in Exhibit IV.C-1.

The first set of reports facilitates monitoring and intervention at the system level. The second set facilitates the management and potential remediation of individual readers.

**Data Exchanges**

As described in Section I.C.3.b, data exchanges are governed by data layouts that described the format and syntactic constraints of data to be collected from Utah’s LEAs and delivered to USOE.

Incoming data, including student enrollment, demographics, and accommodation information, are validated by configurable business rules to be determined in collaboration with USOE. Upon upload, the data are validated and invalid data are rejected along with messages that indicate the specific data that are not valid.

---

**Exhibit IV.C-1: Summary of Reliability and Validity Reports Item-Level Reports Aggregated Across Readers**

| Item-level reports aggregated across readers | Reliability report | Spearman correlation between first scores by two different scorers
| | | Percent adjacent agreement on first score
| | Score point distribution | Frequency distribution of scores assigned
| | Validity report | Average exact agreement on validity papers
| | | Average exact or adjacent agreement on validity papers
| | Throughput report | Number of papers scored
| Individual reader-level reports for each item | Reader monitor report | Spearman correlation between first scores by two different scorers
| | | Percent adjacent agreement on first score
| | Score point distribution report | Frequency distribution of scores assigned
| | Item status report | Status of each response read by reader (e.g., requires second read)
| | Response read by reader report | Identifies all responses scored by an individual reader (This report is useful if any responses need rescoring due to potential reader drift.)
| | Readbehind log | Summarizes outcomes of backreads
valid and custom messages that describe the validation rule that was validated.

Outgoing data are governed by similar data layouts, and automated processes validate each data file against syntactic rules negotiated with USOE. This highly automated process enables us to deliver accurate nightly files to our clients who want them.

Tests administered online are monitored in real time by our QM system. Each completed test runs through QM as soon as the test is complete, where a variety of quality checks is conducted. For example, QM checks attemptedness calculations, raw scores, and other indicators. It periodically aggregates these data to check for items with extreme \( p \)-values and identify any items for which the most common response is not also the correct response.

**Help Desk Operation**

AIR’s customer service representatives must complete a rigorous three-part training program:

1. Training on the functionality of each component of the online testing systems as well as the paper testing program including the policies and procedures set forth in user guides and manuals

2. Training based on past logs of customer queries taken from each stage of an online administration: preparation, startup, and ongoing administration

3. Training on how to use the detailed Visio scripts that enable the customer service agent to ask a series of questions to callers to quickly diagnose the system and point of issue with which the caller is experiencing difficulty

In addition to providing rigorously trained customer service agents, AIR’s help desk is supervised and monitored by a full-time supervisor responsible for monitoring response accuracy.

The supervisor regularly audits calls to the help desk to ensure that help desk agents are providing accurate information. When a question of possible inaccuracy arises, the help desk supervisor obtains a copy of the phone recording and can confirm whether or not the information provided was correct.

Daily and weekly statistical reports enable us to monitor the level of service provided by each tier of our help desk, and project management monitors these reports.

**Training Content and Delivery**

Training content goes through a structured series of reviews, including reviews by AIR specialists, our program management team, and USOE experts. The content is then produced through a well-controlled process that parallels the process we use to develop paper-test forms. This process is described in Section I.C.2.b.

The training itself is continuously reviewed both formally (through participant surveys) and informally through debriefings with USOE and AIR staff. We review what worked, what did not work, and opportunities for improvement. We plan to update training materials annually, as necessary.

**Software Testing, Including Systems Integration Testing and Functional and Nonfunctional Testing**

Quality is built into every phase of design and development. AIR’s process includes peer review of critical documents, code review, and simulations of critical pieces of the system (for example, computer-adaptive algorithm). In this section, we address testing and test plans.
Testing takes place at four levels:

1. **Unit testing**, where a developer verifies that all small parts of the system function properly. The developer compiles and runs the unit (for example, a function or procedure) while visually walking through the code. Both basic and alternative paths and branches of the code are tested.

2. **Integration and system testing**, which verifies whether all systems that interconnect with each other work reliably and completely and work in an environment that mimics the production sites.

3. **Load testing**, which ensures that, with large volumes of data passing through, the system works properly within the required amount of time for user feedback. This testing also determines which modules execute most often or use the most computer time.

4. **User acceptance testing**, which ensures that the system meets its written specifications and requirements and requires sign-off by client representatives. This testing usually occurs after the systems have successfully passed all of AIR’s internal tests (including, at minimum, unit testing, integration testing, security testing, and load testing).

The test plan identifies what is to be tested in all four levels described above, as well as the test strategies that will be used to test each part. The plan aims to measure three objectives: (1) data integrity, which ensures that data (including database validation and data conversion) are consistent and correct; (2) interface reliability, which ensures that the graphical user interface meets its written specifications and is an accessible system (including browser/operating system compatibility testing and usability testing); and (3) operational reliability, which refers to ensuring that the system, including its security and performance aspects, is deployable.

AIR considers the test plan as both a product and a tool. The former ensures that structure, format, and detail of the plan are effective for the testing effort as well as the needs of USOE, while the latter helps AIR manage the testing project and find bugs in the software. The test plan is created by a core team of system representatives; namely the technical lead for the system (see Section IV.D), the software project manager, the test lead, and a subject matter expert (for example, a senior psychometrician or test development specialist). Collaboration results in robust, effective test strategies and efficient testing efforts. The test plan authors determine the following:

- **Risk points**—What features and functions should be prioritized based on importance and likelihood or impact of failure?
- **Test strategies**—What strategies or levels should be performed? Test strategies aim to measure data integrity, interface reliability, and operational reliability. They include testing the following: algorithms, stored procedures/API calls, configuration, database, data conversion, data output, browser/operating system compatibility, graphical user interface, fault injection, security, and performance testing.
- **Test cycles**—How many cycles should be performed? Which tests should be repeated to cover several combinations of user scenarios?

Because the test plan is considered to be a product, it likewise goes through reviews and sign-offs by peers, the technical lead, and USOE (if preferred). Once signed off on, the test plan is implemented in four ways:

1. The test lead or junior developer builds custom drivers that call the stored procedures or APIs directly and generate data for verification. These drivers ensure that all branches and paths of the code are tested,
including those that do not go through the user interface.

2. Automated test scripts built by the test lead are executable procedures that mimic a user’s set of actions and generate a transaction log for verification.

3. Manual testing, in which a tester goes through the test cases by hand and visually verifies if he or she observes the expected result.

4. Full regression testing is performed when any enhancement or modification is added to the system.

**Components of the Test Plan**

To do an exhaustive job of testing, the test plan measures how thoroughly the code is exercised and manages the testing effort and implementation. Below, we briefly address each of the points identified in the RFP, which make up a subset of the information to be found in our test plans.

**User Acceptance Test Plan**

This plan obtains information by client representatives that the overall system or a specific modification meets mutually agreed-on requirements. Components of this test plan include test environment, test period, test cases, user accounts to use, and a sign-off sheet.

**Developing Test Scripts**

To simulate the real-world environment more efficiently, test scripts and custom drivers are built to exercise volume and stress on the system. For automated scripts, usual tools include Quick Test Pro (a record/play tool used for functionality testing), Visual Team Studio (used for load testing), and Watin (used for custom drivers).

**Testing all Software Components**

The unit test plans ensure that each individual component is tested, and is verified to be “fit for use” before integration with other components. Integration testing, on the other hand, ensures that all components work together.

**Parallel Testing**

Development and enhancements are a constant process, and they do not affect the production environment until the following occur: deployments are scheduled and approved, changes are tested and accepted, PCB and client approval are received, and the system is deployed. In the meantime, multiple environments exist that allow development, testing activities, and production activities to occur at the same time but in different locations. At minimum, there are the development site, two test sites (one to hold exactly the same version as the one in production for maintenance purposes and the other to hold the version with enhancements), the load test site, and the production site.

In particular, the load test site must be done on the same equipment as production for the load test to be valid; thus, AIR deploys separate instances of the system on the production servers during maintenance to perform load tests. This approach of co-locating the load test site in the production environment allows AIR to verify, with certainty, that the system properly works under large volume, peak bursts of activity, and peak memory and space capacity. The production environment has server, network, and device configurations and interrelationships between them. This cannot be mimicked in any other test site without incurring large development costs and efforts by developers and system administrators in making sure the sites are in sync.
Security Testing

In the test plan, the section on security testing verifies whether the system (with its enhancements) protects data and maintains data and system integrity as intended; namely, disclosure of information, data received or transferred were not altered in any way, source of data is known, and the specific user is allowed to perform a service.

End-User Activity Testing

The test plan on end-user activity testing ensures that a system is user friendly for the targeted audience. Aside from client representatives involved in implementing user acceptance testing, AIR can work closely with USOE to identify representatives from different user populations including test administrators, district test coordinators, teachers, and students. Test cases that allow these users to walk through the various system interfaces will be provided to them.

Hardware and Network Capacity Testing

This testing ensures that the system can handle heavy loads of data and system processing. AIR builds custom drivers that simulate actual demands on the system. We establish the target (for example, 30,000 concurrent students with an average of 1 second server response time), monitor the latencies and use of bandwidth capacity, and ensure that the system will meet service-level commitments.

Integration Testing

We conduct integration testing with a mix of customized drivers, automated test scripts, and manual testing. Integration testing verifies that all systems work together properly. Integration testing works hand in hand with load testing, which ensures that the systems continue to function properly when simultaneously running on many threads.

Plan by Which Testing Will Be Repeated for Updates to the System

A system update (as opposed to a system change as described below) may not necessarily involve software. Instead, it may involve an update on the hardware, supported browsers, or operating systems, including a new version of Firefox or Internet Explorer, and a security patch on the testing server. Depending on the nature of the update, test strategies and risk points are determined. For example, a new version of Firefox involves GUI testing, functionality testing, and item content review. On the other hand, an update on a security patch may require a load test if AIR’s system administrators believe there is a risk that performance may be affected.

Plan for Full Regression Testing for Changes to the System

A change to the system, as described here, refers to enhancements or modifications on the software itself. Full regression testing is performed when any enhancement or modification is added to the system; this ensures that the change did not break or disturb any other component of the system. It tests the overall integrity of the system, not just the success of the actual change. AIR uses a combination of automated test scripts, custom drivers, load tests, and manual tests, thus allowing complete test coverage to be done efficiently.

QA Reviews and Timely Problem Resolution

Quality assurance is integrated into all our processes and AIR is committed to continuous improvement in all our processes. If quality
issues arise, we will address them immediately in management meetings with USOE and resolve them as quickly as feasible. We will work with USOE to identify valuable process improvements each year at the annual planning meetings.
IV.D. Project Management

Management Overview

Complex, large-scale assessment programs demand organized collaborative project management. AIR’s management structure has supported years of successful implementation of peer-reviewed, multiple-opportunity online assessments in multiple states. More recently, we have helped and rapidly and successfully switch from a paper-based system to similar multiple-opportunity adaptive online systems. In 2011, AIR migrated ‘s state assessment system from the previous online testing service provider.

Our success in these endeavors testifies to the effectiveness of our management processes:

1. Close management in every area of the new online program, using a set of structured management tools and documents designed to facilitate solid planning, tracking, and managing change

2. Foresight in assessing, managing, and reducing the risks inherent in such a complex project

3. Clear and repeated communications, training, and customer support for educators and administrators to reassure and educate stakeholders about the wholesale change involved in administering an online program

Building on a strong foundation of established methods for program management, our team easily adapts to customize services specific to each client. We have successfully applied these methods on state programs and will adapt to meet USOE’s requirements.

In this section we describe our management approach and unique management tools for planning, monitoring, and managing changes to the USOE assessment program and our quality assurance protocols. We describe how we will use our management experience and tools to successfully complete the work under this contract while collaborating with USOE.

In working with USOE to manage the day-to-day operations of the program, AIR will adhere to the principles set forth by the National Council on Measurement in Education in the Code of Professional Responsibilities in Educational Measurement and will work with USOE to adopt the applicable principles set forth in Operational Best Practices for Statewide Large-Scale Assessment Programs (CCSSO & ATP, 2010).

AIR understands that adhering to standard processes and documentation is critical to the program’s successful operation.

Managing this complex project requires a project management team that is experienced at focusing people on the important issues and helping them reach consensus. AIR understands that while the timelines on most assessment projects are challenging, consensus from all the major stakeholders must be obtained in order for the project to be successful.

Management of the CAT System and Its Components

In this section, we introduce our key staff and describe the organization of AIR’s Assessment Program and our proposed project team for USOE.

Our project team will be led by , a senior project director perfectly suited to manage the USOE Test Delivery System effort. Ms. has 10 years of management experience in educational assessment, including managing multimillion dollar accounts. Most recently, she managed Pacific Metric’s online formative assessment program, demonstrating the ability to coordinate seamlessly with tech-
the project, with ultimate responsibility resting with the corporate director of the relevant functional area.

Dr. sets the vision and standards for the program; recruits, develops, and motivates staff; and makes sure 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.

brings over 30 years of assessment experience to his role as AIR Assessment’s 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 are expected to master their capabilities through an apprenticeship system predicated on the gradual increase of individual responsibility. Group training sessions and collaborative interaction that encourage and promote 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
- Online Testing Systems
- Online Reporting

**Air's Assessment Program and Project Structure**

The organizational structure of each of AIR’s projects follows the functional organization of AIR’s Assessment Program. We summarize the organizational structure and the depth of capability of our staff below. AIR’s corporate directors work directly on our state projects more than 90% of the time. AIR’s Assessment Program 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 technology team members, lead the User Support team, and work with users at all levels from state users to teachers, and conduct training workshops. She has also managed reporting lines and handscoreing activities for multiple state and district assessment projects.

Our staff also includes the core software team at AIR that has been responsible for the first and only NCLB-approved, full-year, multiple-opportunity online adaptive assessments, now operational in four states.

The project team will work under the direction of...
Specialist
Project Management

Director, Online Testing:

Project Lead:

The Project Management team is responsible for orchestrating these resources to deliver and report on multiple assessments. (Online) directs our Project Management team.

Ms. has more than 15 years of experience in online systems, including managing ‘s online testing project and document- ing and transferring best practices in online, adaptive, full-year, multiple-administration testing from to new state clients. She received her J.D. from the University of Iowa.

AIR is proposing to serve as the senior project director. She will lead all internal management meetings and all communications with USOE management staff. Her team will maintain schedules, produce minutes and other documentation, and implement the project management and quality assurance procedures to ensure fidelity to USOE goals and policies.

, M.S., is joining AIR as a senior project director with 10 years of management experience in educational assessment, including managing multimillion dollar accounts. She most recently managed Pacific Metric's online formative assessment program, demonstrating the ability to coordinate seamlessly with technology team members; lead the User Support team, working with users at all levels from state users to teachers; and conduct training workshops. She has also managed reporting lines and handscoring activities for multiple state and district assessment projects. In addition to being a highly organized manager with budget and client services skills, Ms. is involved in the substantive aspects of assessment projects. She has presented frequently. Especially relevant to Utah’s project was a session she led at CCSSO’s 2012 Large Scale Assessment Conference on different models for, and issues associated with, state support for classroom assessment. That session focused on models to support online assessments tied to state standards and curriculum frameworks without unduly intruding on the local dynamics of classroom instruction and learning. Ms., a certified Project Management Professional (PMP), earned her M.S. (with a minor in Statistics) from Virginia Tech and her B.S. from the University of Vermont.

Ms. will be assisted in the management of this project by . Ms. has six years of experience on AIR’s Psychometrics and Statistics team and has spent four of those years as a statistical support associate. Her responsibilities include providing research and logistical support to senior staff in the Assessment Program and composing and editing technical reports, analysis specifications, and reporting documents related to the statistical integrity of state assessments. Ms. has a B.A. in Psychology from the University of Maryland.
Computer and Statistical Sciences Center (CSSC)

CSSC is responsible for both our software infrastructure and our software products. CSSC takes as its mission the development of innovative products and systems that will keep AIR’s Assessment Program at the forefront of our field. CSSC combines a structured software development unit (developers and project managers) with a team of statisticians, psychometricians, and mathematicians. It is organized into groups supporting families of software systems:

- Preproduction 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 80 staff members, six of whom have Ph.D.s ranging from computer science to statistics and engineering. Thirty-eight of our software project managers and other engineers have M.B.A.s 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.

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 for operational tests (, and ). Prior to joining AIR, Ms. 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. 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. received her M.B.A. from the Wharton School of Business, University of .

Dr., Ph.D., will provide technical leadership for this project. A senior information technology professional, he has had extensive experience in managing diverse software and hardware projects at AIR, both as a technical lead and as a project manager. As principal software engineer, he serves as technical lead for the online Test Delivery System (TDS) in support of AIR’s Assessment Program. In prior positions, as software technical lead, he successfully designed, implemented, and managed numerous large enterprise systems, utilizing standard methodologies in every phase of the software development life cycle. As hardware/software project lead, he successfully led the hardware design and
implementation, including the transfer into manufacturing and development of software for complex optical networking systems. Dr. holds a Ph.D. in Computer Science from the University of Maryland.

**Psychometrics and Statistics**

AIR’s Assessment Program 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 examinees. 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. leads the Psychometrics and Statistics team. Dr. 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 include overseeing the NAEP, the National Adult Literacy Study (NALS), and the Third International Mathematics and Science Study (TIMSS). Dr. was the architect and served 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 models, and Bayesian inference. He is nationally and internationally known for his expertise in large-scale assessments and complex surveys.

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 item response theory (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

Dr. , Ed.D., will lead psychometrics for the Utah project. Dr. brings more than 25 years of experience as a psychometrician with experience in computer-adaptive, computer-mastery, linear-on-the-fly, computer linear, and paper-and-pencil testing. He has worked in licensure and certification testing,
K–12 assessments, on the Graduate Record Examinations (GRE), and on the Test of English as a Foreign Language (TOEFL). He has experience with innovative item types and automated scoring models applied to essays, short text-based items, and mathematics items. He currently serves as the senior psychometrician on the Smarter Balanced Assessment Consortium pilot studies. Dr. received his Ed.D. in Educational Measurement and Statistics from Rutgers University.

**Test Development**

Developing nearly 10,000 items annually, our test development staff has been developing and aligning existing items to Common Core State Standards since their inception. This team is led by . Ms. has been the director of test development for the Assessment Program at AIR for the past 12 years.

Ms. 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 the development of the assessments. Before joining AIR, Ms. worked at ETS for 10 years, managing test development for numerous projects, including the GRE, GMAT, LSAT, SAT, NAEP, and Praxis assessments. Ms. was also involved in project management for the reading, writing, and mathematics development for the ISEE and the CTP IV. Ms. 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, AIR Test Development staff are responsible for the following tasks:

- Framework development
- Test blueprint and item specification development
- Item development and 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

will lead item development for this project. He brings over 11 years of experience in mathematics assessment. Mr. currently serves as a mathematics test development specialist at AIR where he is responsible for the development, review, and management of multiple-choice and constructed-response items on numerous projects. He also has been named the mathematics lead for grades 6–8 of the Achievement Assessments. Before joining AIR, he worked as the mathematics lead for Measured Progress for projects including high school end-of-course tests and alternative assessments. Mr. also has two years of experience as a high school math teacher. He has a B.A. in Mathematics from Augustana College and a Teaching Certificate in Secondary Mathematics Education from St. Olaf College.

Reporting to Mr., our subject area development will be led by

- Dr., Ph.D. (mathematics),
Ms. (English language arts), and
Dr. , Ph.D. (science).

Dr. , Ph.D., has experience teaching elementary math methods and instructional technology courses for both pre-service and in-service teachers. Further, Dr. has facilitated content and data reviews and edited items and test forms through all steps of the publishing process. Currently he is the math content lead for one of AIR’s statewide assessment project for all grade levels. He holds an M.A. in Computer Science and Mathematics Education from Concordia University and a Ph.D. in Mathematics Education Leadership from George Mason University.

After six years as an eighth-grade English teacher, , M.Ed. has spent the past 10 years working in the field of test development. She led AIR’s ELA team before her recent assignment as the senior program manager for ’s K–8 assessments, where she has been responsible for managing the phases of development, administration, scoring, scaling and equating, and reporting test results. Mrs. also has experience developing certification and licensure examinations for the CFP Board, the Foundation for Rehabilitation Education and Research (FRER), Convention Industry Council, English Language School (ELS), and Homeland Security First Responders for the State of Florida. She holds an M.A. in English Literature and Creative Writing from West Chester University and a M.Ed. in Secondary Education in English.

Our proposed science lead, Dr. , Ph.D., brings five years of high school physics and mathematics education experience, including five years working in educational assessment for AIR. For four of those years, he has served as the science content development lead for a variety of state assessments such as the Comprehensive Assessment, the State Assessments, and the High School Transformation Initiative. Dr. holds an M.A. in Education and a Ph.D. in Applied Physics, both from Stanford University.

**Operations and Scoring**

The AIR Operations and Scoring team includes seven full-time and more than 200 part-time professionals and is responsible for warehousing, distributing, collecting, security processing, scanning, editing, performance scoring, and preparing data files. The supporting software systems interface directly with our pre-ID and ordering systems, student tracking, print tracking, and other AIR systems.

is the senior manager of processing operations and leads AIR’s Operations and Scoring team. Mr. is responsible for all hardware, software, staffing, and implementation of process and procedures. He combines his 15+ years of experience in various operations and processing management roles with AIR’s state-of-the-art technology to deliver the most accurate performance in the industry.

AIR employs patent-pending technology and a series of progressive computer and human quality assurance processes to ensure virtually flawless processing of answer documents and complete security processing of all test materials.

Mr. has taken a lead role in process improvements and in AIR’s technology for scanning and scoring. Mr. reports to , vice president and executive director of Assessment. Mr. has 20+ years of industry experience, including six years as a senior executive in charge of the largest testing organization in the world. He has taken a lead role in project management and in AIR’s technology for scanning and scoring.
Online Reporting

AIR goes beyond simply reporting scores and other numbers to provide deep analyses of the data, reported in a way that is clear, appealing, and actionable. Although not required as part of this RFP, as an option, AIR offers printed family reports. 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.

is the director of our Online Reporting area, bringing 14 years of experience to this role. Ms. has led the development and on-time delivery of more than 30 million pages of score reports for AIR clients in , and various city assessment systems. Prior to joining AIR, Ms. designed and developed the messaging strategies for several national and state education groups, including the National Education Association, 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. worked for five years as managing director for a Vancouver-based publishing group.

Our proposed reporting lead, , M.A., has successfully led statewide reporting for AIR projects for , and . He previously managed test development and psychometrics for the Green Building Certification Institute’s credentialing program, where he oversaw the development and maintenance of several credentialing assessments for the Leadership in Energy and Environmental Design (LEED) Professional Credentialing Program. Before working in assessment, he worked in program evaluation at the World Bank’s Independent Evaluation Group. Mr. has an M.A. in Measurement, Statistics, and Evaluation from the University of Maryland and a B.A. in Mathematics and Sociology from Boston University.

Operations Scoring

For this project, AIR has teamed with Data Recognition Corporation to provide handsoring services. While Ms. will directly manage this subcontract, will provide organizational oversight. Vice president and executive director of AIR’s Assessment Program, Mr. combines his 20+ years of experience in senior management positions at Pearson with AIR’s state-of-the-art technology to deliver the most accurate performance in the industry. AIR employs patent-pending technology and a series of progressive
computer and human quality assurance processes to ensure virtually flawless processing of answer documents and complete security processing of all test materials. Mr. joined AIR as executive director of operations in 2007 after serving for six years as a senior executive in charge of the largest testing organization in the world. He has taken a lead role in project management and in AIR’s technology for scanning and scoring.

Our partner DRC’s team will be led by Ms. , M.S., vice president of performance assessment services at DRC. Ms. brings more than 19 years of assessment experience to her position at DRC. She has managed performance assessment activities for numerous state clients, including . In addition, Ms. leads DRC’s product development team for strategic planning and execution in innovative items, automated scoring, and artificial intelligence.

Ms. has been instrumental in assisting several of DRC’s state clients with transitioning to online assessments and is keenly interested in the pedagogical implications of online learning and assessment environments. She recently gave a presentation titled “Constructed-Response Scoring and Associated Technologies” to the Technology in Large-Scale Assessment (TILSA) committee of the Council of Chief State School Officers (CCSSO).

Ms. brings a blend of experience and varied perspectives to her role. She began her career as an instructional tutor for students in a K–8 school and also worked as an ELA middle school teacher. Her classroom experience helped her discover the power of formative assessment as a way to uncover gaps in understanding and increase student achievement, and it led her to participate in scoring assessments for the National Assessment of Educational Progress (NAEP). She also worked as a scoring specialist at National Computer Systems.

Ms. holds an M.S. in Human Resource Development from Xavier University in Cincinnati, , and a B.S in Elementary Education from Miami University in Oxford. She is a member of Pi Lambda Theta, the International Honor Society and Professional Association in Education.

**Production**

Manager:

AIR will produce item graphics, paper test materials, user guides, administration manuals, and communication brochures associated with implementation of the online testing system in accordance with the requirements outlined in the RFP. AIR is committed to delivering materials as USOE expects. We envision a fully collaborative process, and USOE will have review and signoff authority at each stage.

Our production processes include multiple editorial reviews, and each document must receive senior staff sign-off. Exhibit IV.D-2 provides an overview of our editorial review process. Each document’s content is fully drafted before the document enters the production system.

, senior production manager, leads AIR’s production group. Ms. has 13 years of production/production management experience, including 7 years as AIR’s production manager responsible for test administration/coordinator manuals, user guides, test booklets, and answer documents. The production group comprises 17 graphic
designers/desktop publishers and 3 production editors. Graphic designers are proficient in popular applications, such as Adobe Creative Suite (Illustrator, Photoshop, InDesign), QuarkXPress, and Adobe Acrobat Professional.

AIR will update the manuals and guides annually to reflect changes in the system.

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 USOE, during which we discuss issues of both design and content. We view the drafting process as iterative. Especially with new mate-

---

Exhibit IV.D-2: Overview of Our Editorial Review and Revision Process

- Draft Document
  - Senior Review
    - Editorial Review
      - Initial Layout Enter Revision Process
        - Blackline 1
          - Senior Review
  - Yes
  - No

- Client Review
  - Change
    - Blueline Proof
    - Initial Revision Process
      - Yes
      - No
      - Senior Review
        - Change
          - Blackline

- Revision Process Detail
  - Layout
    - Yes
    - No
  - Proofreading and Editorial
    - Yes
    - No

- Yes
- No
rial, we understand that it is not possible to make all final decisions prior to a first draft. 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 multiple drafts of each product prior to entering the production process.

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.

Communications Plan

AIR understands that in order for this project to be successful, we will need to work and 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 USOE and the USOE management partner 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 the project’s organizational chart that will include all key team members, their contact information for immediate access, and chain of command contact information. AIR will assist USOE in designing a comparable project-based organizational chart, if desired, so that contact information is at everyone’s fingertips.

AIR has developed a series of tools that enable our project directors and manager to analyze the project requirements. These tools form the basis for clear, open, and accessible communication about the project schedule, deliverables, and specific requirements. Our project directors function 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.

Management Documents

The suite includes four types of tools that project directors use to orchestrate the work of the project:

- Schedules,
- Planning Documents
- Tracking Documents
- Specification Documents

We produce schedules on three levels. At the project level, we produce the overall project calendar. The project calendar documents work-start and completion/delivery dates for key activities, milestones, and deliverables. We typically maintain the project calendar in Microsoft Project. A Gantt Chart representing the preliminary project calendar for this project appears in Appendix F.
At the next level, we have “team schedules.” For example, our Test Development team will work with key content personnel at USOE to develop detailed committee meeting schedules, batch delivery schedules that document the dates that batches of items will be delivered for review. The Psychometrics and Statistics team develops an analysis schedule, which identifies milestones in the data receipt, analysis, and delivery schedules. Some of these schedules are entirely internal, designed to enable us to meet the milestones on the project schedule. Other schedules include tasks that require interaction with USOE, and these schedules will be developed in collaboration with appropriate USOE staff.

Planning documents kick off complex activities. They outline the general approach to be taken to the activity, assign responsibilities, and establish milestones. Some of these milestones may be further specification documents, or elements needed to complete the plan. For example, the project director produces an administration summary for each administration, outlining the basics of the administration, including characteristics such as the 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. The Test Development team uses this document as a starting point for their item development plan, and the Psychometrics and Statistics team uses this document to begin their field-test plan and analysis schedule.

Our project management team relies on several tracking documents to monitor project status, coordinate work among internal teams, and communicate progress toward milestones to our clients. For example, to coordinate the multitude of deliverables exchanged between AIR and our clients, our operations lead maintains (and shares with our clients) a priority list of deliverables that are currently in play at the moment (e.g., blackline 1 of the test administration manual). 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. The document also assigns a priority level to help clients manage and prioritize work during busier times in the schedule.

Our project management team maintains an issues log, which tracks cross-team issues, notes progress toward their resolution, identifies risks, and documents changes to scope, if necessary. We also recognize that plans and objectives change. Therefore, we have a structured process for gathering, tracking, and disseminating changes to management documents. These include regular planning meetings and project updates, issues logs, and other tools.

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 USOE staff to develop a detailed reporting specifications document including lists of reports to be produced, descriptions of report recipients and packaging specifications for paper reports, calculation rules for each data element, and other fine details of the reporting system. This document is usually created in collaboration with the lead psychometrician and technical, content, and leadership staff at USOE.

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
## Exhibit IV.D-3: Schedule Documents

<table>
<thead>
<tr>
<th>Team</th>
<th>Document</th>
<th>Description</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management</td>
<td>Project Calendar</td>
<td>Project milestones, start dates for activities</td>
<td>project director</td>
</tr>
<tr>
<td>Management</td>
<td>Production schedule</td>
<td>Timelines for review and approval of items, paper/pencil forms, user guides, and ancillary materials</td>
<td>project director</td>
</tr>
<tr>
<td>Psychometrics and Statistics</td>
<td>Analysis schedule</td>
<td>Schedule of data flows and analysis activities</td>
<td>lead psychometrician</td>
</tr>
<tr>
<td>Operations</td>
<td>Delivery schedule</td>
<td>Milestones for packaging and shipping materials</td>
<td>operations lead</td>
</tr>
<tr>
<td>Operations</td>
<td>Receipt and scoring schedule</td>
<td>Milestones for document receipt, scanning, and scoring</td>
<td>operations lead</td>
</tr>
<tr>
<td>Test Development</td>
<td>Batch delivery schedule</td>
<td>Dates for delivery of item batches and item reviews</td>
<td>item development manager</td>
</tr>
<tr>
<td>Test Development</td>
<td>Delivery schedule</td>
<td>Dates and milestones for the preparation and convening of annual committee meetings</td>
<td>item development manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Receipt and scoring schedule</td>
<td>Schedules for development and deployment</td>
<td>software project manager</td>
</tr>
</tbody>
</table>

## Exhibit IV.D-4: Specifications Documents

<table>
<thead>
<tr>
<th>Team</th>
<th>Document</th>
<th>Description</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychometrics and Statistics</td>
<td>Analysis specifications</td>
<td>Describes the details of analyses, including software settings, locations of data files, and many others</td>
<td>lead psychometrician</td>
</tr>
<tr>
<td>Psychometrics and Statistics</td>
<td>QC specifications</td>
<td>Identifies quality control-specific analyses and comparisons to be undertaken</td>
<td>lead psychometrician</td>
</tr>
<tr>
<td>Psychometrics and Statistics</td>
<td>Scoring engine specifications</td>
<td>Summarizes scores to be reported at various levels, attemptedness rules, and scoring computations that are performed by the test scoring engine for all tests</td>
<td>lead psychometrician</td>
</tr>
<tr>
<td>Psychometrics and Statistics</td>
<td>Test specifications</td>
<td>Includes test blueprints, item specifications, and style guides</td>
<td>lead psychometrician</td>
</tr>
<tr>
<td>Test Development</td>
<td>Field-test summary</td>
<td>Shows numbers and types of items to be field-tested each year and form construction details</td>
<td>item development manager</td>
</tr>
</tbody>
</table>
## Exhibit IV.D-4: Specifications Documents (continued)

<table>
<thead>
<tr>
<th>Team</th>
<th>Document</th>
<th>Description</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Development</td>
<td>Item Tracking System specifications</td>
<td>Defines the attributes to be collected for each item, the content review sequences, and the item layout templates</td>
<td>item development manager</td>
</tr>
<tr>
<td>Operations</td>
<td>Packaging specifications</td>
<td>Describes materials to be shipped, shipping carrier, expected packing list format, materials to be packed, and details of packaging</td>
<td>packaging and shipping manager</td>
</tr>
<tr>
<td>Operations</td>
<td>Receipt specifications</td>
<td>Details secure and nonsecure documents to be received, expected quantity, receipt start and end dates, and special instructions</td>
<td>packaging and shipping manager</td>
</tr>
<tr>
<td>Operations</td>
<td>Scoring specifications</td>
<td>Identifies human-scored constructed-response items, the dimensions and points on which items are scored, condition codes for unscored responses, backread and second-read rates, frequency of validity papers, and target reliability and validity statistics</td>
<td>scoring director</td>
</tr>
<tr>
<td>CSSC</td>
<td>Requirements documents</td>
<td>Defines the detailed requirements for the software or software modification to be implemented</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Test Information Distribution Engine (TIDE) specifications</td>
<td>Defines detailed requirements for materials ordering, student management, and user management, including user roles within the suite of online testing applications</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Data file generator specifications</td>
<td>Defines detailed requirements for student and item score files delivered after testing</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Learning Point Navigator specifications</td>
<td>Defines detailed requirements for AIR's Learning Point Navigator, which provides instructional resources targeted to identified areas of strengths and weaknesses</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Online Project Portal specifications</td>
<td>Defines the layout, skin/theme, user cards, and content sections</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Test Delivery System Specifications</td>
<td>Defines the test settings for the Proctor and the Student Interfaces and the business rules that govern test opportunities</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Online Reporting System specifications</td>
<td>Lists the reporting settings for student and aggregate reports</td>
<td>software project manager</td>
</tr>
</tbody>
</table>
### Exhibit IV.D-4: Specifications Documents (continued)

<table>
<thead>
<tr>
<th>Team</th>
<th>Document</th>
<th>Description</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting</td>
<td>Reporting specifications</td>
<td>Details the reports to be delivered; printing, packaging, and shipping specifications; data elements to be included; calculation rules for each data element; special rules for merging and cleaning data; and so forth</td>
<td>reporting lead</td>
</tr>
<tr>
<td>Reporting</td>
<td>Report mockups</td>
<td>Presents annotated graphic mockup of report, providing the look and feel and identifying the critical elements of the visual appearance of the reports</td>
<td>reporting lead</td>
</tr>
</tbody>
</table>

### Exhibit IV.D-5: Tracking Documents

<table>
<thead>
<tr>
<th>Team</th>
<th>Document</th>
<th>Description</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management</td>
<td>Issues log</td>
<td>Updated weekly; tracks issues and near-term deliverables; and notes progress, problems, risks, assignment, and agreed-upon changes to scope or progress</td>
<td>project director</td>
</tr>
<tr>
<td>Management</td>
<td>Priority list</td>
<td>Lists the testing and administration documents currently being exchanged between AIR and client; assigns a priority value to each document, identifies the document’s review level and due dates</td>
<td>project director</td>
</tr>
<tr>
<td>Management</td>
<td>Risk register</td>
<td>Identifies and classifies project risks, organized by function and area; summarizes proposed mitigation strategies</td>
<td>lead psychometrician</td>
</tr>
<tr>
<td>Test Development</td>
<td>Item release list</td>
<td>Lists specific items to be released each year for each subject</td>
<td>operations lead</td>
</tr>
<tr>
<td>Test Development</td>
<td>Batch delivery tracking sheet</td>
<td>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</td>
<td>operations lead</td>
</tr>
</tbody>
</table>
Exhibit IV.D-6: Planning Documents

<table>
<thead>
<tr>
<th>Team</th>
<th>Document</th>
<th>Description</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSSC</td>
<td>Project charter</td>
<td>Description of the goals, approach, and nature of new systems or system modifications</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Testing design</td>
<td>Design document describing the testing strategies to be employed for a new system or modifications to an existing system; lays the groundwork for the test plan</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Test plans</td>
<td>Provides detailed plans for load testing, individual test cases, and regression tests to be implemented</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Software release plans</td>
<td>Lists step-by-step activities for Software Development team and Network Engineering team to deploy software systems for scheduled releases</td>
<td>software project manager</td>
</tr>
<tr>
<td>CSSC</td>
<td>Production control forms</td>
<td>Details the steps required by AIR’s Production Control Board for evaluating and approving the test plans for a configuration change in the middle of the year in production systems</td>
<td>software project manager</td>
</tr>
<tr>
<td>Reporting</td>
<td>Reporting goals</td>
<td>Describes the objectives of the elements to be included in the report; identifies the goals and the types of actions that the report might motivate, and suggests data elements that might support those actions</td>
<td>Reporting Lead</td>
</tr>
</tbody>
</table>

In this case, client staff members are stakeholders in these documents as well.

Exhibits IV.D-3 – IV.D-6 summarize our standard schedule documents, planning documents, tracking documents, and specification documents.

**Ongoing Communication and Changes to Management Documents**

We recognize that plans and objectives can change frequently. Therefore, we have a structured process for gathering, tracking, and disseminating changes to these documents.

- Weekly project updates—These weekly meetings cover near-term deliverables and address issues as they arrive.
- Special-purpose meetings—If issues arise that require special attention, we will schedule these meetings with USOE as needed and in person when necessary.

AIR will deliver the weekly project update via email. Decisions are documented in the meeting minutes and on issue logs. If any decisions affect any of the communication documents, a section in the meeting minutes reflects this. The project director is responsible for communicating the changes to the document owner.
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.

All major project documents are stored on KnowledgeTree.

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.

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 kickoff will increase productivity, increase efficiency, reduce error and miscommunication, provide a single reference point for all documents that house key project decisions, and lead to a higher degree of collaboration across internal teams and across our organizations.

**Weekly Meetings**

Led by the project director, AIR will host the weekly USOE–AIR program update conference calls to discuss the period’s progress and to identify upcoming deadlines and possible challenges. 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, USOE will receive general progress updates from each team leader. USOE will have the opportunity to speak directly with AIR’s Technical, Systems Engineering, Item Development, Operations, and Project Management team leaders.

Decisions taken 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 USOE stakeholders before the decision becomes final. In this way we ensure that (a) the entire team is aware of the change and (b) each functional group has an opportunity to identify risks or issues that may arise as a result of a change in another functional area.

AIR will maintain an *issues log* to document issues across the weekly meetings. The issues log will be updated continuously and will document the resolution of each issue.

**Communication Plan: Access to and Contact Information for the Project Management Team and Chain of Command Contact Information**

At the kickoff meeting, AIR will provide contact information for the project team as well as an escalation plan including contact information for AIR’s senior management.

**Tracks Work and Provides Progress Reports Against Key Deliverable Dates**

AIR will provide weekly status reports tracking progress against key deliverables. The format
of the status reports will be negotiated with USOE at the kickoff meeting.

**Communication Media**

AIR maintains a suite of communication tools and services to support project communication. AIR will provide a toll-free conference line for all USOE–AIR project-related conference calls. AIR offers web conferencing solutions that enhance a virtual meeting.

To conduct virtual meetings, appropriate technology needs to be in place. 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.

AIR offers all of the services listed below. We have a nimble conference services team and will adapt to existing virtual conferencing systems as needed. Because AIR has offices around the world and already works with stakeholders in multiple locations, staff at AIR are very familiar with these solutions. The offerings below require minimal support. During the course of the project and the different types of meetings, AIR will help select the right tool for the right event.

**CISCO’s WebEx** 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. Web-Ex permits us to reach thousands of attendees and connects us with real-time, online webinars and telephone conferencing. Whether we are meeting with USOE staff or assisting in training state personnel or even local staff, Web-Ex will assist us in connecting with stakeholders in a variety of media through interactive events delivered right to their desks. Besides sharing presentation slides, participating in live video, and recording events, attendees can interact with each other, regardless of the size of the meeting, using a live Q&A.

In addition to WebEx, which offers video and audio conferencing, instant messaging, and desktop sharing, AIR offers the following virtual conference services solutions:

**Global Crossing Conference Phone Line:** This traditional audio conference line may be suited for smaller USOE–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:** For complete video conferencing between multiple locations, AIR offers Tandberg Video Conferencing 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.

**Microsoft Lync:** AIR staff across the globe can connect through instant messaging and video conferencing and can share documents using Lync. Lync is installed on every employee’s computer.

AIR has discussed above our plans for providing information to key stakeholders using management documents, webinar presentations, and virtual meetings. AIR will use the secure web-based systems we have used in previous projects, including a secure FTP site where documents can be posted and then downloaded by multiple users. The method(s) chosen for communicating information regarding each deliverable will be documented in the Communications Plan.
Kickoff Meeting and Annual Planning Meetings

Although the RFP refers to a single kickoff meeting, AIR anticipates a few initial meetings in the early stages of the project to successfully transition to and launch the new assessment system. Exhibit IV.D-7 presents a sample agenda from an initial planning meeting.

AIR will work with USOE and the previous 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 all USOE technical staff who have specific, relevant knowledge.

- **Vendor Meeting:** A follow-up meeting with USOE and the previous contractor. During this meeting, both project leadership and technical/operations staff from the previous contractor will work with AIR and USOE 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, USOE, and the previous contractor will monitor the transition timeline and delivery compliance as defined in the transition specification document. All three types of meetings work best when participants attend in person.
Exhibit IV.D-7: Sample agenda of an initial planning meeting

AMERICAN INSTITUTES FOR RESEARCH®
Utah Statewide Computer Adaptive Assessment System
Initial Planning Meeting
Sample Agenda
TBD, 2012

LOCATION: Utah State Office of Education (USOE)
ATTENDING: USOE and AIR Project Team Members
HANDOUTS: Project Schedule, Technical, Specifications Documents, Organizational Chart, Item Development Plan, Sample Field Test Plan, slides for each presentation

Day 1
Attendees: USOE & AIR

MEETING GOALS AND EXPECTATIONS: AIR will demo the Item Tracking System and the Test Information Distribution Engine (TIDE) and identify decisions required in the specifications documents so these systems can be configured according to USOE’s requirements. AIR will lead specifications gathering information for data exchange, the online assessment system and the online reporting system, presenting configurations and feature options. Through discussion and information gathering, AIR and USOE will seek to either complete the online testing specifications and item tracking specifications or take the necessary steps to ensure that work can begin on the development of these two systems.

Light Breakfast (8:00 – 8:15)
Introductions & Roles on the Project (8:15 – 8:30)

1. Item Tracking System (ITS) (8:30 – 9:15)
   a. ITS Demo and Overview
   b. Item Review Process
   c. Test Specifications
   d. Batch Review Process

2. Item Tracking System Specifications (9:15 – 10:45)
   a. Discuss and finalize ITS specifications

BREAK 10:45 –11:00

3. Test Information Distribution Engine (TIDE) (11:00– 12:30 pm)
   a. Demo of TIDE
   b. Proposed TIDE Functionality
   c. TIDE Specifications
      o User Roles

Working Lunch (12:30 – 1:00)

4. Data Exchange (12:30 – 1:30)
   a. Specifications
b. USOE-AIR file layout  
c. AIR-USOE file layout

5. Online Assessment System (1:30– 2:30)  
   a. Demo and Overview  
   c. Feature Selection  
   d. User Acceptance Testing  
   e. Accommodations overview

6. Online Testing Deployment Specifications (2:30 – 4:00)  
   a. Deployment Specifications and Business Rules

7. Score Reporting – Paper and Electronic (4:00 –5:00)  
   a. Deployment Specifications and Business Rules  
   b. Reporting Deliverables  
      i. Electronic Reports – Educator reports for districts, schools, and teachers  
      ii. Paper Reports – Family reports summarizing individual student performance over the year  
   c. Score Reporting Approach  
   d. Focus Groups Overview and Logistics  
   e. Schedule

Day 2 – Discussion on Online Systems/Finalize Specifications  
Attendees: USOE & AIR

MEETING EXPECTATIONS – During the morning, AIR will present Proposed Milestone Dates for the 2013 - 14 year, a discussion of the proposed field test design, and then a discussion of content-related issues. We will discuss paper and pencil tests: the forms review process, logistics around paper and pencil tests, ordering materials and paper processing. The last part of the day includes a discussion around communications, a demo of the online portals, and a demo of the online reporting system. The goal of Day 1 is to provide context and make decisions in each area so work can begin.

8. Proposed Milestone Dates for 2013-14 (8:45- 9:15)

   a. Reading, Math Field Test  
      i. Field test algorithm  
         1. Horizontal linking  
         2. Vertical linking  
      ii. Comparability

10. Item / Passage Display Templates (10:15 – 10:45)  
    a. Templates by subject  
    b. Templates by item type

BREAK: 10:45 – 11:00

11. Formative Assessments (11:00-11:30)

12. Review Processes (11:30 – 12:30)  
    a. Item Development and Review Process  
    b. Online Forms Review  
    c. Paper Forms Review  
    d. Review Periods
Exhibit IV.D-7: Sample agenda of an initial planning meeting (continued)

   e. Electronic Review and Approval

LUNCH (12:30 – 1:15)

13. Logistics for Paper Testing (1:15 – 2:30)
   a. Process for ordering paper materials
   b. Deciding to test on paper or online
   c. Pre-ID books
   d. Selecting sort option
   e. Submitting initial orders for test materials
   f. Submitting additional orders for test materials
   g. Adding students
   h. Modifying student information
   i. Shipping, Logistics

14. Paper Processing (2:30 – 3:30 pm)
   a. Confirm key delivery / pick up dates
   b. Hand scoring rules
   c. Processing and editing rules
   d. Appeals and verifications

BREAK 3:30 – 3:45

15. Communications (3:45 – 4:15)
   a. Portal
   b. Brochures
   c. Email blasts
   d. Newsletters

16. Score Reporting Demo – (4:15 – 4:45)

17. Wrap-up and Closing
Exhibit IV.D-7: Sample agenda of an initial planning meeting (continued)

**Annual Planning Meeting: USOE and AIR**

**Utah State Office of Education**

TBD, 2013

8:00 A.M. – 4:00 P.M.

**AGENDA**

**ATTENDING:** USOE and AIR Project Team Members

**HANDOUTS:** Project Schedule, Technical, Specifications Documents, Organizational Chart, Item Development Plan, Sample Field Test Plan, slides for each presentation

**Day 1**

**Attendees:** USOE & AIR

I. **Welcome & Introductions (8:00 – 8:15)**

II. **Review of 2013 – 14 School Year (8:15 – 10:30)**

A. Test Administration Summary

B. Lessons Learned

C. Process Improvements

III. **Project Management (10:30 – 12:00)**

A. Project Schedule

1. Milestone Dates

   a. Test Windows (Online/Paper)

   b. Proposed Rubric Validation Dates

   c. Proposed Data Review

   d. Batch Schedules

   e. Proposed Content Advisory Committee Dates

   f. Proposed Fairness/Sensitivity Committee Dates

B. Paper/Pencil Production Schedule Overview

C. Training Program for 2013 – 14

IV. **Technical (1:00 – 3:00)**

A. Review of Technical Report Schedule/Process

B. Field Test Plan for 2013 – 14

C. Planned Research Studies for 2013 – 14

V. **Help Desk (3:00 – 4:00)**
Exhibit IV.D-7: Sample agenda of an initial planning meeting (continued)

A. Review of most frequent help desk tickets
B. Identification of help desk trends and areas for improvements

Day 2
VI. Content (8:00 – 10:00)
   A. Item Development
      1. Review/Approval of test blueprints
      2. Number of items per grade and subject
      3. Review of existing batch review process
      4. Modifications to existing batch review process
      5. Batch schedule review
      6. Committee dates

VII. Score Reporting (10:30 – 11:30)
      A. New Features of Online Reporting System
      B. Discussion of future enhancements to Online Reporting System
      C. Review of content development schedule for paper score reports

VIII. Learning Point Navigator (11:30 – 12:00)
       A. New Features of Online Reporting System
       B. Discussion of future enhancements to Online Reporting System

IX. Break Out Sessions (1:00 – 4:00)
    A. Project Management
    B. Technical
    C. Content
    D. Score Reporting
    E. Data Management
**Invoicing**

AIR will provide quarterly invoices (or more frequently as directed) for services rendered coinciding with USOE’s fiscal year, which ends June 30, reflecting the budget presented in the proposal and finalized at contract signing. Invoices will itemize the work completed.

**Workshop Management**

AIR will be responsible for meeting space, materials, and associated cost requirements for workshops, conferences, and district meetings requested by USOE in the RFP. AIR staff will manage the committee meetings, training sessions, outreach, and project meetings over the length of the contract. Unless otherwise noted, AIR assumes costs for hotel, travel, and honoraria for qualified meeting participants as well as meeting materials, on-site copying, audiovisual needs, and meeting meals. In addition, AIR will assume the cost for substitute teachers, as required to facilitate training, at the current rates.

We will create a seamless logistics package for participants and facilitators, which will allow meeting attendees to focus on the goals of their committees, training, and workshops. Our meeting planner will be the main point of contact consistently throughout the logistics planning of each meeting.

The meeting planner will also arrange travel for attendees in accordance with state guidelines. During the meeting, AIR staff will manage on-site logistics for the training sessions. Following the workshop, our meeting planner will work with participants, vendors, and others to reach closure on post-conference service and budget issues.

AIR typically uses evaluation forms completed by workshop participants to evaluate the training. These evaluations will be discussed with USOE. The findings from the workshop evaluations will be used to inform the design of subsequent workshops and will be shared with future focus groups to aid in future workshop planning.
IV. E. Technical Manuals

In Section I.D, Psychometrics, we discussed in detail our plan for obtaining federal peer review. Technical manuals will be an important resource as USOE and AIR work together toward this goal.

We propose to produce a technical manual for the Utah statewide assessments that consists of five volumes, which are outlined below. AIR understands that USOE and its Technical Advisory Committee will review the annual technical manual prior to publication and that USOE will be allowed sufficient time to review the manual. AIR will provide USOE with electronic copies of the technical manual in both PDF and MS Word formats. Exhibit IV.E-I details some of the information to be included in the technical manual.

**Volume 1: Annual Technical Report**

The annual technical report will provide the basic information on the technical aspects of the Utah adaptive assessment program, including a description of the inner workings of the adaptive algorithm. Most of the sections are static, but a few of the sections present data on the tests administered for the year, and the statistics will vary with the composition of the item pool and the proficiency of the students tested. The variable sections will be updated each year. A preliminary report outline follows:

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Adaptive Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability of scale scores</td>
<td>Standard at each realized score</td>
</tr>
<tr>
<td>Test information</td>
<td>1. Information for complete item pool</td>
</tr>
<tr>
<td></td>
<td>2. Average information value for tests administered and yielding each score</td>
</tr>
<tr>
<td>Validity of scores</td>
<td>1. Estimated correlation between scale score and latent proficiency</td>
</tr>
<tr>
<td></td>
<td>2. Estimates of classification consistency</td>
</tr>
<tr>
<td>Content validity</td>
<td>1. Percentage of tests administered that match blueprint</td>
</tr>
<tr>
<td></td>
<td>2. Average range of content (number of benchmarks) seen by each student</td>
</tr>
<tr>
<td>Success of form equating</td>
<td>1. Estimates of average year-to-year equating error for individual test</td>
</tr>
<tr>
<td></td>
<td>2. Estimates of correlations among equating error across tests</td>
</tr>
<tr>
<td></td>
<td>3. Estimates of average equating error for aggregate statistics</td>
</tr>
</tbody>
</table>

Exhibit IV.E-1: Some Information to be Included in the Technical Manual
2. Recent and Forthcoming Changes to the Test
3. Summary of Operational Procedures
   3.1 Test Administration Procedure
      3.1.1 Testing Windows
      3.1.2 Test Administration
   3.2 Summary of Simulation Studies Performed Prior to the Operational Testing Window
      3.2.1 Testing Plan
      3.2.2 Statistical Summaries
      3.2.3 Summary of Statistical Analyses
   3.3 Summary of Adaptive Algorithm
   3.4 Accommodations
4. Maintenance of the Item Bank
   4.1 Overview of Field-Testing, Item Release, Reuse, and Retirement Policies
   4.2 Field-Testing
      4.2.1 Administration
      4.2.2 Sample Selection and Item Administration Selection Algorithm
5. Embedded Field-Test Item Analyses Overview
   5.1 Rubric Validation
   5.2 Field-Test Item Analyses
   5.3 Field-Test Item Data Review Committee Meetings and Results
6. Item Calibration and Scaling
   6.1 Methodology
   6.2 Item Calibration
      6.2.1 Item Fit Index
      6.2.2 Item Dependency
   6.3 Reporting Scale
      6.3.1 Overall Performance
      6.3.2 Reporting Category Performance
7. Summary of Form Development/Administration Algorithms
   7.1 Information Function From Adaptive Test
   7.2 Estimates of Classification Consistency
   7.3 Statistical Summary of the Content Distribution and Measurement Characteristics of the Tests Delivered
      7.3.1 Blueprint Match
      7.3.2 Item Exposure Rates
      7.3.3 Student Ability–Item Difficulty Distribution for the Operational Item Pool
8. Scoring
   8.1 Online Scoring
   8.2 Attemptedness Rule
9. Summary of Student Performance
10. Quality Control for Data, Analyses, Scoring, and Score Reports

Volume 2: Test Development

Much of the validity of a test rests on the procedural validity with which the test is constructed. This volume documents the procedures used to create the test and validate its alignment to the standards. A preliminary report outline follows:

1. Overview
2. Test Specifications
   2.1 Target Blueprints
      2.1.1 Score Reporting Categories
   2.2 Item Selection Algorithm
3. Item Development Procedures
   3.1 Summary of Item Sources
   3.2 Alignment of Online Item Banks to the CCSS and Utah Core Standards
   3.3 Development and Review Process for New Items
3.3.1 Development of New Items
3.3.2 Department Item Review and Approval
3.3.3 Committee Review of Item Pool
3.3.4 Rubric Validation

**Volume 3: Test Administration**

Standard setting takes place as an isolated event, rather than an annual process. Therefore, the standard-setting volume is created once and modified only when standards are revisited. A preliminary report outline follows:

1. Overview
2. Utah Statewide Assessments
   2.1 Testing Options
      2.1.1 Computer Adaptive Test Administration
   2.2 Testing Windows
3. Test Administration
   3.1 Administrative Roles
   3.2 Online Administration
   3.3 Braille Test Administration
   3.4 Allowable Resources for Online Testing
4. Training and Information for Test Coordinators and Administrators
   4.1 Online Training
      4.1.1 TA Certification Course
      4.1.2 Webinars
      4.1.3 Training Sites
   4.2 Manuals and User Guides
5. Test Security
   5.1 Student-Level Testing Confidentiality
   5.2 Test Security
      5.2.1 System Built-in Test Security
      5.2.2 Security of the Testing Environment
      5.2.3 Test Security Violations

5.3 Student Illness, Disruptiveness, and other Testing Incidents
   5.3.1 Reporting Testing Incidents
   5.3.2 Consequences of Testing Impropieties

6. Student Participation
   6.1 Home-Schooled Students
   6.2 Exempt Students
   6.3 Online Testing Features and Testing Accommodations
      6.3.1 Online Testing Features for ALL students
      6.3.2 Accommodations for Special Populations

**Volume 4: Evidence of Reliability and Validity**

Utah will have to produce evidence of the reliability and validity of the test. Below, we outline what the volume presenting this evidence might look like:

1. Overview
   1.1 Introduction to Technical Quality
2. Purpose of Utah State Assessments
3. Reliability
   3.1 Test Information Curves and Standard Error of Measurement
      3.1.1 Marginal Reliability
      3.1.2 Theoretically Minimum Standard Error
      3.1.3 Standard Error Curves
   3.2 Reliability of Achievement Classification
   3.3 Reporting Category Reliability and Precision at Cut Scores
4. Evidence of Content Validity
   4.1 Content Standards
   4.2 Test Specifications
Volume 5: Summary of Test Administration Procedures

Peer review is also concerned with administration procedures and the security of the test. We recommend compiling the relevant material in a technical report summarizing the administration procedures. A preliminary report outline follows:

1. Overview
2. Introduction
3. Score Reports
   3.1 Uses of Scores
   3.2 Home Page
   3.3 Subject Detail Page
   3.4 Content Strands Detail Page
   3.5 Roster Page
   3.6 Trend Page
   3.7 Student Detail Page
   3.8 Paper Family Score Reports
   3.9 Confidentiality of Student Data
4. Subgroup Reporting
5. Interpretation of Reported Scores
IV.F. Contract Finalization

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.

During the past seven years, AIR has successfully transitioned from other contractors 11 separate testing programs in , and . All of these projects demanded aggressive transition timelines, massive item and item bank development, new standards or assessment designs, extensive field-testing, and delivery of online testing systems. In accomplishing these significant goals, AIR learned that there are four core requirements for a successful transition:

- **Identification**—The first step is to develop a transition plan that clearly identifies all requisite transition information, as well as the data and format in which it 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 vendor must support the new vendor in replicating work and running analyses to ensure that data and trends are accurate.

Below, we discuss our experience and processes for ensuring smooth contract transitions.

Successful transition can be facilitated through an ongoing transfer of program documents, data, and content from the contractor to USOE. This ensures that USOE has access to, and control of, the necessary products for Utah’s system.

In addition, we recognize that integrating USOE data and content with the next vendor’s systems will require support from the outgoing vendor. We have been on both sides of assessment program transitions, and we understand the factors that can contribute to a successful transition. Below, we outline key transition meetings that help facilitate the process.

**Transition Specifications and Meetings**

AIR will work with USOE 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 all USOE technical staff who have specific relevant knowledge.

- **Vendor Meeting**—A follow-up meeting with USOE and the new contractor. During this meeting, both project leadership and technical/operations staff from the new contractor will work with AIR and USOE to define information format and delivery schedules and to finalize the transition specification document.

- **Transition Meetings**—In meetings specific to relevant functional areas, staff from AIR, USOE, and the new contractor will monitor transition timeline and delivery compli-
ance as defined in the transition specification document.

All three meeting types 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 these elements:

- A description of the item to be transferred (e.g., item content, item statistics, data from previous administrations, etc.)
- The format in which the item is available
- The reference source, which consists of the definitive source against which the transferred item could be verified
- The quality assurance plan, which describes how the item will be checked against the reference source
- The planned transition date
- Transition progress
- Additional information and contents

These specification are necessarily quite detailed. For example, these are some of the items that might appear under paper score reports:

- Scoring and reporting processing rules
- Statistical code
- Technical specification for statistical calculations
- Input data layouts
- Report static content
- Report variable content

- Training support materials and document
- Printing specifications
- Focus group/research plans and results

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 USOE on an ongoing basis and work with the incoming vendor to provide the data in a format that may be integrated with the vendor’s systems. If we are the selected vendor, we commit to support the next vendor’s transition in the same way as we will hope to be supported as we begin.

As part of contract finalization, AIR will also

- deliver a transition plan and final report,
- return and/or purge data from our systems,
- supply a final invoice designated as such.

This will ensure that Utah retains the value of the investment that it makes in this new testing program.
V. Proposal Content Requirements
Based in Washington, DC, the American Institutes for Research (AIR) is a not-for-profit firm providing research, assessment services, and technical assistance. We are 1,600+ people working to improve lives in communities at home and around the world. Our mission calls for us to turn the best research into practice, which we do in every arena in which we work.

AIR has demonstrated the ability to be the Utah vendor-partner that reaches beyond state-of-the-art practices and delivers innovative models, materials, and strategies that advance the field of measurement and improve the validity of student scores. In student assessment, our commitment is evidenced by our success delivering innovations, including the following:

- The first and only NCLB-approved, full-year, multiple-opportunity, online adaptive assessments, now operational in four states
- The first statewide online adaptive formative assessments delivered in the same system as the summative assessments and measuring the same constructs on the same scale
- The first NCLB-approved adaptive alternate assessments that measure growth for students with significant disabilities that are as psychometrically sound as the assessments for the general population

AIR has 60 years of experience in aptitude and proficiency testing, including 37 years in K–12 educational assessments. Over the years, AIR staff have worked on 40+ state and national assessments (e.g., NAEP, the National Assessment of Adult Literacy, the Voluntary National Tests), on international assessments (e.g., TIMSS, PISA), and on assessments in

, and other countries. AIR has successfully supported NCLB-compliant state assessment projects in , and . These contracts involve more than 2 million students per year and are valued at more than $700 million. Our Assessment Program also includes paper/pencil projects supporting both grades 3–8 achievement assessments and high school graduation tests. Other smaller Assessment Program projects include tests of English language acquisition on paper in and via computer in , multiple research grant-funded projects with states on modified assessments, and other research and support projects.

Exhibit V.A-1 presents a summary of relevant AIR and DRC assessment projects. Below, we highlight several projects as evidence of AIR’s relevant significant expertise in large-scale test development and record of accomplishment in the development and refinement of high-quality assessment instruments.

### Relevant Project Experience

**Assessment of Knowledge and Skills ( ) 2007–2014.** AIR administers ’s Assessment of Knowledge and Skills Online, or Online, the first NCLB-approved online adaptive state assessment system. AIR delivers approximately 2 million adaptive tests annually to more than 335,000 students in grades 3–12 in mathematics, reading, science, social sciences, and writing in a 9-month testing window each year. AIR also helped develop and delivers ’s Online English Language Proficiency Assessments (ELPA). Beyond excellent performance are these highlights:

- AIR developed and implemented the first adaptive Braille testing.
Exhibit V.A-1: Summary of Relevant Assessment Projects, AIR

<table>
<thead>
<tr>
<th>Program</th>
<th>Relevance to Utah</th>
<th>American Institutes for Research</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smarter Balanced Assessment Consortium Test Delivery System</td>
<td>Administering online, adaptive, pilot and field tests for ~6m students in 23 states; development and delivery of the open-source Test Delivery System</td>
<td></td>
<td>2012–2014</td>
</tr>
<tr>
<td>Smarter Balanced Assessment Consortium Item Developing and Machine Scoring</td>
<td>Developing pilot and field-test items for Smarter Balanced; responsible for development of open-source machine scoring engines</td>
<td></td>
<td>2012–2014</td>
</tr>
<tr>
<td>Assessment of Knowledge and Skills</td>
<td>Full-year, multiple-opportunity, online adaptive summative and interim assessments for all students</td>
<td></td>
<td>2007–2014</td>
</tr>
<tr>
<td>Comprehensive Assessment System</td>
<td>Full-year, multiple-opportunity, online adaptive summative, interim, and formative assessments for all students</td>
<td></td>
<td>2009–2014</td>
</tr>
<tr>
<td>End-of-Course Assessments</td>
<td>Online adaptive summative assessments for all students</td>
<td></td>
<td>2012–2014</td>
</tr>
<tr>
<td>Graduation Tests</td>
<td>Full paper/pencil testing program, with online adaptive formative assessments</td>
<td></td>
<td>2007–2014</td>
</tr>
<tr>
<td>Achievement Assessments</td>
<td>Full paper/pencil testing program, with online adaptive formative assessments</td>
<td></td>
<td>2002–2016</td>
</tr>
<tr>
<td>Assessment System</td>
<td>Online adaptive summative assessments for all students</td>
<td></td>
<td>2011–2015</td>
</tr>
<tr>
<td>National Assessment of Educational Progress (NAEP)</td>
<td>Test development, psychometric analysis, validity studies</td>
<td></td>
<td>2002–2015</td>
</tr>
<tr>
<td>Alternate Assessment</td>
<td>Task-based adaptive assessment for students with disabilities</td>
<td></td>
<td>2004–2014</td>
</tr>
<tr>
<td>Alternate Performance Assessment</td>
<td>Task-based, adaptive assessment for students with disabilities</td>
<td></td>
<td>2006–2014</td>
</tr>
<tr>
<td>Comprehensive System's Alternate Assessment</td>
<td>Task-based adaptive assessment for students with disabilities</td>
<td></td>
<td>2011–2014</td>
</tr>
<tr>
<td>State Alternate Assessments</td>
<td>Task-based adaptive assessment for students with disabilities</td>
<td></td>
<td>2012–2016</td>
</tr>
<tr>
<td>Public Schools High School Transformation</td>
<td>Formative and summative assessments</td>
<td></td>
<td>2006–2012</td>
</tr>
</tbody>
</table>
### Exhibit V.A-1: Summary of Relevant Assessment Projects, DRC

<table>
<thead>
<tr>
<th>Program</th>
<th>Relevance to Utah</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Recognition Corporation</td>
<td>Handscoring for paper/pencil summative assessments (ELA and mathematics)</td>
<td>2011–2014</td>
</tr>
<tr>
<td>Reading and Mathematics Test</td>
<td>Handscoring for paper/pencil summative assessments (ELA and mathematics)</td>
<td></td>
</tr>
<tr>
<td>Standards-Based Assessments and High School Graduation Qualifying Exam</td>
<td>Handscoring for paper/pencil summative assessments (ELA, mathematics, and science)</td>
<td>2004–2016</td>
</tr>
<tr>
<td>Educational Assessment Program, Graduation Exit Examination, and Alternate Assessments</td>
<td>Full-year, multiple-opportunity, online adaptive summative and interim assessments for all students</td>
<td>2001–2015</td>
</tr>
<tr>
<td>Assessments</td>
<td>Handscoring for paper/pencil and online summative assessments (ELA, mathematics, and science)</td>
<td>2011–2015</td>
</tr>
<tr>
<td>State Accountability Writing Assessment</td>
<td>Handscoring for paper/pencil and online writing assessments</td>
<td>2010–2013</td>
</tr>
<tr>
<td>Graduation Tests</td>
<td>Handscoring for paper/pencil graduation tests (ELA, mathematics, and science)</td>
<td>2007–2014</td>
</tr>
<tr>
<td>System of School Assessment</td>
<td>Handscoring for paper/pencil summative assessments (ELA, mathematics, and science)</td>
<td>1999–2014</td>
</tr>
<tr>
<td>Keystone Exams</td>
<td>Handscoring for online and paper/pencil end-of-course assessments (ELA, mathematics, and science)</td>
<td>2009–2015</td>
</tr>
<tr>
<td>Palmetto Assessment of State Standards</td>
<td>Handscoring for paper/pencil summative writing assessment (formerly known as the Palmetto Achievement Challenge Tests)</td>
<td>1998–2013</td>
</tr>
<tr>
<td>High School Assessment Program</td>
<td>Handscoring for paper/pencil exit examinations (ELA and mathematics)</td>
<td>2008–2013</td>
</tr>
<tr>
<td>English Language Development Assessment</td>
<td>Handscoring for ELDA paper/pencil assessment</td>
<td>2006–2017</td>
</tr>
<tr>
<td>Comprehensive Assessment Program</td>
<td>Handscoring for summative assessments (paper/pencil and online), high school proficiency exams, and end-of-course tests (ELA, mathematics, and science)</td>
<td>2008–2015</td>
</tr>
</tbody>
</table>
AIR’s Psychometrics and Statistics team worked to help develop technical reports and design and implement studies that were submitted for the state’s ultimately successful NCLB peer review.

From 2007 to 2012, AIR field-tested approximately 15,000 items with a rapid turnaround time; some field-test items were analyzed and inserted into the operational item pool within a month of being field-tested.

AIR’s item banking system is used by teachers to develop items.

**Comprehensive Assessment System ( ) 2009–2014.** AIR developed and delivered the online Comprehensive Assessment System ( ). Includes assessments for 100,000 students in grades 2–10 in reading, mathematics, science, and social studies and end-of-course exams in multiple subjects. It includes fully integrated benchmark and formative assessments and expansive learning supports. AIR also developed and delivers the adaptive, task-based Alternate Assessment for Students with Disabilities. These are some highlights:

- More than 99% of students have tested online since the initial statewide field test.
- Fully-adaptive Braille testing has been implemented.
- AIR developed 2,400+ new items, including hundreds of technology-enhanced, machine-scorable, and natural language items.

**State Assessments (HSA) 2006–2014.** AIR delivered the paper-based State Assessments (HSA) from 2006 through 2010, before the HSA moved completely online for the 2010–2011 school year. We currently test 120,000 students each year in reading, mathematics, and science at grades 3–8, 10, and 11, including ~1,000 technology-enhanced, machine-scored constructed-response items. We also administer the assessment for students taught in the Language Immersion Program at grades 3 and 4. Earlier in 2012, AIR won contracts to develop online, adaptive high school end-of-course assessments and our adaptive, task-based Alternate Assessment for Students with Disabilities. Both of these new assessments will be operational by spring 2013. Beyond excellent performance are these highlights:

- More than 99% of students have tested online since the initial statewide field test.
- AIR imported nearly 12,000 English and Spanish existing items and aligned them to new standards.
- AIR developed 4,000+ new items aligned to standards, including hundreds of technology-enhanced, machine-scorable, and natural language items.
- There has been exponential growth in the use of AIR’s Learning Point Navigator formative system for instructional resources and diagnostic assessments.

**Assessment System 2011–2015.** AIR successfully completed the initial year of supporting the online and paper/pencil Assessment System for approximately 500,000 students. We support testing for the Comprehensive Assessments (s) in reading, mathematics, and science in grades 3–8, 10, and 11; monthly Graduation-Required Assessments for Diploma (GRAD) in writing, reading, and mathematics; and alternate assessments ( -Modified and the Test of Academic Skills ([ ])). All of these have online components, including multiple-attempt adaptive testing in long windows and fixed-form testing in short windows, based on subject. Beyond excellent performance are these highlights:
AIR delivered ’s first adaptive assessments.

AIR delivered ’s highly technolo- gy-enhanced science assessments with 90% less bandwidth requirements than the previous contractor, significantly enhancing school satisfaction with the assessment.

**Smarter Balanced Assessment Consortium 2012–2014.** Through two prime contracts and one subcontract, AIR is supporting the Smarter Balanced Assessment Consortium in seven major activities, covering research into new item types; development of several thousand items; delivery of the pilot and field tests to up to 6 million students in 20+ states; and development of the core, open-source test delivery system and automated scoring technology. To date, most of the work has been focused on item research and development and the automated scoring engines. These are some highlights:

- Working with teacher focus groups from across the Consortium, AIR developed approximately 1,700 math and ELA items for the spring 2013 pilot administration.
- AIR and its partners have conducted approximately 700 cognitive labs in the Consortium states of California, , , , Michigan, Connecticut, South Carolina, Iowa, and Idaho.

**Assessment System 2002–2016.** AIR successfully completed more than a decade of continuous support to , including item development and test delivery for more than 800,000 students. We support testing for the achievement tests in reading, mathematics, and science in grades 3–8, graduation tests in five subjects; English Language Proficiency testing; and alternate assessments for students with disabilities. Beyond excellent performance are these highlights:

- In 2012, AIR scored 1.75 million tests for more than 800,000 students.
- AIR implemented for the first time our enhanced, full-color, variable paper score reports.
- AIR delivered ’s first online pilot tests. We implemented the transition from a portfolio alternate assessment to our adaptive alternate assessment measuring growth.

We have included detailed project descriptions at the end of this section.

As a result of the quality of our research, assessment, technical assistance, and dissemination efforts, as well as the dedication and hard work of our employees, AIR has experienced tremendous growth in the past decade. In FY 2012, our revenues exceeded $300 million, and, as a not-for-profit corporation, all of our net assets are retained by the organization to support future work. Unrestricted net assets currently exceed $120 million.

**Letters of Reference**

Exhibit V.A-2, over the next few pages, presents references from our relevant statewide work.
Exhibit V.A-2: Letters of Reference (continued)
Exhibit V.A-2: Letters of Reference (continued)
Specific RFP Experience Requirements

As this point, the RFP asks for examples of experience in specific areas.

History of Meeting Deadlines and Satisfying Contract Requirements

AIR is proud of our track record and references for meeting deadlines and satisfying contract requirements.

Beyond just meeting requirements and deadlines, AIR has a record of extraordinary support, particularly in the transition of new testing programs. Delivering fast transitions with quality results is not unusual in AIR’s experience. During the past seven years, AIR has successfully transitioned more than a dozen separate testing programs in , , , , , , and from other contractors, and AIR met all key deliverables. For example,

- awarded AIR the contract for the assessment system in April 2011. AIR was able to meet all deliverables, starting with a November 2011 online administration, despite not receiving the item bank in a standard electronic format from the previous contractor.

- intended to make a contract award for its statewide, adaptive online testing system in September or October 2006. The award was made to AIR on March 22, 2007. Despite the delay, AIR delivered the testing system for operational testing to begin in October 2007.

- Twice AIR supported in rapid quality fashion. Despite a four-month delay in the contract award for its high-stakes high school graduation test and a four-month delay in the contract award for achievement testing, AIR completed all item development on time, including committee reviews, and made the first testing windows without incident.

- made the contract award for its new alternate assessment contract six months later than anticipated. Even though this was an entirely new assessment design requiring AIR and to develop new extended standards and completely new types of items, AIR met all deliverables and timelines of the original RFP.

- made the contract award for the State Assessment 6 months later than anticipated. The award was made to AIR on July 19, 2006. AIR met all deliverables, including creating ‘s first-ever electronic item bank (importing thousands of items) and delivering 2,100 new items aligned to newly approved state standards for committee review meetings within three months.

As these and many other projects demonstrate, AIR meets our timelines and deliverable requirements, including situations when aggressive timelines for project startup/transition...
tion involve large item development efforts, creating new item banks from other sources, transitioning to entirely new standards or new assessment designs, meeting field-test windows, and delivering online testing systems. We offer this track record as evidence to support our commitment and our plan to meet Utah’s timelines and requirements.

**Minimum Required Skills**

**Item Development Subject Area Expertise at Target Grade Levels**

AIR takes great care in all project assignments, including targeting item development specialists with experience in the grade levels they will support. AIR has assigned an overall Item Development Manager (Robert Hodgman) and then created three teams by subject area. Each subject area has an overall lead, and then a lead by grade band: 3–5, 6–8, and high school. As shown in Exhibit V.A-3, each of the assigned individuals has specific, relevant, and recent experience in item development at the proposed grade-band.

AIR presented qualifications of all staff in the Project Management plan, and all resumes are included in Appendix E.

**Advanced Psychometrics and Statistics Expertise and Ability to Develop Technical Manuals**

AIR’s Psychometrics and Statistics team brings three specific qualities that uniquely position us to support Utah as it implements this state-of-the-art system and seeks approval from the U.S. Department of Education (USED):

- Our Psychometrics and Statistics team brings experience with online, adaptive, multiple-opportunity systems including technology-enhanced items in , , , and . This experience has afforded us the opportunity to explore and understand the different operational psychometric issues that arise in such a system and to build tools and procedures to address them.
- AIR alone has experience in helping a state obtain USED peer review approval for a testing system that is online and adaptive and offers multiple opportunities. Prior to implementing our online delivery system, contracted with AIR to design studies and prepare technical documents to support the state with its peer review with USED. These studies were instrumental in supporting the state in getting peer review approval.
- Our psychometricians and software engineers have developed the tools necessary to efficiently fine-tune the adaptive algorithm for specific blueprints and item pools. These tools can show the state how the system will perform before students start testing and provide the ability to make adjustments before problems arise.

AIR’s Assessment Program 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 examinees. AIR combines expertise in sampling and psychometrics; all of our samples are optimized, and our statistics accurately reflect the complexities of the sample designs.

In addition to developing ’s technical report for the first multiple-attempt adaptive system even before having the contract for the testing system, AIR regularly develops and
### Exhibit V.A-3: Item Development Experience by Grade-Band

<table>
<thead>
<tr>
<th>Name, Role</th>
<th>Competence, Experience, and Expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**English Language Arts**
delivers technical reports for all of our testing programs.

**Organizational Capability to Manage Large-Scale Test Development Projects with Relatively High-Stakes for Accountability**

AIR brings experience with both traditional and technology-based items for standards-based assessment using evidence-centered design, permissioned and commissioned stimuli, and a unique understanding of the nature of building an item bank for online adaptive testing.

AIR annually develops more than 10,000 items, including an ever-increasing number of innovative online interactive items. We have protocols for training staff and item writers that result in faithful adherence to each client’s content standards, style guides, and other policies and to the principles of universal design and language accessibility, bias, and sensitivity. Some of our clients have established detailed, published documentation of these requirements (e.g., ) before contracting with AIR, and we maintain or create this documentation as necessary for all clients.

AIR is part of the team developing the first set of next-generation assessment items and performance tasks for the Smarter Balanced Assessment Consortium pilot tests. AIR is working with educators to create assessment items that measure the critical-thinking and problem-solving skills that students need to be successful in college and careers. AIR is leading the research into new item types for the Consortium through cognitive laboratory studies and online small-scale trials; leading the development of technology-enabled test items; and developing the Consortium’s multiple new open-source scoring engines.

For more than a year, nearly 100% of AIR’s item development efforts with , , , South Carolina, , , and other clients have focused on developing items that measure the Common Core State Standards (CCSS). This has involved extensive workshops and training.
with our staff; client colleagues; outside item writers, and content, bias, and accessibility review committees. We have reviewed existing client item banks consisting of tens of thousands of items to support the realignment to the CCSS.

AIR delivers high-quality items. We have consistently experienced item acceptance rates above 90% in our state assessment contracts. On NAEP, we completed item development with NAGB acceptance rates above 95%.

AIR develops items that can be deployed in online testing systems in all schools, without placing large technology demands on the schools. Having developed and currently operating the only statewide, multiple-attempt adaptive online testing systems (now in four states), AIR understands the true requirements for computer-based test administration. AIR’s online testing state clients test nearly all of their students online. currently tests 99.99% of students online, and and tested online in all schools and nearly all students during the first year of online testing. We continue to innovate, adding online, adaptive testing in Braille during the most recent school year.

**Resumes**

Resumes are included in Appendix E.
As described in the introduction to our proposal, AIR offers

- a robust, online adaptive test delivery system that has proven effective, requiring little equipment, bandwidth, expertise, or support;
- a standards-based adaptive engine capable of supporting sophisticated blueprints, ensuring that all students have a fair chance to show what they know and can do across the full range of the curriculum;

- 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 through simulations; and

- the richest collection of embedded supports and accommodations in use in a statewide testing system to ensure access for all students.

We have proven this experience in the only four states with statewide online adaptive testing, as shown in Exhibit V.B-1. In addition to these states, AIR also supports online components of programs in other states, such as pilot and formative assessments in . Please note that AIR’s demonstrated experience in adaptive testing is from secure, statewide testing programs across multiple grades and subjects in real-world school environments similar to those in Utah.

### Exhibit V.B-1: Experience with Online Adaptive Statewide Testing

<table>
<thead>
<tr>
<th>Client/Program</th>
<th>Summative, Interim, and/or Formative</th>
<th>Grade Levels (K, 1, 2, etc.)</th>
<th>Volume of Tests Taken</th>
<th>Type of Test Administration Period (e.g., single date, test window)</th>
<th>Content Area(s) Tested</th>
<th>Item Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment of Knowledge and Skills ( )</td>
<td>Summative and Interim</td>
<td>3 through 12</td>
<td>~2m/year</td>
<td>9-month testing window, with multiple attempts (adaptive)</td>
<td>Math, Reading, Science, Social Sciences, Writing, and English Language Proficiency</td>
<td>MC, CR, TE</td>
</tr>
<tr>
<td>Comprehensive Assessment System</td>
<td>Summative, Interim, and Formative</td>
<td>2 through High School</td>
<td>~500k/year</td>
<td>Multiple attempts in three distinct testing windows (adaptive)</td>
<td>Math, Reading, Science, Social Studies, EOC, and Alternate Assessment for Students with Disabilities</td>
<td>MC, CR, TE, PT</td>
</tr>
<tr>
<td>State Assessment System</td>
<td>Summative and Interim</td>
<td>3 through 12</td>
<td>~550k/year</td>
<td>9-month testing window, with multiple attempts (adaptive)</td>
<td>Math, Reading, Science, EOC, and Alternate Assessment for Students with Disabilities</td>
<td>MC, CR, TE, PT</td>
</tr>
<tr>
<td>Assessment System</td>
<td>Summative and Interim</td>
<td>3 through High School</td>
<td>~1.4m, rising to 4+m/year</td>
<td>Multiple-attempt adaptive testing in long windows and fixed-form testing in short windows (based on subject)</td>
<td>Math, Reading, and Science for general population; Graduation testing; and Modified (2%)</td>
<td>MC, CR, TE, PT</td>
</tr>
</tbody>
</table>
Development and Implementation in a Variety of Technical Environments

As shown in our Technical Proposal, we support Mac, Windows, and Linux OS versions that go back more than a decade, and our minimum requirements for current computers are 64-128 MB RAM and 52-200 MB hard drives on very old processors. Starting in 2013, AIR will also support tablet devices.

We understand that schools often do not have the best or fastest equipment or networks, and they often have limited technical support available. AIR’s Test Delivery System has the smallest possible technology footprint in the industry and low end-user maintenance. We accomplish this with an architecture that

- is a fully web-based, adaptive test system;
- requires no external caching servers to manage;
- requires no extra third-party software;
- supports a wide range of OS and HW; and
- works on wired and wireless networks.

History of Meeting Requirements

High System Availability

AIR has delivered near 100% availability and above 99.9% for all clients. Further, our system is available at very high system performance. Students perceive no delays in navigating between items on a test. Specifically, average client latency ranges from .268 to .549 seconds across our testing programs, and with many thousands of simultaneous users.

<table>
<thead>
<tr>
<th>Client</th>
<th>Average Client Latency (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.549</td>
</tr>
<tr>
<td></td>
<td>.268</td>
</tr>
<tr>
<td></td>
<td>.315</td>
</tr>
<tr>
<td></td>
<td>.440</td>
</tr>
</tbody>
</table>

Based on AIR’s unique system design, AIR’s contract with our managed hosting provider (Rackspace) guarantees 100% uptime. We believe this is unique in the assessment industry.

Industry Standard Redundancy

We have designed our test delivery system 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 include the following:

- 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 state’s installation is served by multiple web servers, any one of which can take over for an individual test upon failure of another.
Active/passive and active/active clusters of database servers are configured so that the second node takes over in the event of failure of the first 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.

Our systems are monitored by a series of real-time performance monitoring and alerting systems. Our network engineers receive real-time alerts as server or network resources approach (configurable) critical use levels, experience exceptions or special events, or fail.

**High Concurrent User Capacity**

Our test delivery and related systems are entirely scalable by adding hardware. They can be configured to serve any number of students. Our database has been load-tested, and one redundant pair of database servers can serve over 40,000 simultaneous users. For each client, we project load and then add database and web servers to meet the projected demand, with enough excess capacity to provide comfort. For example, in , which has approximately 40,000 students per grade, the client requirement for simultaneous users (number of students concurrently taking the test) is 10,000. We configured the system to support in excess of 25,000, and the actual demand has not yet exceeded 8,000, with peaks of about 60,000 students tested per day.

---

**Effective Quality Assurance Processes**

In the Quality Assurance Section of the Technical Proposal, we detailed our quality assurance processes for all aspects of a testing program, including the special needs of computer-adaptive testing, and we addressed the roles of AIR and USOE staff. In the Project Management Section of the Technical Proposal, we described the system of Schedules, Planning Documents, Tracking Documents, and Specification Documents that are core to our quality assurance processes.

These processes have been proven effective on all of AIR’s Assessment projects.

**Expertise in Application Lifecycle Development**

As evidenced by our industry-leading test delivery and related systems, AIR brings extensive expertise in software development across the application lifecycle. Led by and three senior Ph.D.-level system architects, AIR’s Computer and Statistical Sciences Center (CSSC) includes 80+ staff members dedicated to software development, structured into teams.

- Each software system has a lead application developer and a lead software project manager and team of developers assigned.
- CSSC has a 10-person network engineering team.
- AIR maintains a substantial quality assurance team, and our quality assurance process involves everyone from the project director through the developers and quality assurance engineers and leads.
- A full-time chief security officer holds a Certified Functional Continuity Professional (CFCP) from Disaster Recovery Institute (DRI) International professional certifications.
The CSSC team developed and maintains the following sets of software systems:

- Item banking
- Student registration
- Test delivery
- Quality monitoring
- Database of record
- Handscoring
- Online reporting
- Learning Point Navigator (formative assessment and instructional resources)
- Value-added modeling

**Proven Ability to Develop and Troubleshoot in Windows, Macintosh, and Linux Environments**

As noted above, we support Mac, Windows, and Linux OS versions that go back more than a decade, and are successfully delivering tests in our client states without requiring caching servers or other special hardware or software in the schools. Based on public information, we believe our client-side software makes the lowest demands on schools and districts relative to hardware, software, and technical expertise.

**Proven Ability to Support Large-Scale Help Desk Needs**

AIR’s help desk can be measured by our performance statistics:

- First-call resolution rates above 80%.
- Average call is answered in 22 seconds.
- Less than 1.3% of calls go to voicemail.

Of course, our first goal is to avoid calls and emails through providing training and documents and by continuously improving these tools. To that end, we have the goals of decreasing the number of inquiries every year of a contract. For our online testing projects active more than one year, the number of inquiries have decreased 30% in the past year.

Based on our track record, AIR was awarded a sole source contract for Smarter Balanced Assessment Consortium Tier 1 help desk support for the spring 2013 pilot test. This program anticipates 2 million students.

We proactively research, track, and train for the kinds of questions we get. For example, in fewer than 13% of the help desk inquiries are related to the online testing system. The highest number of inquiries are related to student registration, reporting, and the limited paper/pencil testing. With analysis and collaboration with our client, we are able to answer 84% of questions at Tier 1.

Our help desk software is web-based, so it is infinitely scalable and geographically distributed.
Detailed Project Descriptions
V.C. Development of a Project Plan

Our Detailed Response presents a specific plan for completing the requirements of this contract. Throughout the proposal, we present the detailed sequence of steps and quality assurance procedures that will lead to the successful implementation of Utah’s CAT assessment program.

A schedule of activities completes the project plan, and this detailed schedule appears in Appendix F. A summary of key activities and milestones appears in section IV.B. Our project plan provides an appropriately ambitious implementation timeline that will allow Utah to transition online starting with formative assessments in fall 2013 and operational field-testing of interim/summative assessments in spring 2014.
V.D Cost Proposal

Price Proposal

Exhibit V.D-1 presents AIR’s Fixed Prices. We have based our proposal on the following documents:

- Utah Statewide Computer Adaptive Assessment System Solicitation PR13015 dated August 20, 2012, including all attachments
- Attachment_1_ANSWER_TO_GENERAL_QUESTION_1
- FINAL_Adaptive_RFP_Q__A_MASTER
- Addendum_3__Budget_Year_Clarity

**Base Price $49,361,631**

This price matches our technical proposal.

**Optional Price-100% Machine-scored Assessments - $39,303,646**

As discussed in the Technical Proposal, AIR believes that hand-scoring is required to meet all of Utah’s content and DoK requirements. For example, despite claims to the contrary by advocates or vendors, validating the logic and reasoning open-ended, possibly multi-mode math communication is beyond the current capabilities of artificial intelligence scoring.

In our base bid:

- For ELA, the ECR item in the performance task will be scored by the essay scoring engine, but for the higher-stakes summative assessment will be evaluated and potentially routed to human scoring based on heuristics that we use to identify suspect scores. We have included resources to handscore 25% of the summative responses.

- For math, the ECR item on the performance task will always be human-scored, interim and summative.

This handscoring is a major portion of AIR’s price.

If Utah believes all of the content and DoK requirements can be met with machine scoring and considers such proposals an acceptable response to the RFP requirements, AIR stands ready and would expect to be evaluated on an equal basis. AIR can machine-score any type of item available from other vendors, and more. Our section on scoring describes our scoring engines, and our proposal includes constructed response items in which the student responds with a drawing, equation, written statements, sequence of interactions and choices or longer written texts.

AIR’s price, replacing hand-scored ECR items above, would be $39,303,646

Responses to RFP Requirements

Contract Amount

We have replaced the CRT administration in spring 2014 with an operational field test of the new Utah Statewide Computer Adaptive Assessment System, which will yield the scores for Utah’s accountability reporting for the 2013–2014 school year. As such, we have assumed the 4.7m from the CRT will be available.

Price Guarantee Period

Per the RFP and the Q&A (Question 108), AIR has prepared and submitted a cost reimbursement budget with a fixed fee.
AIR would be pleased to enter into an entirely fixed-price contract for the Utah Statewide Computer Adaptive Assessment System. In a fixed-price contract, as in all of AIR’s other statewide testing program contracts, our prices would be completely guaranteed for the entire term of contract.

In either case, we understand requests for adjustments must include sufficient justification.

**Work Made for Hire**

Consistent with the RFP and Q&A (75, 106, and elsewhere), AIR will retain ownership of all software and other preexisting intellectual property. In our proposal, AIR also has included item-sharing agreements with other states to build Utah’s item pools. These items will be available to Utah under the Item Sharing Agreement.

With these understandings, AIR understands that custom-developed test items, manuals, data, and other similar deliverables will be Work for Hire and become the property of USOE upon acceptance and payment.

---

### Standard Contract Terms and Conditions

The RFP and Q&A provide the potential Vendor the opportunity to submit alternatives to the Standard Contract Terms & Conditions. In reviewing the standard template that will become the basis for the contract award, AIR respectfully requests the opportunity to negotiate the following issues.

In Clause 8, Indemnification, we agree with the spirit of this section but take limited exception to the wording as too broad and lacking some of the customary indemnification procedures. We will agree to indemnification against all third-party claims for injuries or death to the extent such claims arise out of or result from our negligence while performing under this contract. We believe that contract claims related to performance are covered elsewhere. Further, for any indemnification, we could like to include customary related provisions that allow us to be notified promptly of any such claims, to receive reasonable cooperation in defending of such claims, and to control any such defense.
As allowed by Clause 8, AIR would request the inclusion of a Limitation of the Contractor’s Liability Clause, including a limitation of liability for anyone for whom the Contractor is responsible. We propose all damages, including Liquidated Damages, shall be capped at 10% of the annual Contract value for the year in which the event causing damages occurs. Under no circumstances shall either party be liable for any lost profits or other special, consequential, or incidental damages that may arise. We understand that the resulting Limitations of Liability clause will not apply to injuries to persons, including death, or to damages to property.

In Clause 16, AIR is not licensing, contracting, or selling any hardware, firmware, or software.

Finally, we request the force majeure language in Clause 25 be expanded to be consistent with the RFP definition included on Page 4 of the RFP under Liquidated Damages.

**Additional Terms and Conditions**

**Disclosure**

AIR has had no claims of nonperformance or breach of contract filed within the past five years. DRC has had no claims of nonperformance or breach of contract filed within the past five years. We would be happy to provide additional detailed information on our performance.

**Subcontractors**

AIR has proposed entering into subcontracts with DRC. We will provide USOE with a copy of the resulting subcontract within 60 days of the execution of the contract.

**Liquidated Damages**

AIR accepts the RFP’s Liquidated Damages, with the understanding that Utah is not intending to assess multiple or cumulative remedies for a single incident.

**Protected Information**

AIR has submitted the Confidentiality Claim Form and Redacted Versions of our proposal.

**Pricing for Selected Sections**

Following the RFP directions, we have completed the Budget Worksheet for each Section. We have made the following assumptions:

- The majority of AIR’s Project Management and Quality Assurance work (under “Other Contract Requirements”) is embedded within AIR’s work processes under Sections I-III.

- AIR is proposing a single set of data exchange specifications for student and administrator data for Interim and Summative (Section I) and Formative (Section III) assessment. If we are just awarded Section III, then we would need to shift these costs from Section I to Section III.

- There are multiple hand-offs between Sections I and II, including from item development (Section II) to field-testing and scoring (Section I) to data review (Section II) to Standard Setting (Section I). We have included the specific costs where the RFP lists the tasks, such as including Braille and text-to-speech costs under Section I.

- We assume all item transfers of new content developed by any vendor under Section II will be in QTI format. Thus, if we are awarded Section I, we will receive items from the Section II vendor in QTI format.
Assumptions Underlying Our Fixed Prices

We have documented our assumptions throughout this proposal and have relied on the Questions and Answers in every case where they clarify or further define the requirements of the RFP.

We have also based our proposal on the following assumptions:

- 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.

- In the event of a reduction or change in the scope of work or partial termination for convenience, AIR assumes 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.

- Any conflicts between the provided documents (RFP, Contract Form, 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 to determine the contract requirements.

Specific Assumptions

Meetings/Training Detail

The details for each type of meeting are shown in Exhibit V-1

Overall assumptions:

- We have assumed that 75% of meeting participants drive daily (mileage reimbursement), while 25% of meeting participants travel and will require reimbursement of long distance travel, hotel, per diem, and local transportation costs.

- All meetings fall in Section II except Specifications Review and Data Review, which fall in Section I, and LPN Formative Training, which falls in Section III.

Handscoring Detail

As noted above, our proposal includes significant handscoring resources and associated costs for both interim and summative assessments.

Operational Field Test, Spring 2014

ELA

- Each student will respond to one ECR item. AIR will score all responses to 20 ECRs for each of 9 grades. Handscores will be used (1) for the operational results in spring 2014 and (2) to train and validate scoring engine

- Each student will respond to one or two natural language items. AIR will score 500 responses for 20 natural language items for each of 9 grades to use to examine the comparability of human scores versus the revised rubrics (after rubric validation).

Math

- Each student will respond to one ECR item. AIR will score all responses to 20 ECRs for each of 12 grades/courses. Handscores will be used (1) for the operational results for spring 2014 and (2) for training and validity papers for future operational scoring.

Science

- Each student will respond to one or two natural language items. AIR will score 500 responses for 18 natural language items for each of 9 grades/courses to use to examine
## Exhibit V-1

<table>
<thead>
<tr>
<th>Meeting</th>
<th>Math</th>
<th>ELA</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Specifications Review</strong></td>
<td>4 groups of 8 participants</td>
<td>4 groups of 8 participants</td>
<td>4 groups of 8 participants</td>
</tr>
<tr>
<td>Year 1 only</td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>AIR staff, 5</td>
<td>AIR staff, 5</td>
<td>AIR staff, 5</td>
</tr>
<tr>
<td></td>
<td>3 days</td>
<td>3 days</td>
<td>3 days</td>
</tr>
<tr>
<td><strong>Passage Review</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 groups of 8 participants</td>
<td>4 groups of 8 participants</td>
<td>3 groups of 8 participants</td>
</tr>
<tr>
<td></td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>AIR staff, 4</td>
<td>AIR staff, 4</td>
<td>AIR staff, 4</td>
</tr>
<tr>
<td></td>
<td>2 days in year 1, 1 day in subsequent years</td>
<td>2 days in year 1, 1 day in subsequent years</td>
<td>2 days in year 1, 1 day in subsequent years</td>
</tr>
<tr>
<td><strong>Item Writer Training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For all years</td>
<td>40 teachers (no teacher travel)</td>
<td>40 teachers (no teacher travel)</td>
<td>40 teachers (no teacher travel)</td>
</tr>
<tr>
<td></td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>AIR staff, 2</td>
<td>AIR staff, 2</td>
<td>AIR staff, 2</td>
</tr>
<tr>
<td></td>
<td>1 day</td>
<td>1 day</td>
<td>1 day</td>
</tr>
<tr>
<td><strong>Content Review</strong></td>
<td>4 groups of 3</td>
<td>4 groups of 3</td>
<td>1 group of 4 in year 1, 4 groups of 3 in subsequent years</td>
</tr>
<tr>
<td></td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>AIR staff, 1 per grade/course, plus 1 IDM</td>
<td>AIR staff, 1 per grade/course, plus 1 IDM</td>
<td>AIR staff, 1 per grade/course, plus 1 IDM</td>
</tr>
<tr>
<td></td>
<td>4 days in year 1, 1 day in subsequent years</td>
<td>4 days in year 1, 1 day in subsequent years</td>
<td>3 days in year 1, 1 day in subsequent years</td>
</tr>
<tr>
<td><strong>Fairness Review</strong></td>
<td>3 groups of 8 participants in year 1, 1 group in future years</td>
<td>3 groups of 8 participants in year 1, 1 group in future years</td>
<td>3 groups of 8 participants in year 1, 1 group in future years</td>
</tr>
<tr>
<td></td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>AIR staff, 1 per group</td>
<td>AIR staff, 1 per group</td>
<td>AIR staff, 1 per group</td>
</tr>
<tr>
<td></td>
<td>5 days in year 1, 3 days in future years</td>
<td>5 days in year 1, 3 days in future years</td>
<td>5 days in year 1, 3 days in future years</td>
</tr>
<tr>
<td><strong>Parent Review</strong></td>
<td>2 groups of 15 participants in year 1, 1 group in future years</td>
<td>2 groups of 15 participants in year 1, 1 group in future years</td>
<td>1 group of 15 participants in year 1, 1 group in future years</td>
</tr>
<tr>
<td></td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
<td>Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>AIR staff, 1 per group</td>
<td>AIR staff, 1 per group</td>
<td>AIR staff, 1 per group</td>
</tr>
<tr>
<td></td>
<td>14 days in year 1, 4 days in future years</td>
<td>16 days in year 1, 4 days in future years</td>
<td>4 days in year 1, 4 days in future years</td>
</tr>
<tr>
<td>Meeting</td>
<td>Math</td>
<td>ELA</td>
<td>Science</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------</td>
</tr>
<tr>
<td>Rangefinding</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
</tr>
<tr>
<td></td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>■ AIR staff, 3</td>
<td>■ AIR staff, 3</td>
<td>■ AIR staff, 3</td>
</tr>
<tr>
<td></td>
<td>■ 2 days in year 1, 1 day in future years</td>
<td>■ 2 days in year 1, 1 day in future years</td>
<td>■ 2 days in year 1, 1 day in future years</td>
</tr>
<tr>
<td>Rubric Validation</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
</tr>
<tr>
<td></td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>■ AIR staff, 6</td>
<td>■ AIR staff, 6</td>
<td>■ AIR staff, 6</td>
</tr>
<tr>
<td></td>
<td>■ 3 days in year 1, 2 days in future years</td>
<td>■ 3 days in year 1, 2 days in future years</td>
<td>■ 3 days in year 1, 2 days in future years</td>
</tr>
<tr>
<td>Data Review (All subjects held concurrently)</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
</tr>
<tr>
<td></td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>■ AIR staff, 4</td>
<td>■ AIR staff, 4</td>
<td>■ AIR staff, 4</td>
</tr>
<tr>
<td></td>
<td>■ 4 days in year 1, 1 day in future years</td>
<td>■ 4 days in year 1, 1 day in future years</td>
<td>■ 4 days in year 1, 1 day in future years</td>
</tr>
<tr>
<td>Performance Level Descriptors (All subjects held concurrently) Year 1 only</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
<td>■ 3 groups of 8 participants</td>
</tr>
<tr>
<td></td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>■ AIR staff, 4</td>
<td>■ AIR staff, 4</td>
<td>■ AIR staff, 4</td>
</tr>
<tr>
<td></td>
<td>■ 1 day</td>
<td>■ 1 day</td>
<td>■ 1 day</td>
</tr>
<tr>
<td>Standard Setting (All subjects held concurrently) Year 1 only</td>
<td>■ 91 participants</td>
<td>■ 45 participants</td>
<td>■ 87 participants</td>
</tr>
<tr>
<td></td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
<td>■ Substitute or participant payments, $100 per day</td>
</tr>
<tr>
<td></td>
<td>■ AIR staff, 18</td>
<td>■ AIR staff, 18</td>
<td>■ AIR staff, 18</td>
</tr>
<tr>
<td></td>
<td>■ 2.5 days</td>
<td>■ 3 days</td>
<td>■ 2.5 days</td>
</tr>
</tbody>
</table>

- All subjects and grade/courses held during a single week on the schedule presented in the Technical Proposal
- 18 AIR staff across content areas and psychometrics
the comparability of human scores versus the revised rubrics (after rubric validation).

**Operational Interim Testing, all years, beginning 2014-15**

Math
- Each student will respond to one ECR item on each of two interim attempts. AIR will score all responses.

**Operational Summative Testing, all years, beginning 2014–2015**

ELA
- USOE will select three essay items per grade for the summative pool each year. Each student will respond to one essay (human-scored constructed-response) item, which will be scored by the automated scoring engine that was trained and validated after the field test. AIR will re-score 25% of the responses that appear to be outliers, based on the criteria described in the technical proposal.

**Math**
- Each student will respond to one ECR item. AIR will score all responses.

**Adaptive Braille Testing Option**

Our proposal includes online, fixed-form summative testing for Braille that would be used for the course of the contract. This includes 50 items in each of the 27 grade/subjects or courses. As a cost option, Utah could choose to follow and prepare the entire item pool for adaptive interim and summative testing in Braille. The additional one-time cost for this option would be $676,947 to Braille all of the remaining items.

**Item Development and Item Sharing**

As described in our proposal, AIR is combining new item development with item sharing with (at least) and , which operate similar online, adaptive testing programs. Both and are developing math and ELA items for the common core, and therefore items are being directly developed to
measure the same core standards that form the basis of the Utah Core Standards.

While the agreement allows each state to select up to 40 items per year per grade/subject or course from each other state, or a total of 80, we only assumed Utah selected only 20 shared items per year for Years 2-5. AIR will develop new the remaining 30 items per grade/subject or course.

**Cost Elements**
Offerors bidding on this section must also complete the budget proposal for Addendum A, below.

### I. SUMMATIVE & INTERIM TECHNOLOGY

#### C. Technology Requirements

<table>
<thead>
<tr>
<th>Contract Year 1</th>
<th>Contract Year 2</th>
<th>Contract Year 3</th>
<th>Contract Year 4</th>
<th>Contract Year 5</th>
</tr>
</thead>
</table>

1. **Item Bank**
2. **Administration**
   - a. **Testing Time**
   - b. **Training**
   - c. **Test Administration Requirements**
   - d. **Helpdesk Support**
   - e. **Paper-Based Administration**
   - f. **Test Administration Manuals (TAMs)**

3. **Delivery System**
   - a. **Overall Approach**
   - b. **Data Exchanges**
   - c. **Adaptive Test Engine**
   - d. **Scoring**
   - e. **Report System**

4. **Technology Requirements**
   - a. **Application Installation and Maintenance**
   - b. **Security and Integrity**
   - c. **Supported Devices and Operating Systems**
   - d. **Hosting**

#### D. Psychometrics

<table>
<thead>
<tr>
<th>Contract Year 1</th>
<th>Contract Year 2</th>
<th>Contract Year 3</th>
<th>Contract Year 4</th>
<th>Contract Year 5</th>
</tr>
</thead>
</table>

1. **Overall Approach**
2. **Blueprint**
3. **Standard Setting**
   - a. **Proficiency Levels**
   - b. **Performance Level Descriptors**

Totals for this section if awarded separately.

### II. SUMMATIVE AND INTERIM TEST CONTENT

Offerors bidding on this section must also complete the budget proposal for Addendum A, below.

#### B. Requirements

<table>
<thead>
<tr>
<th>Contract Year 1</th>
<th>Contract Year 2</th>
<th>Contract Year 3</th>
<th>Contract Year 4</th>
<th>Contract Year 5</th>
</tr>
</thead>
</table>

1. **Overall Approach**
2. **Parental Review**
3. **English Language Arts** Item Development/Procurement
   - a. Items from an existing vendor system
   - b. Items from USOE's existing CRT item banks
   - c. New items developed for this proposal
   - d. Available open source consortium items

2. **Mathematics** Item Development/Procurement
   - a. Items from an existing vendor system
   - b. Items from USOE's existing CRT item banks
   - c. New items developed for this proposal
   - d. Available open source consortium items

2. **Science** Item Development/Procurement
   - a. Items from an existing vendor system
   - b. Items from USOE's existing CRT item banks
   - c. New items developed for this proposal
   - d. Available open source consortium items

3. **Annual Analysis, Review, and Revisions**
4. **Technology Requirements**

### ITEM DEVELOPMENT REQUIREMENTS (from Addendum A)

<table>
<thead>
<tr>
<th>Contract Year 1</th>
<th>Contract Year 2</th>
<th>Contract Year 3</th>
<th>Contract Year 4</th>
<th>Contract Year 5</th>
</tr>
</thead>
</table>

- **Workshops**
- **Item Preparation**
- **Item Style Guide**
- **Item Characteristics**
- **Item Alignment**
- **Universal Design Accessibility**

Totals for this section if awarded separately.
### III. Formative Assessments

#### B. Requirements

1. Overall Approach
2. Educator Functionality
3. System Functionality
4. Additional Requirements
   - 5. Formative Assessment Technical Requirements
     - a. Application Installation and Maintenance
     - b. Security and Integrity
     - c. Supported Devices and Operating Systems
     - d. Hosting

**Totals for this section if awarded separately:**

#### OTHER Contract Requirements

- **A. Return and Destruction of Items and Data**
- **B. Timetables**
- **C. Quality Assurance**
- **D. Project Management**
- **E. Technical manuals**
- **F. Contract finalization**

**Totals for this section:**

*Cost savings if awarded multiple sections: (See accompanying text)*

**Yearly totals:**

*Offerors bidding on multiple sections shall provide a narrative explanation of how an award on multiple or all sections of the RFP will result in enhanced value and cost savings to the State. This includes an explanation of where the cost savings will be derived.*

**Contract Total:**

---
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</th>
<th>II. SUMMATIVE AND INTERIM TEST CONTENT</th>
<th>III. FORMATIVE ASSESSMENTS</th>
<th>OTHER CONTRACT REQUIREMENTS</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Test Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychometrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online Testing and Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online and Paper Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Total Labor Costs         |                                                     |       |      |       |      |       |      |       |      |               |       |
### BUDGET - YEAR 1
**Sep-12 to Aug-13**

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</th>
<th>II. SUMMATIVE AND INTERIM TEST CONTENT</th>
<th>III. FORMATIVE ASSESSMENTS</th>
<th>OTHER CONTRACT REQUIREMENTS</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fringe Benefits - Grp 1</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Direct Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Direct Costs (see detail)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total ODCs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G &amp; A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Cost</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Price</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
<td>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</td>
<td>II. SUMMATIVE AND INTERIM TEST CONTENT</td>
<td>III. FORMATIVE ASSESSMENTS</td>
<td>OTHER CONTRACT REQUIREMENTS</td>
<td>TOTAL</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------</td>
<td>----------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Test Development</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Psychometrics</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Project Management</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Online Testing and Reporting</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Online and Paper Reporting</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
</tbody>
</table>

Total Labor Costs
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</th>
<th>II. SUMMATIVE AND INTERIM TEST CONTENT</th>
<th>III. FORMATIVE ASSESSMENTS</th>
<th>OTHER CONTRACT REQUIREMENTS</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fringe Benefits - Grp 1</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Direct Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Direct Costs (see detail)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total ODCs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G &amp;A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Cost</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Price</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### BUDGET - YEAR 3
Sep-14 to Aug-15

<table>
<thead>
<tr>
<th>Name</th>
<th>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</th>
<th>II. SUMMATIVE AND INTERIM TEST CONTENT</th>
<th>III. FORMATIVE ASSESSMENTS</th>
<th>OTHER CONTRACT REQUIREMENTS</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Onsite Overhead</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Test Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychometrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online Testing and Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online and Paper Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total Labor Costs

---

Solicitation: PR13015  
Title: Utah Statewide Computer Adaptive Assessment  
American Institutes for Research  
AIR Proposal #: 12-AST-010
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</th>
<th>II. SUMMATIVE AND INTERIM TEST CONTENT</th>
<th>III. FORMATIVE ASSESSMENTS</th>
<th>OTHER CONTRACT REQUIREMENTS</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fringe Benefits - Grp 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Direct Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Direct Costs (see detail)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total ODCs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G &amp;A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Cost</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Price</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
<td>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</td>
<td>II. SUMMATIVE AND INTERIM TEST CONTENT</td>
<td>III. FORMATIVE ASSESSMENTS</td>
<td>OTHER CONTRACT REQUIREMENTS</td>
<td>TOTAL</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Test Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychometrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online Testing and Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online and Paper Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total Labor Costs
### BUDGET - YEAR 4
**Sep-15 to Aug-16**

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</th>
<th>II. SUMMATIVE AND INTERIM TEST CONTENT</th>
<th>III. FORMATIVE ASSESSMENTS</th>
<th>OTHER CONTRACT REQUIREMENTS</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fringe Benefits - Grp 1</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Direct Costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Direct Costs (see detail)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total ODCs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G &amp; A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
<td>I. SUMMATIVE &amp; INTERIM TECHNOLOGY</td>
<td>II. SUMMATIVE AND INTERIM TEST CONTENT</td>
<td>III. FORMATIVE ASSESSMENTS</td>
<td>OTHER CONTRACT REQUIREMENTS</td>
<td>TOTAL</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
<td></td>
</tr>
<tr>
<td>Test Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychometrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online Testing and Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online and Paper Reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total Labor Costs: 0

BUDGET - YEAR 5  
Sep-16 to Aug-17
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fringe Benefits - Grp 1</td>
<td></td>
</tr>
<tr>
<td>Central Onsite Overhead</td>
<td></td>
</tr>
<tr>
<td><strong>Other Direct Costs</strong></td>
<td></td>
</tr>
<tr>
<td>Other Direct Costs (see detail)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
</tr>
<tr>
<td><strong>Total ODCs</strong></td>
<td></td>
</tr>
<tr>
<td>G &amp;A</td>
<td></td>
</tr>
<tr>
<td><strong>Total Cost</strong></td>
<td></td>
</tr>
<tr>
<td>Fee</td>
<td></td>
</tr>
<tr>
<td><strong>Total Price</strong></td>
<td></td>
</tr>
</tbody>
</table>
### Other Direct Costs

<table>
<thead>
<tr>
<th>Other Direct Costs</th>
<th>Section</th>
<th>Units</th>
<th>Rate</th>
<th>Contract Year 1</th>
<th>Contract Year 2</th>
<th>Contract Year 3</th>
<th>Contract Year 4</th>
<th>Contract Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handscoring Subcontractor</td>
<td>I</td>
<td></td>
<td></td>
<td>See text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staff Travel Management</td>
<td>I</td>
<td></td>
<td></td>
<td>See text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity Studies</td>
<td>I</td>
<td></td>
<td></td>
<td>See text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web Hosting and Licenses</td>
<td>I</td>
<td></td>
<td></td>
<td>See text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Webinars</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage Procurement</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web Hosting and Licenses</td>
<td>III</td>
<td></td>
<td></td>
<td>See text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meetings/Trainings</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meetings/Trainings</td>
<td>II</td>
<td></td>
<td></td>
<td>See attached chart</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meetings/Trainings</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consultants/Temporary Staff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Import Support</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text-to-Speech</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text-to-Speech</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text-to-Braille</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform Review</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help Desk</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Animation/Production</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Writers/Reviewers</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Writers/Reviewers</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject Matter Experts</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Import Support</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help Desk</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help Desk</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Subtotal by Sections

<table>
<thead>
<tr>
<th>Subtotal by Sections</th>
<th>Section</th>
<th>Units</th>
<th>Rate</th>
<th>Contract Year 1</th>
<th>Contract Year 2</th>
<th>Contract Year 3</th>
<th>Contract Year 4</th>
<th>Contract Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtotal</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal</td>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary of Meeting/Training Costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Year 1</td>
<td>Year 2</td>
<td>Year 3</td>
<td>Year 4</td>
<td>Year 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Year</td>
<td>Meeting Name</td>
<td>Number of Trips</td>
<td>Days per Meeting</td>
<td>AIR Staff</td>
<td>Participants</td>
<td>Travel Costs</td>
<td>Facility/AIR</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>---------------------------------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LD Travel @</td>
<td>MAVetc @ $50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$/day</td>
<td>attendees</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Specifications Review - Math</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>32</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Specifications Review - ELA</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>32</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Specifications Review - Science</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>32</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Passage Review - ELA</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Passage Review - Science</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Item Writer Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>120</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Content Review - Math</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>12</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Content Review - ELA</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>12</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Content Review - Science</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Fairness Review - Math</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Fairness Review - ELA</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Fairness Review - Science</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Parent Review - Math</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td>30</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Parent Review - ELA</td>
<td>1</td>
<td>16</td>
<td>2</td>
<td>30</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Parent Review - Science</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>30</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 1</td>
<td>LPN Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>120</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Range Finding - Math</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Range Finding - ELA</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Rubric Validation - Math</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Rubric Validation - ELA</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Rubric Validation - Science</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2</td>
<td>Data Review - Math</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2</td>
<td>Data Review - ELA</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2</td>
<td>Data Review - Science</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td>$/trip</td>
<td>$/day</td>
</tr>
<tr>
<td>Section</td>
<td>Year</td>
<td>Meeting Name</td>
<td>Number of Trips</td>
<td>Days per Meeting</td>
<td>AIR Staff</td>
<td>Participants</td>
<td>LD Travel @ $/trip</td>
<td>Travel Costs @ $/day</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>--------------------------------------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>-------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Performance Level Descriptors - Math</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Performance Level Descriptors - ELA</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Performance Level Descriptors - Science</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Standard Setting - Math</td>
<td>1</td>
<td>2.5</td>
<td>18</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Standard Setting - ELA</td>
<td>1</td>
<td>3</td>
<td>18</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Standard Setting - Science</td>
<td>1</td>
<td>2.5</td>
<td>18</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Train the Trainer</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Face to Face Training</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 2</td>
<td>LPN Training - ELA</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 2</td>
<td>LPN Training - Science</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Passage Review - ELA</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Passage Review - Science</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Item Writer Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Content Review - Math</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Content Review - ELA</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Content Review - Science</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Fairness Review - Math</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Fairness Review - ELA</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Fairness Review - Science</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Parent Review - Math</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Parent Review - ELA</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Parent Review - Science</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 2-5</td>
<td>LPN Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Range Finding - Math</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Range Finding - ELA</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Year</td>
<td>Meeting Name</td>
<td>Number of Trips</td>
<td>Days per Meeting</td>
<td>AIR Staff</td>
<td>Participants</td>
<td>LD Travel @ $/day</td>
<td>Travel Costs @ $/day</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-----------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Rubric Validation - Math</td>
<td>1</td>
<td>2</td>
<td></td>
<td>6</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Rubric Validation - ELA</td>
<td>1</td>
<td>2</td>
<td></td>
<td>6</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Rubric Validation - Science</td>
<td>1</td>
<td>2</td>
<td></td>
<td>6</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 3-5</td>
<td>Data Review - Math</td>
<td>1</td>
<td>1</td>
<td></td>
<td>4</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 3-5</td>
<td>Data Review - ELA</td>
<td>1</td>
<td>1</td>
<td></td>
<td>4</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 3-5</td>
<td>Data Review - Science</td>
<td>1</td>
<td>1</td>
<td></td>
<td>4</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Train the Trainer</td>
<td>4</td>
<td>1</td>
<td></td>
<td>3</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Year</td>
<td>Meeting Name</td>
<td>Number of Trips</td>
<td>Days per Meeting</td>
<td>AIR Staff</td>
<td>Participants</td>
<td>Airfare @ $</td>
<td>Daily Travel Costs @ $/day</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>--------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>--------------</td>
<td>------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Specifications Review - Math</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Specifications Review - ELA</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Specifications Review - Science</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Passage Review - ELA</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Passage Review - Science</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Item Writer Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Content Review - Math</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Content Review - ELA</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Content Review - Science</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Fairness Review - Math</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Fairness Review - ELA</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Fairness Review - Science</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Parent Review - Math</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Parent Review - ELA</td>
<td>1</td>
<td>16</td>
<td>2</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Parent Review - Science</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 1</td>
<td>LPN Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Range Finding - Math</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Range Finding - ELA</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Rubric Validation - Math</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Rubric Validation - ELA</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Rubric Validation - Science</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2</td>
<td>Data Review - Math</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2</td>
<td>Data Review - ELA</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2</td>
<td>Data Review - Science</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Year</td>
<td>Meeting Name</td>
<td>Number of Trips</td>
<td>Days per Meeting</td>
<td>AIR Staff</td>
<td>Participants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>------------------------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>--------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Performance Level Descriptors - Math</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Performance Level Descriptors - ELA</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Performance Level Descriptors - Science</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Standard Setting - Math</td>
<td>1</td>
<td>2.5</td>
<td>18</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Standard Setting - ELA</td>
<td>1</td>
<td>3</td>
<td>18</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Standard Setting - Science</td>
<td>1</td>
<td>2.5</td>
<td>18</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Train the Trainer</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 2</td>
<td>Face to Face Training</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 2</td>
<td>LPN Training - ELA</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 2</td>
<td>LPN Training - Science</td>
<td>1</td>
<td>2</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Passage Review - ELA</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Passage Review - Science</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Item Writer Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Content Review - Math</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Content Review - ELA</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Content Review - Science</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Fairness Review - Math</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Fairness Review - ELA</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Fairness Review - Science</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Parent Review - Math</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Parent Review - ELA</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 2-5</td>
<td>Parent Review - Science</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>Year 2-5</td>
<td>LPN Training - Math/ELA/Science - 40 participants each discipline</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Range Finding - Math</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Range Finding - ELA</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Year</td>
<td>Meeting Name</td>
<td>Number of Trips</td>
<td>Days per Meeting</td>
<td>AIR Staff</td>
<td>Participants</td>
<td>Airfare @ $/day</td>
<td>Daily Travel Costs @ $/day</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>----------------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>-----------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Rubric Validation - Math</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Rubric Validation - ELA</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Rubric Validation - Science</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 3-5</td>
<td>Data Review - Math</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 3-5</td>
<td>Data Review - ELA</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Data Review - Science</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 3-5</td>
<td>Train the Trainer</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 3-5</td>
<td>Train the Trainer</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>