Educational Testing Service
Response to Request for Information
for the Race to the Top Assessment Program
Assessment Technology Standards

January 2011

Prepared for:

U.S. Department of Education
Office of Innovation and Improvement
400 Maryland Avenue, SW
Washington, DC 20202

TABLE OF CONTENTS

3.2 Questions About Assessment Technology Standards – General and Market Questions

3.2.5 Customizing 3
3.2.6 Conformance and Testing 7
3.2.7 Best Practices 8

Technological Questions Regarding Assessment Technology Standards

3.2.8 Interoperable Assessment Instruments 11
3.2.9 Assessment Protection 12
3.2.10 Security and Access 13
3.2.11 Results Validity 15
3.2.12 Results Capture 16
3.2.13 Results Privacy 17
3.2.14 Anonymization 18
3.2.15 Scoring and Analysis of Results 21
3.2.15.1 Results Aggregation and Reporting 22
3.2.16 Sequencing 24
3.2.17 Computer-Driven Scoring 24
3.2.18 Formative, Interim, and Summative Assessments 26
3.2.19 Learning and Training 28
3.2.20 Repositories 29
3.2.21 Content Lifecycle 29
3.2.22 Interfaces and Services 30
3.2.23 Internal Transparency and Ease of Use 32
3.2.24 Discovery and Search 33
3.2.25 Metadata 34
3.2.26 Recommendation, Rating, and Review 35
3.2.27 Content and Media Diversity 36
3.2.28 Accessibility 40
3.2.29 English Learners 45

NOTICES

Certain information and data contained herein are proprietary and are marked as confidential in accordance with the RFI instructions. This information and data shall not be duplicated, used, or disclosed other than to evaluate this proposal. This restriction does not limit the government's right to use the information or data obtained from another source without restrictions.

Any request for release of information or data under any freedom of information request will be for the use of the individual requestor only, and will not constitute a license to publish or distribute, or a transfer of any rights in or to the information or data. Paragraphs and/or pages stamped "confidential" must be redacted from the final released version of the ETS proposal.

Copyright © 2011 Educational Testing Service

Copyright in this material is owned by Educational Testing Service, and limited dissemination has been made for the above purpose only. Except as provided above, no reproduction, use or distribution of any part of this material may be made without the specific authorization of Educational Testing Service.

3.2 Questions About Assessment Technology Standards — General and Market Questions

Educational Testing Service (ETS) respectfully submits this response.

Description of ETS's Involvement in Statewide Assessment Practices

ETS has been involved in K–12 assessment for decades. At the federal level, we have held contracts since 1984 to develop, administer, and report the National Assessment of Educational Progress (NAEP). Under contract to the College Board, we also develop national assessments that play important roles in K–12 education. These assessments include the *Advanced Placement Program*® (AP®), the *Preliminary SAT/National Merit Scholarship Qualifying Test* (PSAT/NMSQT®), and the SAT®. ETS develops national-level assessments for other clients, most notably for the Educational Records Bureau (ERB) and the Southern Regional Education Board (SREB). Moreover, at the state level, we are either developing or have developed statewide assessments for California, Florida, Georgia, Indiana, Maryland, Mississippi, New Jersey, Puerto Rico, Tennessee, Texas, Virginia, and Washington. In California, we also work with the California State University system to develop the Early Assessment Program (EAP), a college readiness supplement to end-of-course components of that state's assessment system.

In support of the assessment programs outlined above, ETS provides a full range of services to our clients. These services include psychometric research and statistical analysis, assessment development, program management, production and delivery (we have extensive experience in both paper- and computer-based delivery), communications, and policy analysis. We have participated significantly in the advances in technology as applied to assessment development and delivery, and have observed the impacts that technology plays in providing a richer and higher quality assessment experience.

ETS and the Development and Application of Technology Standards

ETS has been involved in standards development since the late 1990s, and we are a member of several leading standards organizations. Specifically, ETS has been a Contributing Member of the IMS Global Learning Consortium since 2000, and we were a key participant in the evolution of the IMS Question & Test Interoperability (QTI) specification. As described on **www.imsglobal.org**, IMS Contributing Members are the voting members and strategic partners of IMS who lead and ratify IMS standards. ETS also is a vendor member of the Schools Interoperability Framework (SIF) Association; a founding charter member of the Postsecondary Electronic Standards Council (PESC); and an associate member of the DAISY (Digital Accessible Information System) Consortium. Additionally, an ETS staff member co-edited the User Agent Accessibility Guidelines, which became a World Wide Web Consortium (W3C) Recommendation (**www.w3.org/TR/UAAG10/**).

ETS has experience in using technology standards in both high- and low-stakes environments. In high-stakes environments, we use the IMS QTI specification to transfer data between our enterprise-wide test registration, scoring, and reporting applications. In the K–12 market, ETS supplies approximately 64,000 items from our formative assessment item banks to a variety of vendors and several school districts in the QTI 2.0 format. The ETS item banking system also exports items and assessments in the QTI 2.1 standard for use in state assessments, and then delivers these items and assessments to vendors for both paper and online delivery. Presently, we are enhancing this system to support exchanging QTI 2.1 items and assessments bidirectionally, thus enabling item review in third-party systems in support of ETS state contracts. As a result of our participation in standards development, membership in standards bodies, and incorporation of standards into our own software, we have long-standing experience with the advantages and challenges of using assessment related standards.

ETS's responses to the "Questions About Assessment Technology Standards" incorporate more than a decade's worth of development and use of assessment standards in various testing systems and environments. The responses reflect our use of the standards in support of summative, interim, and formative assessments. We also provide examples of our projected future use of standards that include accessibility, gaming, and virtual environments. In addition, we have included examples of how standards provide value and areas where they fall short of the mark, with recommendations for improvement. With the Race to the Top Assessment (RTTA) program, the U.S. Department of Education has a unique opportunity to work with states, vendors, and standards organizations to accelerate the assessment standards development process, facilitate adoption of the standards, and strengthen the standards certification process so that they are implemented correctly by the parties involved. Ultimately, the RTTA program — supported by effective assessment technology standards — will decrease costs, improve efficiency, and increase quality.

In the pages that follow, ETS provides answers to the questions regarding assessment technology standards, as posed by the U.S. Department of Education.

3.2.5 Customizing

Can assessment tools developed under existing technology standards be customized, adapted, or enhanced for the use of specific communities of learning without conflicting with the technology standard under which a particular assessment tool was developed? Which technology standards provide the greatest flexibility in permitting adaption or other enhancement to meet the needs of different educational communities? What specific provisions in existing technology standards would tend to limit flexibility to adapt or enhance assessment tools? How easy would it be to amend existing technology standards to offer more flexibility to adapt and enhance assessment tools to meet the needs of various communities? Do final technology standards publications include flexible IP rights that enable and permit such customizations? What are the risks and the benefits of permitting such customization within technology standards? When would it make sense to prevent or to enable customization?

If we define "technology standards" as standards that allow the exchange of assessment and learning data between members of a community, then there is ample room in current standards to allow for the enhancement, modification, and adaptation of assessment tools and technologies. The primary reason for technology standards is for ease of interoperability and data exchange. As long as the communities involved in this exchange are sharing data and a more comprehensive certification process is in place (as recommended in **Section 3.2.6, Conformance and Testing**), their respective assessment tools and applications can be independently enhanced to accommodate new innovations.

Technology standards inherently do not limit assessment tools and applications, because the standards themselves are put in place for interoperability and exchange. The exchange via technology standards should be conceived of as an output of sending applications and as an input to receiving applications, and not as a limiting force to the inherent architecture of the applications themselves.

Which technology standards provide the greatest flexibility in permitting adaption or other enhancement to meet the needs of different educational communities?

QTI specification. A standard that has shown great flexibility for ETS is the Question & Test Interoperability (QTI) specification from IMS (**www.imsglobal.org/question/**). The schemas of the QTI specification are specific to assessment data, but remain abstract and generic. The combination of this specificity of the assessment domain with the abstract nature of the supported data elements provides a robust framework for interoperability and exchange between learning vendors and customers.

Application profiling. The IMS organization also provides specifications that learning vendors can use to define specific instances of interoperability. These specifications are referenced as "application profiling." Application profiling provides a framework for mutual definition and use of the QTI specification. The particular use of the QTI specification between each party within a learning community can be defined, developed, supported, and maintained by using the

application profiling process and tools. Please refer to **www.imsglobal.org/profile/profileapproach.cfm** for more information.

Through the use of application profiling, the community can adapt and enhance the QTI specification to their particular use while remaining in conformance to the standard itself. ETS has successfully leveraged the flexibility in QTI standards to distribute more than 64,000 unique assessment questions to many other organizations. The more than 20 entities that consume the QTI generated by the *ETS Formative Assessment Item Bank* range from local education agencies (LEAs) with homegrown platforms to major assessment companies with large presences in the K–12 formative market. The items have been loaded into online assessment systems, "clicker" response systems, pencil-and-paper solutions, and hybrid solutions. Using QTI, each company was able to receive a standardized delivery and then process the items. ETS was able to accomplish this task by creating a QTI standards document that described precisely how the standard was to be implemented. Sample item data in QTI format was provided to each entity for them to use as test data during their implementation of the standard. In addition, ETS provided advice and responded to questions to confirm the proper implementation of the QTI specification. The practical experience we gained from implementing QTI with multiple entities informed our response to **Section 3.2.6, Conformance and Testing**.

What specific provisions in existing technology standards would tend to limit flexibility to adapt or enhance assessment tools?

Given the collaborative nature of educational interoperability standards, only those elements that are deemed important to be included in the standard are actually authorized and supported. The risk is always present that assessment elements deemed important to a learning community are not contained in the standard. Although the QTI specification is under continual review and enhancement, the fact remains that as current or new assessment elements are identified, the standards body has not yet considered the element for inclusion. The rate of technological innovation in assessment may outpace the standards body's ability to authorize new elements. However, even if this is a continual risk, technology standards do allow for wide-ranging adaptation and customization.

Customization, in turn, introduces new risks (higher maintenance costs, less generic interoperability), but these risks are counteracted by the common point of departure that a learning community gains by adopting the standards. The standards represent a common starting point and not, necessarily, an immovable end point. As detailed in **Section 3.2.7, Best Practices**, changes to the standards can cause upgrade cycle issues, particularly for smaller companies with less capital. Any material change to the standards in use by a large group of companies will require an upgrade cycle window that is sufficient for the member with the fewest resources to implement.

How easy would it be to amend existing technology standards to offer more flexibility to adapt and enhance assessment tools to meet the needs of various communities?

A starting point. The standards themselves represent, as noted above, a starting point for sharing data. As communities interact and refine the standards for their particular use, the adaptation and customization of the standards will evolve. Standards do not dictate strict adherence as much as they provide a common starting point for interoperability. Assessment technology standards possess extensions that can be used as needed to provide flexibility for particular use cases.

Review and authorization process. The standards bodies must authorize formal amendments. The review and authorization process can take extended periods of time, even years, to resolve. The risk with any technology standard is that the rate of review and authorization does not keep pace with new assessment innovations and requirements. It may make sense for vendors participating in the RTTA program to agree upon their own extensions to the standard so that innovation is not compromised as the needs of the RTTA program evolve. These extensions can be made available to the standards organization that supports them.

Modifications. At the same time, mutual users of the QTI specification can modify the standard within the QTI parameters as needed. ETS has introduced a number of changes to the QTI 2.0 standard during the past two years. These changes include adding a manifest file for assessment passages and updating/adding metadata fields. Based on experience with a large group of companies, even a minor change to the standards requires a 9- to 12-month notification period. Typically, ETS informs companies 9 to 12 months out that a change is coming, and what the change will encompass. We give more details at 6 months, and a draft specification update at 3 months. One month before the final release, sample sets of items are delivered, and then after release, the companies have up to 3 months to finalize their import. Total deployment cycle for minor changes is 15 months.

Customization and adoption. Many companies that ETS works with could do this much more quickly, but in order to keep all companies on the same timeline, this extended period is needed. As a community customizes the standards to meet their objectives, the further away from the baseline standards the community veers. If, on the other hand, the community implements the standards as closely as possible to the published use, then the wider and more readily the adoption can be applied. Ease of adoption is proportional to the level of customization: As customization increases, easy adoption declines. These risks must be weighed against the value that the use of the standards possesses in the judgment of the community. Additionally, some customizations will cause QTI packages to fail standard validation checks using available tools. As changes to standards are needed, formal adoption of those changes and associated validation tools will be required.

When would it make sense to prevent or to enable customization?

Enabling and using customization would be recommended within a highly specialized arena of assessment delivery with few learning community vendors. In such an arena, the likelihood that the data being transferred must be contained in highly refined markup is high. The sending and receiving systems must use the data in highly specialized ways, and the need to customize based off of the standards is required.

Do final technology standards publications include flexible IP rights that enable and permit such customizations?

Technology standards are built to provide flexibility for many uses; they are conceived of as guides, not a set of inflexible rules. The IP considerations, therefore, are liberal in nature and provide for customization and extension. Amendments to QTI standards can be made by agreement between users, and are allowed under the IP rights. ETS has successfully used this ability when needed for our deliveries.

What are the risks and the benefits of permitting such customization within technology standards?

The further a community customizes the standards to meet its objectives, the further away from the baseline standards the community veers. If, on the other hand, the community implements the standards as closely as possible to the published use, then the wider and more readily the adoption can be applied. Ease of adoption is proportional to the level of customization: As customization increases, easy adoption declines. These risks must be weighed against the value that the use of the standards possesses in the judgment of the community. Additionally, some customizations will cause QTI packages to fail standard validation checks using available tools. As changes to standards are needed, formal adoption of those changes and associated validation tools will be required.

When would it make sense to prevent or to enable customization?

Enabling and using customization would be recommended within a highly specialized arena of assessment delivery with few learning community vendors. In such an arena, the likelihood that the data being transferred must be contained in highly refined markup is high. The sending and receiving systems must use the data in highly specialized ways, and the need to customize based off of the standards is required.

Prevention or limitation of customization would be recommended when the data being transferred are relatively simple, and the use of the data is envisioned to be transferred between many vendors and customers within a growing or changing learning community. In this instance, maintaining adherence to the prescribed and well documented uses of the standards would allow for ease of adoption and signup. ETS has minimized our changes to the QTI specification for the *ETS Formative Assessment Item Bank* for this reason. Experience has shown that changes to the specification can be supported by a diverse group of companies as long as sufficient lead time is given, and the changes remain minor.

3.2.6 Conformance and Testing

Do existing technology standards or technologies include specifications or testing procedures that can be used to verify that a new product, such as an assessment tool, meets the technology standards under which it was developed? What specifications or testing procedures exist for this purpose, e.g., software testing suites, detailed specification descriptions, or other verification methods? Are these verification procedures included in the costs of the technology standards, or provided on a free or fee-basis, or provided on some combination of bases?

Existing standards and technologies. Yes, well documented standards and technologies for product conformance testing exist. Three well known organizations that produce these are IMS, **www.imsglobal.org**; the Schools Interoperability Framework (SIF) Association, **www.sifinfo.org**; and the Advanced Distributed Learning (ADL) Initiative, **www.adlnet.gov**. All three organizations produce well documented specifications, which you can find on their websites, along with the descriptions of their certification procedures below:

IMS. IMS currently offers certification for two of its specifications: Common Cartridge and Learning Tools Interoperability. A self-service testing tool for Common Cartridge is available for free. However, to obtain an official conformance seal, organizations must join the Common Cartridge & Learning Tools Interoperability Alliance, which requires a fee. Under the Alliance, IMS provides extensive support to implement conformant products. When testing is completed, IMS issues a conformance mark that the organization stands behind. As a result, IMS will get directly involved in any issues that arise, and will work across organizations to resolve those issues.

SIF. SIF has a test suite that vendors can use to test their products before submitting those products for official SIF Certification. The SIF Certification is fee based, and is performed by a third party SIF Certification Authority, the Open Group (**www.opengroup.org**). A vendor is not required to be a member of SIF, but these fees apply to both members and nonmembers of SIF. However, prior to commencing certification activity, nonmembers who desire SIF Certification must pay an annual surcharge to SIF. This surcharge is equal to the corresponding annual association membership dues based on the type of supplier and revenue level. Once the SIF Certification Authority has successfully completed the certification, the product is entered into the SIF Certification Register.

ADL. ADL has a test suite that vendors can use to test their products for compliance to the appropriate Sharable Content Object Reference Model (SCORM) standard before submitting those products for official SCORM certification. The SCORM certification is fee based, and is conducted by an approved ADL Certification Testing Center. This testing center evaluates the product to certify that it has achieved SCORM Compliance.

Strengthening the certification process. In addition to the certifications that we have described above, the RTTA program represents an excellent opportunity to strengthen the certification process. In general, current certification processes focus on how applications conform to a standard, but they do not address the quality of the data that they produce. Significantly, the quality of the data will be central

to the success of the RTTA program. Accordingly, we recommend strengthening the certification process by incorporating items, test forms, and candidate results that represent the Common Core State Standards.

Examples. For example, tasks and candidate results could be packaged according to the IMS Question & Test Interoperability (QTI) specification, and the candidate information packaged according to the IMS Learner Information Package (LIP) specification. Due to the complexity of the data transported in these formats, there is built-in flexibility and extendibility (and thus ambiguity) to the specifications. Because a common set of items will be shared among states, an enhanced certification process will establish that applications produced by multiple vendors will be based on a consistent interpretation of the standards. In addition, this enhanced certification process will help produce consistent results across vendor products, from item authoring through test delivery and results reporting. This improved certification will minimize the guesswork needed by a vendor who is adopting the specifications.

Please see **Section 3.2.5, Customizing**, for an example of how ETS achieves consistent interpretation and implementation of the QTI specification across the multiple organizations we interoperate with.

3.2.7 Best Practices

What are best practices related to the design and use of assessment interoperability technology standards? Where have these best practices been adopted, and what are the general lessons learned from those adoptions? How might such best practices be effectively used in the future?

What are best practices related to the design assessment interoperability technology standards?

We have provided the following examples of such best practices:

1. Participating vendors and customers in the learning community must share common agreement that the standards form the basis for interoperability.
2. Equal weight and consideration should be given to both stakeholders and the standards bodies to make decisions concerning formal amendments to standards. Typically, the stakeholders are closer to real life commercialization of the standards, and they have firsthand knowledge of the types of changes that will provide the most value. The standards bodies possess the platform for receiving the disparate suggestions. The closer the cooperation between stakeholders and the standards bodies, the greater the opportunity for widespread and value laden adoption.
3. Work groups across the learning communities must possess both information technology staff and subject matter experts (assessment development and business staff) to provide the decision making process for the adoption, use, and possible customization of the standards.

What are best practices related to the use of assessment interoperability technology standards?

The work groups must use a predefined software development lifecycle approach for development and production release (waterfall, agile, etc.). The type of approach is less important, than that one be used. A methodical approach to implementation is required because interoperability, by its very nature, requires tight integration and well defined connection points.

In addition, there must be a testing and conformance process in place, as recommended in **Section 3.2.6, Conformance and Testing.**

Where have these best practices been adopted, and what are the general lessons learned from those adoptions?

ETS selected the QTI specification two years ago for use with the *ETS Formative Assessment Item Bank*. In the processes of delivering up to 64,000 items to 23 companies and up to three times that to some companies, we have learned a number of general lessons. Even with an interoperable technology standard in place, there are issues to address when delivering assessment items between companies.

Differing technology levels. Differing technology levels among the companies is the first issue when implementing technology standards. With ongoing upgrades to established specifications, all companies must be using the same version of specifications in order to communicate effectively. When ETS initiated use of the QTI 2.0 specification for use with the *ETS Formative Assessment Item Bank*, some companies were still using Version 1.2. Accordingly, they needed to upgrade their systems before they were able to consume assessment items from ETS.

This upgrade required significant effort and time for some companies, which delayed timelines for transfers. As ETS considers upgrading to QTI 2.1 for the *ETS Formative Assessment Item Bank*, the main impediment is confirming that all the companies we work with are ready to upgrade as well.

Limited technical resources. While larger companies in the K–12 market have the resources to upgrade to newer versions of specifications quickly, it can be a daunting prospect for smaller companies with limited technical resources. Any common technology standard being used by a large number of companies has the potential to require lengthy upgrade cycles when specifications are updated. As the community that is using the standard grows larger and more diverse, a more structured upgrade path with supporting documentation and tools needs to be provided to shorten upgrade times.

How might such best practices be effectively used in the future?

Thought should be given to making any standard as robust as possible to minimize the frequency and impact of future changes. At the same time, consideration must be given to the technology currently available in platforms. An example of this in QTI is the use of Mathematical Markup Language (MathML) to transmit MathType. While this is allowable under the QTI 2.0 specification, there are a number of companies that cannot support MathML, and these companies require graphics files instead of MathType. Considerations such as this one should be built into a specification to make it as universal as possible. In this particular case, realizing that not all systems can support MathType — and requiring that mathematical expressions be included in QTI as both artwork and MathML — would resolve the issue.

Technological Questions Regarding Assessment Technology Standards

3.2.8 Interoperable Assessment Instruments

What techniques, such as educational markup or assessment markup languages (see also http://en.wikipedia.org/wiki/Markup_language), exist to describe, package, exchange, and deliver interoperable assessments? How do technology standards include assessments in packaged or structured formats? How can technology standards enable interoperable use with resources for learning content? How can technology standards permit assessment instruments and items to be exchanged between and used by different assessment technology systems?

What techniques, such as educational markup or assessment markup languages (see also http://en.wikipedia.org/wiki/Markup_language), exist to describe, package, exchange, and deliver interoperable assessments?

The QTI specification uses XML schemas, manifest files, and extensions to communicate the framework. The manifest files can be viewed as a catalogue of everything contained in the QTI package. Individual or combined XML files then contain the actual item and test level content.

How do technology standards include assessments in packaged or structured formats?

The QTI schema defines item level elements and test level elements. The schema possesses a detailed hierarchy of item and test level elements that can be used to place assessment content into. As noted below, the schema allows for HTML and MathML markup to be incorporated as content, as needed. Images (art objects and files) can be incorporated into the XML as serialized objects with presentation rules or as separate images.

How can technology standards enable interoperable use with resources for learning content?

The Common Cartridge and Learning Tools Interoperability standards enable the interoperable use of learning content resources. You may find detailed information about these two standards at **www.imsglobal.org/cc/alliance.html**.

How can technology standards permit assessment instruments and items to be exchanged between and used by different assessment technology systems?

With the use of common markup languages like XML, HTML, and MathML, there is a low learning curve for companies wishing to exchange items using the QTI specification. Even with the specification in place, however, there can be issues with recognizing tags and allowable values for the tags. The tags should be as fully developed as possible from the start, to minimize the need to upgrade systems to support tag changes. Also, while HTML is a primarily standard markup language, limits on allowable tags in the QTI specification can cause issues with upgrade cycles if

those tags are later introduced. Ideally, the HTML support in any technical specification would be as robust as possible from the start.

With MathML, the main limitation, as previously noted, is the inability for some systems to processes it (or to even support MathType). In part, this is due to Internet Explorer's inability to display MathType without a browser plug-in. Most other major browsers support this functionality natively.

Cascading style sheets also are used to allow for contextual markup of assessment items. This use allows items adhering to the QTI specification to be formatted as required by end users. Some examples would be the title of a passage that might be left justified on one customer's tests and centered on another customer's tests. The use of cascading style sheets allows content from multiple vendors to be quickly integrated into a cohesive whole for use by customers.

3.2.9 Assessment Protection

For this RFI, "Assessment Protection" means keeping assessment instruments and items sufficiently controlled to ensure that their application yields valid results. (See also paragraph below, "Results Validity.") When assessment instruments or content are re-used or shared across organizations or publicly, are there capabilities or strategies in the technology standards to assist in item or instrument protection? What mechanisms or processes exist to ensure that assessment results are accurate and free from tampering? Do examples exist of public or semi-public assessment repositories that can provide valid tests or assessments while still sharing assessment items broadly?

When assessment instruments or content are re-used or shared across organizations or publicly, are there capabilities or strategies in the technology standards to assist in item or instrument protection?

The *ETS Formative Assessment Item Bank* is considered a low-stakes assessment tool, but our distributors use various forms of item tagging to provide protection of items as needed by the end user. One common use case is a local education agency (LEA) that wishes to use the *ETS Formative Assessment Item Bank* for LEA interim assessments, school level monthly assessments, and class level weekly assessments. The host platform prevents users from modifying original items, and also allows different user access levels to restrict items. Consequently, the LEA might build a series of test forms and restrict all those items from all other users. Then, the remaining items are released to school personnel, who then create monthly tests and restrict those items from all other users. Next, the remaining items are released to teachers to use in individual classes. This process provides each level of users who are accessing the system with a unique set of items for the test forms or quizzes they are creating.

What mechanisms or processes exist to ensure that assessment results are accurate and free from tampering?

The accuracy of the results can help to be confirmed by tracking changes to items and preventing changes after an item has been placed on an assessment. Access restrictions as detailed in **Section 3.2.13, Results Privacy**, also can be used as a tool to prevent tampering.

Do examples exist of public or semi-public assessment repositories that can provide valid tests or assessments while still sharing assessment items broadly?

Currently, the *ETS Formative Assessment Item Bank* is used by nearly 1.5 million students across the U.S. in more than 20 platforms. Due to this wide distribution, ETS requires that all distributor platforms make ETS items available to authorized users only, who can access the items only by using their specific login credentials from a secure connection. Also, we require ETS items to be protected from changes in all distributor systems so that items remain reliable. These requirements allow for a secure, stable pool of items for educators to use.

3.2.10 Security and Access

In what ways do technology standards provide for core security issues, such as access logging, encryption, access levels, and inter-system single-sign-on capabilities (i.e., one login for systems managed by different organizations)?

Common security standards. The most commonly referenced security standards are the ISO/IEC 27000 series (industry/commercial) and NIST SP 800-53 (government). Both are foundational and most useful as frameworks and "management systems" in support of comprehensive security programs. Layered upon this foundation are a number of technology specifications outlined by the International Standardization Organization/International Electrotechnical Commission (ISO/IEC) and the National Institute of Standards and Technology (NIST), including Information Technology Standards, Special Publications (SP), and Federal Information Processing Standards (FIPS). These specify protocols and configurations for authentication and security of access, and for collection, transmission, processing, storage, and output of data. At the enterprise level, ETS products and services are implemented in alignment with such public standards expectations.

Sensitive data transmissions. ETS uses the "Advanced Encryption Standard (AES) — 128" algorithm for data encryption in storage and during transfers recommend by FIPS-197. Sensitive data transmissions and web user interactions are always done using secure protocols such as SSLv3 and HTTPS. ETS applications authenticate users to a central Active Directory services database via Lightweight Directory Access Protocol (LDAP), which is an industry standard for accessing directory services over a network. Having the central user repository enables our organization to easily manage access control and also supports centralized logging, as well as event monitoring. The Common Event Format (CEF) standard attempts to improve the interoperability of various event- or log-generating devices by aligning the logging output from various technology vendors. ETS has implemented ArcSight's Security Information

and Event Management (SIEM) solution, which employs the CEF standard and allows for correlation of output from key devices for more effective and actionable logging and monitoring.

Because there are multiple aspects to security, beyond technologies, ETS has put in place processes, testing procedures (such as vulnerability and penetration testing), and management controls (such as policies, best practices, and standard operating procedures) to confirm that our applications, infrastructure, and data are adequately secured.

Single sign-on standards. For single sign-on, several standards such as SAML, Shibboleth®, OpenID, and OAuth are available. Each of these standards has strengths and weakness, but generally the standard to use in a software solution is primarily driven by the user community of the solution and the authentication authority (user account repository) of that user community. Considering the wide industry adoption, maturity, and product supportability, ETS chose to use SAML 2.0 for our authentication scheme whenever ETS is the authentication authority. Web applications offered to ETS users are integrated via SAML. ETS authenticates the end user and returns the necessary identity data required by each application to further apply the entitlements of the users to control access to functions. For the services offered by ETS, we have a flexible strategy to support various standards as required by the authentication authority.

References

ArcSight. (2011). *Common event format* (SEIM). Retrieved January 4, 2011, from ArcSight's website: www.arcsight.com/solutions/solutions-cef/

International Organization for Standards (ISO). (2010). *ISO/IEC 27000 series*. Retrieved January 4, 2011, from ISO's website: www.iso.org/iso/catalogue_detail?csnumber=41933

National Institute of Standards and Technology (NIST). (2001). *Specification for the advanced encryption standard (AES)* (Federal Information Processing Standards Publication 197). Gaithersburg, MD: Author. Available online here: http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

National Institute of Standards and Technology (NIST). (2009). *Recommended security controls for federal information systems and organizations* (Special Publication 800-53). Gaithersburg, MD: Author. Available online here: http://csrc.nist.gov/publications/nistpubs/800-53-Rev3/sp800-53-rev3-final_updated-errata_05-01-2010.pdf

Oasis. (2005). *SAML V2.0 executive overview* (Committee Draft 01, 12 April 2005). (P. Madsen & E. Maler, Eds.). Burlington, MA: Author. Available online here: www.oasis-open.org/committees/download.php/13525/sstc-saml-exec-overview-2.0-cd-01-2col.pdf

3.2.11 Results Validity

For this RFI, "Results Validity" means protecting the statistical validity and reliability of assessment instruments and items. How can interoperable instruments be managed to ensure they are administered in a way that ensures valid results? Are solutions regarding assurance or management of validity appropriate for inclusion in technology standards, or should they be addressed by the communities that would use the technology standards to develop specific assessments?

Process of validation. According to the widely accepted joint standards of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, & NCME, 1999), validity is the "the most fundamental consideration in developing and evaluating tests" (p. 9). Conceptions of validity are increasingly framed in terms of arguments — specifically, arguments about the proposed interpretations and uses of assessments (Kane, 2006). The process of validation involves integration of evidence from many sources, and it is ongoing, never being regarded as being fully concluded (Messick, 1989). The joint standards identify several different sources of validity evidence (i.e., response processes, internal structure, relations to other variables, consequences of testing) (pp. 11-17). Depending on factors such as the specific purpose of the assessment, the sample size, and other factors, different methods — quantitative or qualitative — may be employed to provide such evidence. Procedures involved might vary widely, ranging from the use of specific statistical analyses (e.g., item analyses) to pilot testing items, conducting cognitive labs (think-aloud studies), and having experts evaluate the content of items.

Challenges due to data diversity. Due to this great diversity of data, it may be challenging to develop overarching interoperability standards for results validity of assessments, though most if not all of the other sections of this document have a valuable role in establishing the validity of assessment results. It should be noted that evidence-centered design (ECD) approaches, which are argument-based, may be suited to integrating and clarifying considerations that bear upon validity (Mislevy, Steinberg, & Almond, 2003). (Please see **Section 3.2.28, Accessibility**, for information on ECD-based argument structures related to accessibility.)

ETS's quality standards. With regard to overarching procedures for establishing validity, currently each testing organization typically has its own standards and procedures. ETS has the *ETS Standards for Quality and Fairness* (2002), which are consistent with the joint standards (AERA, APA, & NCME, 1999), while also being suited to the specific requirements of ETS programs. Currently, the joint standards are being redrafted (**http://teststandards.net/Revision.htm**), and upon their publication, the *ETS Standards* will be revised to remain comparable.

ETS also has an audit process that examines compliance to the *ETS Standards*. Audits are conducted for each product (assessment or nonassessment) within a year after the launch of a product and every three years thereafter. Parts of this RFI response that address data security and integrity would provide essential support for documenting compliance to standards for validity of results from the RTTA assessments. Given the specific requirement of the RTTA program, it may be possible to identify

additional specific areas for interoperability standards, such as standards for exchange of data related to audits or other specific aspects of validity.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (1999). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

Educational Testing Service (ETS) (2002). *ETS standards for quality and fairness.* Princeton, New Jersey: Author. Available online here: www.ets.org/Media/About_ETS/pdf/standards.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington, DC: American Council on Education & National Council on Measurement in Education.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education & National Council on Measurement in Education.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

3.2.12 Results Capture

How can technology standards accurately link individual learners, their assessment results, the systems where they take their assessments, and the systems where they view their results? How do technology standards accurately make these linkages when assessments, content, and other data reside across numerous, distinct learning and curriculum management systems, sometimes maintained by different organizations?

Enhanced certification process. As noted in our response to **Section 3.2.6, Conformance and Testing**, it is an enhanced certification process that will establish that applications produced by multiple vendors will be based on a consistent interpretation of the standards, and will produce consistent results across vendor products from item authoring through test delivery and results reporting. This improved certification will minimize the guesswork needed by a vendor who is adopting the specifications. For example, a certified implementation of the IMS QTI specification and the candidate information packaged according to the IMS Learner Information Package (LIP) specification would provide vendors with the data elements required to link content, assessments, and results in the appropriate ways.

Seamless data sharing. Assessment technology standards such as the IMS Learning Tools Interoperability (LTI) standards could be used to create a seamless data sharing experience across the following: e-learning/course management systems, course authoring applications, collaboration applications, e-book applications, homework applications, assessment applications, e-portfolio applications, and more. Additionally, current conventional methods of transferring data among systems

and companies, as described in **Section 3.2.22, Interfaces and Services**, could be used because the data exchanged would be in assessment technology standards format.

3.2.13 Results Privacy

How do technology standards enable assessment results for individual learners to be kept private, especially as assessments results are transferred across numerous, distinct learning systems? How can such results best be shared securely over a distributed set of systems managed by independent organizations that are authorized to receive the data, while still maintaining privacy from unauthorized access?

The most commonly referenced standard for data privacy is the NIST SP 800-122. Country- and state-specific data privacy legislation and variation of requirements by jurisdiction add complexity. The Generally Accepted Privacy Principles from the American Institute of CPAs (AICPA), in addition to related technical guidance, are intended to assist in navigating this complexity.

The ETS Data Privacy Policy, technology standards, reference architecture, and corporate data model provide implementation guidance, while shared services implement key mechanisms across systems. Data transfer agreements, third-party service provider assessments, and independent audit/certification are key controls.

Technology standards, by themselves, are neutral with respect to assessment results because they are specifically created to form an interoperable base for sharing content between learning communities and systems. The security and privacy of material contained in the transport must be established between communities as a separate activity: The communities must together define a set of privacy and security protocols for the entire transaction: send, transport, and receipt. Many technologies exist to foster strong security and privacy, and the choices made should be governed by the sensitivity of the content and the nature of the applications engaged. Regardless of technologies used, sending and receiving applications must adhere to these shared security rules. Sending and receiving applications must possess internal security measures (authentication, identity management, and role based, to name a few). Additionally, transport mechanisms must be established (encryption, secure socket, application authentication, etc., as discussed in **Section 3.2.10, Security and Access**).

3.2.14 Anonymization

Do technology standards or technologies permit or enable anonymization of assessment results for research or data exchange and reporting? How do various technology standards accomplish these tasks? For example, where a number of students take a test, can their answers be anonymized (through aggregation or other techniques) and shared with researchers to examine factors related to the assessment (e.g., instructional inputs, curriculum, materials, validity of the instrument itself) without revealing the identity of the learners? Is this an area where technology standards can help?

A critical component. Anonymization of data can be a critical component of any assessment delivery and reporting system. There are technical solutions to uniquely identifying a record. Moreover, there are statistical/policy steps that can be used to protect the identity of respondents in a given dataset.

Universally Unique Identifiers. Common, nonproprietary technology standards exist to keep data anonymous for reporting and sharing purposes through the judicious use of Universally Unique Identifiers (UUID).

A UUID is a 16-byte (128-bit) number. The number of theoretically possible UUIDs is, therefore, about 3×10^{38}. In its canonical form, a UUID consists of 32 hexadecimal digits, displayed in five groups separated by hyphens, in the form 8-4-4-4-12 for a total of 36 characters (32 digits and four hyphens). An example is, "7780e8400-d20b-41e4-a873-446455346090."

Guaranteed unique. UUIDs are important because they are guaranteed unique across space and time, given the large number of combinations that are available.

The source system, or the system that is the repository of the data, needs to create and store UUIDs as the data are recorded at the student, and potentially school, level. There are numerous libraries that generate UUIDs that are available for all mainstream programming languages and computing environments.

When transferring data between systems that need anonymous information, student level data would not need to include any data that specifically identifies the individual (e.g., a Social Security number). The only item that would be necessary to send would be the UUID. Accordingly, the information is unique, which is important, but it is not something that can be used to identify a specific person.

Remaining anonymous in the face of other identifiable characteristics. Often, there is a general concern that someone with specific knowledge of students who are attending a given institution (e.g., at a given school, there may be only a certain number of students with a given racial characteristic), or someone who is in possession of other identifiable characteristics of individuals that are part of two dissimilar datasets, might be able to use that information to identify individuals in the dataset. Therefore, it sometimes is necessary to further protect the anonymity of the data in the source database. Generally, there are several ways to accomplish that goal, including data perturbation and policy steps.

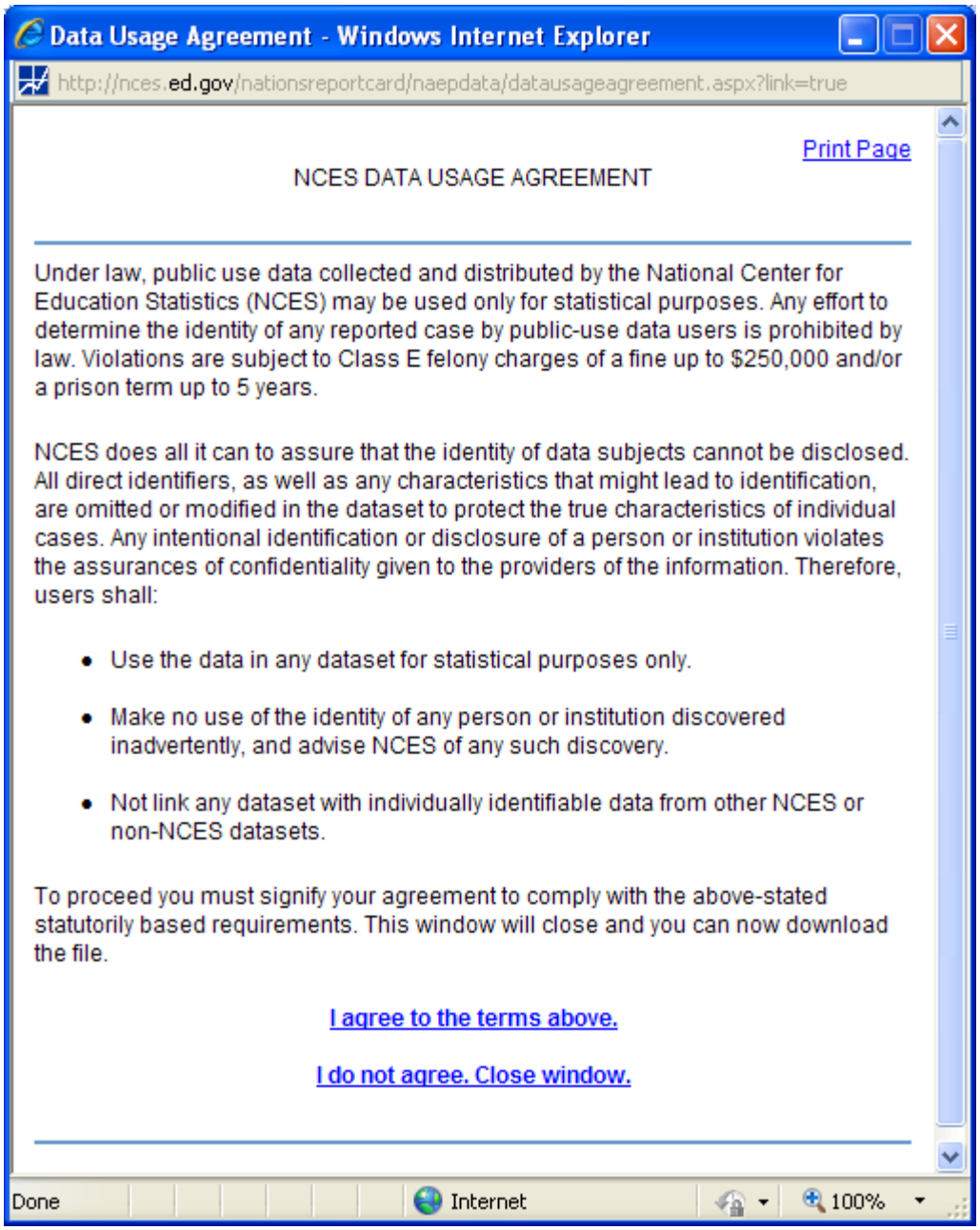
Data perturbation. Statistical methods can be applied to the source data to "perturb" it so that nobody truly knows what record is associated with what student. This is routinely done in assessments — for example, in the National Assessment of Educational Progress (NAEP). An excellent article on the methods employed by NAEP can be found here: **www.ets.org/Media/Research/pdf/RR-07-24.pdf**.

Data perturbation adds noise to the data by "swapping" characteristics and scores around between the records, such that when aggregations are performed, the results are relatively unaffected. However, a "data snooper" has no guarantee that an individual record has not been perturbed.

Policy steps. While not necessarily a technical solution to the problem, there are policy steps that can be taken to further reduce inadvertent disclosure of a person's identity. For example, when reporting group scores, all users of the data stream could be required to follow certain reporting rules. A common rule is that for a given cell, no data are reported if the number of data records in that cell is less than a certain number. Another rule is that percentages are only displayed to the nearest integer, and that the cell counts are only displayed to the nearest 10 or nearest 100. All of these rules could be set based on the size of the dataset.

An additional policy step that can be taken is asking users of the dataset to specifically sign or agree to a statement regarding appropriate use of the data. For example, every time users of the NAEP Data Explorer use the application, they must agree to the National Center for Education Statistics (NCES) Data Usage Agreement (please see **Figure 1** on the following page).

Figure 1: NCES Data Usage Agreement. Users of the NAEP Data Explorer must confirm that they will abide by these guidelines, every time they use the application.



Data Usage Agreement - Windows Internet Explorer

http://nces.ed.gov/nationsreportcard/naepdata/datausageagreement.aspx?link=true

Print Page

NCES DATA USAGE AGREEMENT

Under law, public use data collected and distributed by the National Center for Education Statistics (NCES) may be used only for statistical purposes. Any effort to determine the identity of any reported case by public-use data users is prohibited by law. Violations are subject to Class E felony charges of a fine up to $250,000 and/or a prison term up to 5 years.

NCES does all it can to assure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted or modified in the dataset to protect the true characteristics of individual cases. Any intentional identification or disclosure of a person or institution violates the assurances of confidentiality given to the providers of the information. Therefore, users shall:

- Use the data in any dataset for statistical purposes only.
- Make no use of the identity of any person or institution discovered inadvertently, and advise NCES of any such discovery.
- Not link any dataset with individually identifiable data from other NCES or non-NCES datasets.

To proceed you must signify your agreement to comply with the above-stated statutorily based requirements. This window will close and you can now download the file.

I agree to the terms above.

I do not agree. Close window.

Done | Internet | 100%

3.2.15 Scoring and Analysis of Results

How can technology standards be used for the scoring, capture, recording, analysis or evaluation of assessment results?

Consideration and sharing of results. Assessment results can be considered to include both candidate results (item responses, test scores) and question results (classical test theory, differential item functioning, and item response theory statistics). Assessment results should be shared between systems in common formats so that they can be moved from one vendor's system to a subsequent vendor's system without undue time, cost, and effort.

IMS QTI specification. Notably, the IMS Question & Test Interoperability (QTI) specification was initially released in 2000. Version 1.2 is the most widely adopted version, and focuses on questions. Version 2.1 (public draft) is more recent and covers items, tests, and results.

ETS has made use of versions 1.2, 2.0, and 2.1 to facilitate interactions between internal systems and interactions with external vendors. QTI is valuable in that it allows captured results to be sent to a separate vendor for scoring, analysis, and/or evaluation without requiring information technology (IT) work to accept, process, and return the results.

However, the QTI specification does not cover every detail, and it also enables tremendous flexibility and extension. That flexibility and extendibility impact the ability to transfer results from one vendor to another, as each vendor may use the specification differently (though each conforms to the specification), thus negating much of the value that the shared specification brings. Consequently, IT work is required to process the alternate use of the specification for each vendor.

ETS's lifecycle of assessment results. At ETS, the lifecycle of these assessment results includes a variety of disparate processes which are best supported by using common standards/specifications for data transport. These processes include the following:

- **Capture** — from scanned paper-based assessments as well as computer-based assessments. Although paper-based item responses are much simpler than many response formats supported by computer-based assessments, the use of a single standard allows any scoring system to handle both delivery modes.
- **Rating of constructed-response results** — results may include short- and long-written responses, spoken responses captured as audio, photographs of artifacts like art, and video-like teacher classroom interactions. The rating may be done by humans and/or computer models. Moreover, response data may include a hyperlink to the actual response (e.g., image, audio, or video).
- **Results management** — merging responses from scanning, computer-delivery, and constructed-response rating systems.
- **Scoring of items** — from objective item responses as well as the aggregation of multiple ratings for the same constructed-response item.

- **Scoring of tests** — aggregation of item scores.
- **Statistical analysis of item and test results** — classical item analyses use raw and scored item responses as well as test scores; item response theory analyses use scored item responses.
- **Evaluation of assessment results** — for example, screening items for exclusion from scoring, excluding items from reuse in future assessments, and evaluating the statistical properties of the items in a new assessment for conformance to test specifications. QTI 2.0 and 2.1 contains secondary metadata elements relating to item and distractor statistics, allowing both content and statistics to be transported together.
- **Reporting** — both at the candidate level and at aggregate levels like classroom, school, and local education agency.

Challenges with QTI. Within ETS, some of these processes currently use QTI, and some do not. One of the downsides of QTI and many other specifications is that they are nontrivial to implement. This characteristic makes them difficult for small organizations to implement, and makes even large organizations hesitant to use them in every process. Organizations that do not need to accept QTI data would likely not implement them on their own, as it would impact their time-to-market. Similarly, QTI (partially because it is XML-based) is a verbose specification and can negatively impact processing time.

Image content encoding. For encoding audio, video, and static image content, using industry standard file formats allows disparate vendors to read, understand, and display that content using off-the-shelf tools. For the analysis of audio and video content, software may be developed to identify and tag events within the data stream that can be used either to facilitate human scoring (such as chapters, book marks, event marks), or to do the scoring.

3.2.15.1 Results Aggregation and Reporting

How can technology standards enable assessment results to be aggregated into statistical or other groupings? How can technology standards provide capabilities for results (aggregated or raw) to be reported across multiple technology systems? For example, if a learner takes an assessment in one system, but the results are to be displayed in another, how do technology standards address transferring results across those systems? How do technology standards address aggregation of results for a number of learners who are assessed in one system and whose results are displayed in yet another technology system? Can anonymization controls be included with aggregation and reporting solutions to ensure individual data privacy and protection?

Common manner, common representation. Technology standards exist that allow for systems to interoperate at the data level in a common manner using standards based on the general concepts of web services, using agreed upon messages to communicate from data provider to data consumer. Underlying this entire solution is a common representation of data in XML format. Please refer to **Section 3.2.22, Interfaces and Services**, for our detailed discussion of web services. Using this standard technology, one system could be responsible for collecting the information, another system could be responsible for displaying individual results, and yet another could could be responsible for the aggregation of results.

Results reporting. Web services allow for results, aggregated or raw, to be reported across multiple technology systems. In order for a web service to be effective, the messages that are passed between organizations must be well defined. Version 2.1 (public draft) of the QTI specification has defined some of the structure that could be used to transmit raw/ individual results.

Data aggregation system. While no formal XML structure exists for requesting/responding to aggregate data queries, the U.S. Department of Educaton, through the NAEP contract, has in place a data aggregation system that drives the aggregation of results for the NAEP, which then was immediately reused by the Program for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), and Trends in International Mathematics and Science Study (TIMSS). Generally known within the U.S. Department of Education as "the Data Explorers," the messages they pass back and forth take into consideratrion everything an aggregation service would need to consider, including the following:

- trend results
- multidimensional tables (e.g., a way to request/respond to a table that contains results for race by gender for each jurisdiction that participated across time)
- differing statistical calculations (means, percentages, standard deviations, standard errors, proficiency levels)
- reporting rules

The Data Explorers for NAEP and the international assessments — seven different aggregation services under the U.S. Department of Education's control, in all — can be found at:

- http://nces.ed.gov/nationsreportcard/naepdata/ (for NAEP assessments)
- http://nces.ed.gov/surveys/international/ide/ (for the international assessments)

Security. System security is an important piece of a service-oriented architecture (SOA) based system. Decisions would need to be made regarding whom has access to what service, how callers identify themselves to the service, and how the services should make themselves available to their clients. For example, there might be one service made available that is designed to interact with clients, which requires appropriate access to the underlying data. This would be true for a score reporting service where a student's name is displayed. In another case, where student data are being transmitted for research purposes, another service would need to be defined where individual data are to be made anonymous using Universally Unique IDs as described, in the response to **Section 3.2.14, Anonymization**. Further, for a typical result aggregation service, reporting rules would have to be established and the service would need to be configured so that aggregate data are not emitted that violate one or more of those reporting rules.

Another important consideration is how close each service is to the source data. If there is a need to transmit results for millions of students, then having two services located on the same network, versus being separated by Internet connectivity, would have dramatic consequences on performance and scalability. Network connectivity between two services on the same network or machine can be

optimized, but the connectivity between two widely separated machines that depend on the public Internet for communication cannot.

3.2.16 Sequencing

How do technology standards enable assessment items stored within an assessment instrument to be sequenced for appropriate administration, when the assessment consists of more than a single linear sequence of items? For example, how do technology standards address computer-adaptive assessments? How are the logic rules that define such sequencing embedded within a technology standard?

The IMS QTI specification provides for the simple sequencing of items within a test, by arranging items in sections and subsections. Through the use of selections and branching rules, more complex sequencing can be achieved for linear delivery. Through the use of preconditions and branching, some degree of adaptive testing also can be specified. However, ETS has implemented various linear and adaptive sequencing algorithms that do not fit well within the specification. Further, ETS recently has developed a multistage testing (MST) structure that can accommodate a wide variety of both linear and adaptive sequencing techniques and algorithms. We believe that the MST concept could be useful in extending the specification to handle more complex, real world scenarios.

3.2.17 Computer-Driven Scoring

How do technology standards permit, enable, or limit the ability to integrate computer-driven scoring systems, in particular those using "artificial intelligence," Bayesian analysis, or other techniques beyond traditional bubblefill scoring?

Automated scoring systems at ETS are accessed by means of a J2EE standards based web services interface. The web services interface accepts an XML based input request, via HTTP POST, and an XML document is returned as the result feedback. The XML used to communicate with the interface is based on the IMS QTI assessment technology standard.

ETS's automated scoring capabilities. A number of automated scoring capabilities are available for use at ETS. Significantly, these capabilities represent the state of the art in the automated scoring of constructed-response test items. Moreover, these capabilities have resulted from more than a decade of fundamental research at ETS into natural language processing, speech technology, and related issues in educational measurement. We have published much of this work in scholarly journals and other publicly accessible repositories of research literature, and you can access this research at **www.ets.org/research/capabilities/automated_scoring**.

Below, we have described some of ETS's automated scoring capabilities.

***e-rater*®.** The *e-rater* scoring engine is an automated scoring system that we designed to evaluate the quality of academic essays. The system scores essays based on a variety of aspects of writing, using techniques from a subfield of artificial intelligence called natural language processing. Specifically, it includes scoring features covering aspects of grammar, usage, mechanics, style, organization and development, lexical complexity, and content relevance. These features are used to compute essay scores and also to provide performance feedback to students in learning environments.

Some of these features are based on machine learning models designed to predict specific types of writing errors, while others rely on sophisticated sets of rules devised by dedicated linguists. We subject each of the features that we consider for inclusion in *e-rater* to both rigorous statistical evaluation and careful consideration of its relevance to the writing construct. More than 10 years of system evaluations have indicated that *e-rater* score predictions are comparable to human reader scores. Accordingly, we use this technology in high-stakes assessments such as the *Graduate Record Examinations*® (GRE®) and *Test of English as a Foreign Language*™ (TOEFL®) programs, and in practice and learning environments such as the *Criterion*® Online Writing Evaluation Service product.

***c-rater*™.** The *c-rater* scoring engine automatically scores short free-text English-language responses for the content of the response in a brief (paragraph or less) textual response. Particularly, the *c-rater* engine measures respondents' understanding of specific content by using natural language processing techniques to determine whether responses contain **main points or concepts** that the prompt was designed to elicit. In so doing, *c-rater* is designed to be highly flexible regarding forms of expression or grammatical errors in conveying key concepts, and it gives credit based on the representation of information in the text.

c-rater works by analyzing responses with spelling correction algorithms, a morphological analyzer, and a robust linguistic parser, in order to derive a basic structural representation. Then, a paraphrase engine assesses whether these structures carry the same meaning as the target concepts specified in the rubric for the item. At present, this paraphrase engine is implemented by a maximum-entropy model trained on features related to semantic relationships between entities in each linguistic structure. Ultimately, *c-rater* is capable of providing summary scores for prompts as well as performance feedback based on which concepts were identified in the response.

***m-rater*SM.** The *m-rater* scoring engine is an automated scoring engine that scores constructed-response mathematics items for which the response is a mathematical expression, graph, or geometric figure. By using numerical evaluation, *m-rater* determines whether the examinee's response is mathematically equivalent to the correct response, and therefore accommodates all possible ways that a mathematical relationship may be expressed. *m-rater* also can give partial credit based on multiple criteria, such as being both mathematically equivalent and having an optimal form of the equation. Moreover, it is capable of scoring multipart responses with conditional dependencies in scoring, and it can provide specific performance feedback to examinees along with summary item scores.

***SpeechRater*SM.** The *SpeechRater* scoring engine is an automated scoring system for spoken-response items that elicit spontaneous speech from English language learners. We designed *SpeechRater* for the type of spontaneous spoken responses provided on the TOEFL test. It predicts the scores assigned by expert human graders, and we currently use this scoring engine for TOEFL® Practice Online, a low-stakes practice test.

SpeechRater uses features derived from speech recognition — related to the pace, pronunciation, rhythm, and fluency of recorded speech — to calculate its score estimates. In addition to finely tuned speech recognition algorithms that confirm that the computer transcribes the spoken response as accurately as possible, *SpeechRater* uses statistical classifiers as the basis for some of the scoring features it uses. Then, it combines these scoring features by means of multiple regression. Released in 2006, *SpeechRater* reflects patented research into the challenging problem of automatically deriving meaningful proficiency measures from the spontaneous speech of non-native English speakers, and represents the cutting edge of automated scoring technologies.

3.2.18 Formative, Interim, and Summative Assessments

What technology and technology standards exist that support formative, interim, and summative assessments? What technology standards support nontraditional assessment methods, such as evidence, competency, and observation based models?

QTI specification process. Of the standards produced by ADL, IMS, and SIF, the QTI specification possesses the most robust support for item and assessment level content. (The SIF standard has a much broader application to disparate school and student data, and contains standards for assessment results, not specifically item and assessment content). Regardless of type of assessment, the QTI specification does possess ample tags for many item and test level content components. At the item level, QTI provides direct support for many basic response formats (multiple-choice, constructed-response, hotspot, slider bar, etc.).

Item types and assessment structures. It is possible to use QTI for some complex item types, such as simulations, in which the simulation might be presented as an object with well defined inputs and outputs that can be handled by QTI. However, it is likely that some innovative item types (e.g., involving complex sequencing) cannot be handled by QTI and that amendments will need to be considered and perhaps developed. The IMS Learning Tools Interoperability (LTI) standard may prove valuable in supporting innovative item types that cannot be addressed through QTI. Determining the appropriate scope of application for QTI versus LTI versus other approaches for highly innovative item types is an important area for investigation. At the assessment level, often the structure of formative, interim, and summative assessment content is similar in nature in both online and paper delivery: test level directions, instructions to students, item content, sections, stopping points, and the like. Thus, with respect to overall assessment structure, the QTI specification possesses strong support, and the distinction between types of assessments is not so much significant for interoperability as it is for the purpose and delivery of the assessment.

Standards enhancements. Although the QTI specification contains a solid base for item and assessment content, the standard still needs enhancements for collateral material, metadata, and related instructional content that may pertain specifically to a type of assessment. For instance, while some instructional tags, diagnostic tags, and review and recommendation tags can be included, the QTI specification should be enhanced to support the advanced diagnostics in formative assessments. Tags to communicate advanced feedback to students or teachers in the context of a formative assessment, or links to diagnostic material to support real time learning, do not have detailed support in the QTI specification today. In order to increase interoperability and adoption across types of assessments, standards bodies and stakeholders will need to amend the standards. As these amendments are made and adopted, the opportunity for further interoperability and application integration increases.

Evidence-centered design (ECD). Regardless of the level of innovation (regular or innovative) or the assessment purposes (e.g., formative, interim, summative), we believe that evidence-centered design (ECD) (Mislevy, Steinberg, & Almond, 2003) provides a useful framework for articulating the argument for the validity of the inferences and actions resulting from assessment use. ECD-based approaches offer structured ways of representing key portions of the validity argument for assessment and may be helpful in integrating diverse sources of evidence. Templates for structuring evidence-centered design processes are being used in a variety of contexts (Mislevy & Haertel, 2006; Hansen & Mislevy, 2006) and may be candidates for extending interoperability standards.

In considering how interoperability standards might be developed to address assessments based on observations of complex performances, one might look to next-generation teacher assessments in which samples of teachers' teaching performance are recorded in video and evaluated by trained raters, and how the resulting data is accumulated to provide scores about overall performance as well as identification of areas for improvement.

References

Hansen, E. G., & Mislevy, R. J. (2006). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. Howell (Eds.), *Online assessment and measurement: Foundations and challenges* (pp. 214-262). Hershey, PA: Idea Group Publishing, Inc.

Mislevy, R.J., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*, 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

3.2.19 Learning and Training

What applications or technology standards exist that can apply assessment results to support learning and training? Are there technology standards or applications that support more than one of the following: Early learning, elementary/secondary education, postsecondary education, job training, corporate training, and military training?

In order to implement integrated learning and assessment systems (i.e., assessment systems that include formative, summative, and professional development components), assessments need to be able to share information with other learning application tools (e.g., learning management systems and other computer-based educational tools). Several standards will play an important role in this integration, including two IMS standards: Learning Tools Interoperability (LTI) and Question & Test Interoperability (QTI). These standards should be enhanced to allow for an efficient integration of assessment information with other relevant educational resources. These resources include mechanisms for representing the main design components of assessments; student information (e.g., background information and assessment results); assessment information for teachers, parents, and administrators; and other educational materials (curriculum information and relevant instructional materials). Methodologies such as evidence-centered design (ECD) (Mislevy, Steinberg, & Almond, 2003) provide the structure needed to define the information and evidence needed to support assessment and learning goals (Zapata & Bauer, in press; Hansen & Zapata-Rivera, 2010).

References

Hansen, E. G., & Zapata-Rivera, D. (2010). *Evidence-centered design for an accessible game-based assessment-for-learning system for middle school mathematics*. Presentation at the annual meeting of the National Council on Measurement in Education (NCME) on May 3, 2010, in Denver, Colorado.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Zapata-Rivera, D., & Bauer, M. (in press). Exploring the role of games in educational assessment. In J. Clarke-Midura, M. Mayrath, and D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*.

3.2.20 Repositories

What technology standards-based assessment instruments, questions, or item banks (or repositories and learning management systems) are used to manage and deliver assessments?

Many of the item banks in the formative and interim assessment marketplace, in addition to the *ETS Formative Assessment Item Bank*, are stored in XML format for delivery in online and offline situations. QTI allows items to be passed between assessment platforms because it is a common standard with wide adoption. For those platforms that are not currently QTI compliant, its XML basis allows for relatively easy adoption by companies that are XML based but have not used QTI before. Of the customers that consume the *ETS Formative Assessment Item Bank*, approximately 10 were not able to consume QTI two years ago, but now are able to consume it. The common QTI specification feeds assessment items into a variety of platforms. It has allowed the *ETS Formative Assessment Item Bank* to be loaded into platforms that support student "clicker" response systems, like TurningPoint by Turning Technologies, and pencil-and-paper as well as online, like SchoolNet's formative assessment suite.

3.2.21 Content Lifecycle

How can technology standards be employed to support an assessment content lifecycle (creation, storage, edit, deletion, versioning, etc.)?

Item banking applications. Typically, item and assessment lifecycles are developed, tracked, stored, edited, versioned, and managed in item banks. Please refer to **Section 3.2.26, Recommendation, Rating, and Review**, for more details about how this is accomplished at ETS. Depending upon the complexity of the item banking application, it can extend to and be integrated with other assessment related applications, such as automatic test assembly systems, paper and online publishing delivery systems, assessment registration systems, scoring and reporting systems, and statistical analysis systems. Moreover, technology standards could be employed as an internal data transfer mechanism within and across any of these systems.

Data transfer. If technology standards were used in this manner, then data transfer could be simplified because published, agreed upon schemas would be relied upon across all application areas. Although each application area might employ aspects of the schemas differently, depending upon data transfer needs, a set of published schemas could provide a strong baseline for application data transfer, and reduce the risk of "reinventing the wheel" that arises with application development.

Data sharing. The use of standards in this manner, then, opens up the possibility for sharing data with external vendor and stakeholder applications. At ETS, we have converted item banks to XML and published a QTI schema for import and export in support of state assessments. We also use the QTI schema in the transfer of scoring data between our internal scoring applications. We are finding that the K–12 assessment industry is already moving toward the use of QTI to support the assessment lifecycle. The bulk of the data sharing using technology standards has been with respect to item and assessment content, with less commercialized application applied to other areas of the assessment lifecycle. ETS

sees a great opportunity in the industry for the expanded use of technology standards in these other areas of assessment related data.

For example, ETS provides assessment development services to certain state contracts, and other vendors supply test delivery, scoring and reporting, and statistical analysis. In the course of providing these services, scoring data, statistical analysis results, and assessment content changes must be provided by the vendors back to ETS in order for ETS to maintain the complete item bank of record. The import of advanced statistics from vendor applications is currently not accomplished using the QTI specification because it does not support them (i.e., the standard currently possesses basic elements for statistical analysis, but areas of expansion are needed), and the specific schemas that are used between ETS and other vendors have not been fully articulated in these areas. This is just one area of opportunity and advancement. Notably, there are many other areas in the assessment development lifecycle that this type of opportunity exists. ETS perceives that the initiatives associated with the RTTA program will provide a key impetus to this more advanced use of technology standards within and across the assessment development lifecycle.

3.2.22 Interfaces and Services

What interoperability specifications for application program interfaces (APIs) or Web services interfaces to assessment management, delivery and tracking systems have been developed? How are they organized? What are the best practices related to their design and usage? How broadly have they been adopted, and what are the lessons learned from those who have designed or implemented them?

Web services. The majority of the ETS internal and external interfaces are developed using web services because of their maturity, interoperability, and security capabilities, as well as their broad industry adoption. Many of our partners and customers globally and within the U.S. prefer to interact with our systems using web services. ETS started using web services about six years ago, and our information technology developers are very well experienced now, following the best practices based on our lessons learned. Following, we have provided some of the recommended best practices:

- **Interoperability.** Web service application programming interfaces (APIs) should be implemented with encoding standards in compliance to Web Service Interoperability Specification 1.0 or above. WS-I Basic Profile confirms the interoperability of web services across platforms and languages. For example: A web service developed in Microsoft technology will be able to be consumed by a Java client API. Conforming to WS-I specifications allowed ETS to integrate with several our local and global vendors seamlessly, without worrying about their system implementation platforms.
- **Clarity of the service definitions and samples.** Web service APIs should be designed to be as explicit as possible to minimize misinterpretation. The data flow, protocol, security, validations, interactions, and sequence of messages should be clearly described and agreed upon between the systems. System integrations and implementation will be smoother when all the data flow scenarios are identified with necessary test scenarios between the systems, and sample

messages are exchanged for each of these scenarios. Because ETS deals with several global and U.S. partners across time zones, providing clear specifications and samples helped very much in system implementation. In addition, several of our clients appreciated when ETS took the extra step to provide sample client APIs to showcase the consumption of our web services.

» **Challenges.** Tracking and troubleshooting are challenges when two systems are interacting over a wide area network (WAN), especially across organizations. Because of lack of access to logs files across organizations, it is especially challenging to deal with distributed transaction and the integrity of data. At ETS, we confirm that each request has a unique request identifier across our web service messages. Because there are several steps such as authentication, authorization, marshalling/unmarshalling, and validation in the web service stack before the final processing of a message, it is always recommended to log every message even before the start of processing. In several occasions, these request identifiers have helped ETS to know precisely which message failed and then reprocess/reconcile them with our partners. Beyond the acknowledgement of the receipt and processing of messages, ETS found it very useful to build transaction reconciliation reports as well, to catch any mismatch in transaction status when there is a server side time out.

» **Agreements.** Service level agreements — such as reliability, availability, number of concurrent requests, payload size, performance, time to respond, and standard operating procedures — should be defined upfront and implemented carefully. Design should consider the business impact when implementing real time interactions amongst systems. For example: The service availability of a computer-based test delivery system is crucial when several thousands of test takers are taking exams. In a situation like this, it might be better to follow a design pattern that stocks the necessary data needed by the delivery system upfront, instead of "just in time real time" interactions to the assessment item bank or candidate management systems. This design change helps avoid dependencies on the availability of these other backend systems during the crucial time of delivering the test, thereby providing a smooth test taking experience.

» **Maintainability.** Web service APIS should be designed for stability and maintainability, and abstracted as much as possible, so that systems are not tightly coupled. This could be achieved by designing the web service operations as coarse-grained as possible without exposing atomic operations. This model allows for the addition of new operations without requiring any Web Service Definition Language (WSDL) change, which minimizes impacts to the existing consumers. APIs should be implemented with versioning; this approach minimizes breakage of existing consumer implementations when a new feature is implemented. In addition, public and private data across systems should be clearly distinguished. Avoiding exposure of private data in the web services minimizes impacts when there are internal changes in the systems.

» **Security.** Lastly, web service interactions should be secured for client authentication, authorization, data encryption, and integrity of the data**.** Security could be implemented at the transport level using HTTPS with Basic Authentication scheme, or Message Level security using OASIS Web Services Security Standards. The choice largely depends on the sensitivity of data being exchanged, implementation capability of the systems, and the network security

level/location of these service interactions. ETS has used a combination of both for some sensitive service implementations.

References

Oasis. (2006). Web services security: SOAP message security 1.1. Retrieved January 4, 2011, from Oasis's website: http://docs.oasis-open.org/wss/v1.1/wss-v1.1-spec-errata-os-SOAPMessageSecurity.pdf

Web Services Interoperability Organization. (2004). *Basic profile version 1.0*. Retrieved January 4, 2011, from Web Services Interoperability Organization's website: www.ws-i.org/profiles/BasicProfile-1.0-2004-04-16.html

3.2.23 Internal Transparency and Ease of Use

Are there technology standards and communication protocol implementations that are "human readable?" What are the benefits and risks of "human readable" technology standards? Some technology standards are not comprehensible without tools to unpack, decode, or otherwise interpret the implementation data resulting from use of the technology standard. Other technology standards, such as HTML, RTF and XML, are largely readable by a reasonably sophisticated technical user. Restful-designed Web services are often specifically intended to be readable by, and even intuitive to, such users as well. We ask commenter's to consider the extent to which various technology standards possess native "human readability" and comprehensibility.

QTI 2.0 is the preferred format for delivery of the *ETS Formative Assessment Item Bank*. It is based on XML and is human readable to 23 companies and three local education agencies (LEAs). The fact that QTI/XML are human readable makes it far easier to transfer items between platforms, as both sides can easily see just what is being delivered in an item. This capability allows for quicker troubleshooting and quality checks when needed. There have been several cases where ETS and customer personnel have read over the QTI together to resolve issues during the first integration phase. QTI with concrete XML schema definition provides a clear contract between the consumer and provider of the assessment data, thus avoiding much of the data integration issues.

The ability for humans to read the standard does not impact security in a material way. Even if the QTI data was not human readable, any open source format would have readers available for those that wished to examine items. In any sort of common standard, readability will be a step to facilitate adoption by a wide variety of vendors. Even though the data at the endpoints are human readable, ETS encrypts the data using "Advanced Encryption Standard (AES) — 128 during data transfers recommend by FIPS-197 using secure protocols such as HTTPS, SFTP, and FTPS. Encrypting data during transmission confirms that the data are not compromised in the transmission. Implementation of adequate security and access control methods at the endpoint applications or access to these assessment data is vital.

Reference

National Institute of Standards and Technology (NIST). (2001). *Specification for the advanced encryption standard (AES)* (Federal Information Processing Standards Publication 197). Gaithersburg, MD: Author. Available online here: http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

3.2.24 Discovery and Search

How is the discovery of items or instruments (or other elements) handled within a technology standard or technology? For example, are there search APIs that are provided to permit a search? How are metadata exposed for discovery by search engines or others?

Because technology standards are arranged for ease of transport, they are typically readily available for search and discover methods, whether proprietary, commercial, or universal. The standards are typically schematized, hierarchical, and document based, such that the structure and content are amenable for recursive, linear, and hierarchical search and retrieval.

Items and item metadata are typically discovered and searched while maintained within a centralized item bank, inventory, or database. Depending upon the level of security required, or the level of access among learning communities, search capabilities can be shared across applications or user communities. For instance, item and item metadata can be stored internally in exact or slightly modified technology standard formats, and the items can be searched based upon the standards themselves. If items are stored internal to an application in a format other than the technology standard, then search methods still can be applied and shared if required. Once translation of the internal format to the learning standard is made, search methods could be applied to the standard itself.

In the QTI compliant *ETS Formative Assessment Item Bank*, searching is accomplished entirely through the use of metadata. Attributes such as unique item identification, grade level, standard alignment, cognitive complexity, and difficulty all are provided in the manifest file for each group of items. The specific pieces of metadata are in compliance with the QTI specification, but were selected for this specific item bank. These metadata points can be searched directly within the QTI using the proper tools, or made available to end users once the items have been loaded into a platform for viewing.

Assuming more companies are working together in a collaborative environment, it would be useful to have an industry-wide Globally Unique Identifier (GUID) system for identifying most if not all of the above metadata items, including unique item identification. In addition, a set of required, searchable metadata should be defined.

3.2.25 Metadata

What kinds of metadata about assessments (i.e., information describing assessments) are permitted to be stored within technology standards or technologies? How do technology standards accommodate structured data (such as new State curriculum standards) that were not anticipated when the technology standard was designed? How are metadata describing unstructured (such as free-text input) and semistructured data incorporated within assessment technology standards?

As noted in **Section 3.2.26, Recommendation, Rating, and Review**, typical commercial item banks carry a wide array of metadata about items and tests. At ETS, an enterprise wide and distributed item bank carries comprehensive data at both the item and test levels, and the item bank supports both the export and import of this data in QTI format. To date, the commercial use of this import and export feature has primarily been item and assessment content, as opposed to a large set of item and test metadata. For example, the assessment level QTI supports TestPart.xml and assessment.xml specifications that can include attributes such as test program, test name, test part name, form name, sub form name, delivery mode, item sequence, test and item directions, and item usage, to name some of the typical elements. However, any of the metadata that accompany the items and assessments can be retrieved for QTI import and export if a business need arises. As long as the learning communities clearly communicate the required data elements and share the use of the standard, these data points can be accommodated in the transfer.

When using QTI to transmit the *ETS Formative Assessment Item Bank* to other companies, numerous data points about the items that will make up the assessments are transmitted. These data points for assessment items include historical *p*-value, historical point biserial, cognitive complexity, grade level, learning standard assessed, subject, and grade level. For passages/stimuli that will appear on a test, metadata sent include author, readability, gender, ethnicity, text type, text subtype, grade, and word count. The *ETS Formative Assessment Item Bank* has been checked for alignment to all 50 states, Washington D.C., and the Common Core State Standards. ETS currently places up to 53 pieces of alignment information in the manifest file for each item, which allows use of the items nationwide.

3.2.26 Recommendation, Rating, and Review

Do technology standards or technologies permit rating, review, or recommendations to be incorporated within an item, instrument, or other element? If so, in what ways? How are conflicting ratings handled? Do technology standards or technologies permit "reviews of reviews" (e.g., "thumbs up/down" or "Rate this review 1–5")? Is the rating or review system centralized, or are multiple analyses of the rating data permitted by distributed participants?

Standards flexibility. Technology standards are flexible enough to incorporate the ratings, reviews, and recommendations of items, instruments, and other elements through extensions as required. An extension to the standards consists of schema elements that serve as open ended areas for customization. Rating data is modeled, stored, and transferred within these approved extensions.

For instance, at ETS we use a proprietary item bank that produces QTI 2.1 compliant data as an export format, and we currently are using this feature to support K–12 state assessments. Within the item bank, item reviews, ratings, comments, and recommendations are tracked, versioned, and stored. These data travel with each item and elements of the assessment, and they also are associated with a custom workflow that supports the item development process. During the item development stages, items move through revision; recommendations and rewrites are tracked and discussed; and conflicting recommendations are resolved.

In addition to review and recommendation data, scoring information, rubrics, test form and statistical history, content area classifications, and alignment to state standards (in the case of K–12 programs) all accompany the item within the item banking system. Items also possess a series of statuses that track their use through the item and test lifecyle: initial authoring and review, pilot or pretest, operational use, resting or retirement, and so forth. Any or all of these data that are retained in the item bank, in principle, can be exported via an agreed upon use of a technology standard.

Comprehensive integration. A more comprehensive integration between learning communities could conceivably require standards based sharing of some or all of these elements that are tracked within the item bank to collaborate on item review, test assembly, scoring and reporting, postadministration reviews, and related assessment building tasks. The technology standards, such as QTI and SIF, have extensions whereby this material could be shared. However, currently in the industry, most commercialized interoperability has been associated specifically with item and assessment content as opposed to the wider array of rating, review, recommendation, and the host of other metadata mentioned above that accompany the item and assessment life cycle. Beyond transporting content, the technology standards have been extended for specific uses between learning communities, but the extensions have been targeted to smaller communities and uses. With the RTTA program and state consortia adopting the Common Core State Standards, there is an unprecedented need to elaborate and expand technology standards to more fully support this wider array of material, and that reduces the ad hoc extensions to technology standards. Stakeholders in this arena should apply the standards to more fully support this wider array of data by working closely with the standards boards and with one another. With standards boards and stakeholders mutually enhancing the technology standards for

commercial use, a broader based platform of interoperability that supports the full assessment lifecycle can be realized. Enhanced sets of standards that support the full assessment lifecycle also provide the opportunity to further learning technology application integration. For instance, broad sharing of recommendation, rating, and review data through the standards could support multiapplication, multisite, multivendor item review, test assembly, scoring and reporting, and postadministration reviews.

ETS software used for the *ETS Formative Assessment Item Bank* allows for electronic review of the content of items, their alignment to state standards, all associated metadata, and all associated resources. When using this solution, reviewers are able to electronically comment on all of the above data and review previous comments. Access level controls are used to limit reviewers to making comments, which are retrieved by and acted on by users authorized to make changes. All changes are linked to a user identification with a date-time stamp. Additionally, a comprehensive tracking log of all changes is maintained for on-screen review at any time. Changes are accomplished through versioning, which allows a restoration of any previous version if needed.

3.2.27 Content and Media Diversity

What types of diverse content types and forms of assessment content exist that extend beyond traditional paper-based assessments translated to an electronic delivery medium? We are interested in learning more about electronic delivery and interaction media, such as performance-based assessments, games, virtual worlds, mobile devices, and simulations.

Rich, engaging experience. The computer allows for extending the types of questions and stimuli that can be presented to respondents far beyond what is possible in a traditional paper and pencil assessment. The ability to deliver interactive simulations, video, and other complex electronic media supports a rich and engaging experience for the test taker.

Two content categories. The rich content supported by computer delivery can be divided into two categories. One is content used purely as a stimulus in a test question. For instance, an audio or video clip can be presented to a test taker, and then questions about that stimulus can be asked. An example of this is the TOEFL iBT™ assessment (**www.ets.org/toefl/ibt/about/content/**), where the English language listening skills of a test taker are assessed by presenting on the computer an audio recording of a conversation, along with animated graphics to provide a context for the conversation. Then, a variety of questions are asked about that conversation.

Another use of rich content that can be used as a stimulus in a computer-based assessment is an interactive simulation. For instance, in a science or mathematics assessment, a test taker could experiment with a simulation of some natural phenomenon to investigate the relationships that the simulation demonstrates. Questions about these relationships could then be asked. Use of such simulations provides a means to investigate scientific inquiry skills as well as a means to embed learning within an assessment context.

The second category of rich content supported by computer delivery is content where performance or response data is captured and used in evaluation of a test taker's skills. A simple example of this would be a word processor environment within a writing assessment. The test taker would use the word processor to create their essay, taking advantage of common functionalities such as copying/pasting, spell checking, and formatting. Such a writing assessment was delivered as part of the NAEP Writing Online project (**http://nces.ed.gov/nationsreportcard/studies/tba/wol/assessment.asp**). In this case, writing in such an environment provides a more valid assessment, given that most writing by students today happens on the computer.

Interactive simulations. A more complex example of using rich content to capture test taker performance is interactive simulations. In addition to their use as stimulus content, data can be collected as the test taker interacts with the software, tracking their actions and capturing the work that they create. The data collected can be thought of in two ways. One is the work product created (e.g., a graph of experimental data) and the other is the process used by the test taker in their interactions. Both of these data sets, then, can be used in scoring a test taker's performance in the assessment.

NAEP study. A concrete example of such a complex simulation was part of the NAEP Problem Solving in Technology Rich Environments (TRE) study (**http://nces.ed.gov/nationsreportcard/pubs/studies/2007466.asp**). In this study, two simulated environments were created to assess test takers' ability to use new technology in the process of scientific inquiry. In one environment, a simulated web browser was provided along with a searchable collection of web pages related to a series of questions about the use of balloons in science experiments. The test takers were instructed to do research using the web browser to answer these questions. The system tracked the searches performed by the test takers, the web pages visited, and other uses of the available tools. This rich process information gave a more detailed view of a test taker's strategy and process for answering the questions.

The second environment used in the TRE study was a simulation of helium balloons. Test takers could perform experiments by altering the payload mass and the volume of helium in the balloons, and observe results such as final altitude of the balloons. Data that was collected could be displayed in tables and graphs for analysis. Based on the experimental results, test takers were asked to answer questions about what influences the behavior of the balloons, and the kinds of relationships among the various inputs and outputs in the experiments. Again, the system tracked the actions of the test takers to provide a richer picture of the test taker's performance.

ETS's operational assessments. We have used both rich interactive stimuli and collected rich evidence in operational assessments (e.g., Katz, 2007, **www.ala.org/ala/mgrps/divs/lita/ital/262007/2603sep/katz_pdf.cfm**). The use of these types of content do offer potential benefits for assessment including increased engagement, more robust reflections of competency, and connections to educational practice. There are associated risks and cost, too, that we explore.

Careful assessment design. Actual assessments can include tasks that provide rich, interactive stimuli to students (such as simulations and virtual worlds) or collect rich evidence from students (such as sequences of actions or complex work products), or both. The types of interactions and the evidence collected typically depend upon such factors as the competencies to be measured, and cost, delivery, and accessibility criteria. In our experience, to be of high value and of high validity, tasks that employ these richer types of content and collect complex responses need to be designed carefully. Methodologies such as evidence-centered design (ECD) (Mislevy, Steinberg, & Almond, 2003) and assessment engineering (Luecht, 2002) provide the rigor needed to establish that student performance depends upon skills that are intended to be measured and not upon irrelevant factors. Messick (1994) distinguished between *task-centered* simulations, in which the goal is to replicate a situation as closely as possible, and *construct-centered* simulations, in which only the portions and details that are required to gather needed evidence for an assessment are included within a simulation or tool. By making the connections among tasks, evidence, and constructs explicit, ECD provides a principled way to develop construct-centered simulations and games appropriate for specific assessment purposes. Working through the complexities of the accessibility needs of special populations in such environments may benefit from ECD approaches (Mislevy, Steinberg, & Almond, 2003) — in particular, extensions to ECD specifically focused on accessibility (Hansen & Mislevy, 2006; Hansen, Mislevy, & Steinberg, 2008; Hansen & Zapata-Rivera, 2010). (Please see our response to **Section 3.2.28, Accessibility**, for further details on accessibility.)

There are many technology implications and trends in the development of some specific types of content — specifically virtual worlds, simulations, and games — and their use in assessment.

The growth in virtual worlds for technology-based games has driven many of the advances in virtual worlds for edutainment and education. Online worlds of the past decade were launched in contained environments on servers that allowed players from all parts of the world to play a game together. In this new decade, we are seeing more virtual worlds created with multiple levels of access.

Large-scale virtual worlds. When large-scale games or worlds are launched, there is almost always some kind of mobile application for the game. Specifically, these worlds allow users or players to access the environment from their iPhone, iPad, Android, or website interface. Though the virtual world may not be entered fully by the user, the user can still access different parts of that environment no matter where they are. For example, if you wanted to check the score of a test you took that morning, you may be able to get the result on your mobile device at lunch time. Thus, integration between mobile and hardware devices continues to become more seamless.

Offering tools or virtual worlds to students where they can take tests or conduct activities provides a wide-open landscape for interaction with teachers. Building these environments to connect across both personal computer and mobile devices allows access at all times. It is critical when building a virtual education environment in the next decade that it is available on a wide variety of platforms.

Rapid introduction of new applications. The emergence of these new platforms is fostering the rapid introduction of new applications. The connected nature of platforms like Apple's iOS (iPhone and iPad),

Google's Android (Galaxy Tab), and online portals (Second Life, Facebook) has dramatically accelerated the development cycle. Because applications are smaller in scale, they can be pushed to users in weeks instead of months. Pushing content to users is no longer the end of the development cycle, however, but merely the next stage in a feedback loop that collects real time data on user behavior, difficulty levels, time-on-task, and a host of other data points to identify bottlenecks and other problems in the user's experience. Then, this information can inform revisions and upgrades to the application, which are then redelivered to users, often in a matter of days. With such rapid turnaround, applications can be quickly fine-tuned to collect the best evidence of student abilities and to be adapted to changing needs.

References

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2008). *Evidence-centered assessment design for reasoning about testing accommodations in NAEP reading and mathematics* (ETS Research Report No. RR-08-28). Princeton, NJ: ETS.

Hansen, E. G., & Zapata-Rivera, D. (2010). *Evidence-centered design for an accessible game-based assessment-for-learning system for middle school mathematics.* Presentation at the annual meeting of the National Council on Measurement in Education (NCME) on May 3, 2010, in Denver, Colorado.

Katz, I.R. (2007). Testing information literacy in digital environments: ETS's iSkills™ assessment. *Information Technology and Libraries, 26*, 4-13.

Luecht, R.M. (2002). From design to delivery: Engineering the mass production of complex performance assessments. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments . *Educational Researcher, 23*(2) pp. 13-23.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Zapata-Rivera, D. & Bauer, M. (in press). Exploring the role of games in educational assessment. In J. Clarke-Midura, M. Mayrath, and D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*.

3.2.28 Accessibility

How do technology standards ensure that the platforms are accessible to all persons with disabilities? How can technology standards ensure the availability of accommodations based on the individual needs of persons with disabilities? What factors are important to consider so that accessibility capabilities can be included within an interoperable technology standard, both for end-users, as well as operators, teachers, and other administrators? How are issues related to Universal Design for Learning (UDL) relevant to standards for accessible use? How can technology standards provide for, improve, or enhance Section 504 and 508 of the Rehabilitation Act compliance for assessment technology?

"How do technology standards ensure that the platforms are accessible to all persons with disabilities? How can technology standards ensure the availability of accommodations based on the individual needs of persons with disabilities?"

Technology standards can play a vital role in establishing that platforms are accessible to students with disabilities. Two major areas are particularly noteworthy with regard to what the student with a disability experiences during the assessment.

Accessibility features. First, technology standards can specify the accessibility features that need to be provided by assessment delivery platforms. Features to be provided are a major focus for several guidelines that have been published for improving the accessibility of computer-delivered assessments (Pearson & CAST, 2010; Thurlow, Lazarus, Albus, & Hodgson, 2010; Thompson, Thurlow, Quenemoen, & Lehr, 2002; Hansen & Mislevy, 2006). Across all of these guidelines, several important accessibility features are mentioned including the following: voicing for text and navigation, text descriptions for non-text content, color modification, enlargement/magnification of text and graphics, captions and audio descriptions for video, and navigation through a keyboard interface, as well as practice and familiarization in the use of accessibility features. Recently, considerable progress has been made in implementing such features into a number of assessment delivery systems, but some elements (e.g., refreshable braille displays) are still missing.

Differences in the accessibility features. However, there are differences across platforms in how the features are implemented (e.g., built-in features vs. compatibility with off-the-shelf assistive technologies). If Race to the Top assesment delivery systems focus on integrating off-the-shelf assistive technologies (e.g., commercial screen reader) rather than building assistive technologies into the platform, then there may be a need for improved interoperability standards to address communication between the off-the-shelf assistive technologies and the platforms. Finally, there are significant inconsistencies across testing platforms in user interfaces between the systems. These differences could occur, for example, due to different screen layouts, different availabilities of tools, and different key strokes for navigation. For most test takers, these differences are purely aesthetic and do not impact the construct being measured. However, for individuals with visual, fine motor, learning, and cognitive disabilities, these inconsistencies can significantly reduce their facility in using the interface, thereby resulting in test scores that include additional construct-

irrelevant factors (i.e., competence with the computer-based testing platform). Thus, there may be a need for greater standardardization of layouts, tools, and key stroke commands. There also is a need to establish adequate opportunity for practice and familiarization with user interfaces, including accessibility features. To address this inconsistency, there is a need to develop and adopt common user-interface standards for use across computer-based testing platforms.

An example. For example, several computer-based testing platforms have developed alternative means of navigation by key stroke instead of by mouse. While this alternative navigation is a critical accessibility feature for some individuals, the user interface varies across platforms so that arrow keys may be used to navigate to the next item in one platform, but be used to navigate across the page in another platform. This is particularly problematic for individuals who will be testing across many platforms (e.g., state accountability assessment on platform A, end-of-course assessment on platform B, graduate admissions test on platform C).

Tagging specifications for test content. A second major area where technology standards can improve accessibility is by providing specifications for the tagging of test content. One new and particularly relevant development regarding standards for computer-based testing is the Accessible Portable Item Profile (APIP) (**https://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/APIP/index.html**) project, which is funded by the U.S. Department of Education as an Enhanced Assessment Grant (EAG). The goal of this project was to develop an interoperability standard for the accessibility of assessment content. Led by the Minnesota Department of Education, the APIP project consists of eight member states and three observing states working closely with IMS Global and Nimble Assessment Systems. The project built off the current set of IMS Global Question & Test Interoperability (QTI) specifications (as well as other IMS standards such as Access for All) to develop an integrated set of tags, along with descriptions of expected behaviors, which can be applied to standardize the interoperability and accessibility of test items. The goal is to establish that no matter the test administration tool (platform), adherence to the APIP would allow test questions to look and function the same way. In support of this goal, the APIP standard provides for the tagging of accessibility requirements.

National Instructional Materials Accessibility Standard. Another relevant standard is the National Instructional Materials Accessibility Standard (NIMAS), Version 1.0. While these standards do not currently focus on the accessibility of computer-based tests, this effort may be supportive of and complementary to standards supporting accessible computer-based testing. The NIMAS standard was endorsed by the U.S. Department of Education on July 27, 2004, as a standard to guide the production and electronic distribution of flexible digital instructional materials, such as textbooks, so that they can be more easily converted to braille, text-to-speech, and other accessible formats.

This endorsement was followed by reauthorization of the Individuals with Disabilities Education Act (IDEA) in November 2004, which names NIMAS as the voluntary national standard for providing instructional materials to blind persons and other persons with print disabilities. The Version 1.0 of

the NIMAS standard, which is based on the ANSI/NISO Z39.86 (DAISY 3) standard specification, does not currently support all the interactive capabilities that would be expected for computer-based tests. However, future versions or applications of NIMAS may contain additional interactive capabilities. (Please see the NIMAS section, **http://aim.cast.org/collaborate/NIMASCtr**, of the website for the Center for Applied Special Technology).

Building on earlier, ongoing work. Many of the above guidelines build on earlier and ongoing work, such as the Web Accessibility Initiative (WAI) of the World Wide Web Consortium (W3C). These include, for example, the Web Content Accessibility Guidelines (WCAG Version 1) (Chisholm, Vanderheiden, & Jacobs, 1999), which had a major influence on the development of the Section 508 standards of the Federal government (Architectural and Transportation Barriers Compliance Board, 2000). As technology has progressed, the distinction between websites and other computer applications has blurred, which has resulted in the development of WCAG 2.0 (Caldwell, Cooper, Reid, & Vanderheiden, 2008). Preliminary indications are that WCAG 2.0 will have a significant influence on the update to the Federal Section 508 standards (please see **www.access-board.gov/sec508/update-index.htm**).

Another segment of students. Also, it is important to note that there will be some segment of students who still are unable to use the computer-based assessment, and may need additional or alternate means of accessing the assessment. In developing technology standards for assessments, it may be necessary to consider the means whereby the student may be prompted to use materials (e.g., hard copy braille, manipulables) that are not computer-delivered.

What factors are important to consider so that accessibility capabilities can be included within an interoperable technology standard, both for end-users, as well as operators, teachers, and other administrators?

A few standards exist that seem particularly relevant when considering the accessibility of the assessment systems for individuals in addition to students — such as teachers, operators, and other administrators. For example, the W3C User Agent Accessibility Guidelines (UAAG, **www.w3.org/TR/UAAG20/**) can inform the development of assessment delivery systems. A computer-based test delivery system would be a type of "user agent," where the user may be a teacher, operator, or administrator, or student.

Also, the W3C Authoring Tool Accessibility Guidelines (ATAG, **www.w3.org/TR/ATAG20/**) can help assure both (a) the accessibility of the content that is authored and (b) the accessibility of the interface for the authoring tool. While the UAAG and ATAG guidelines are not specifically focused on assessments, these guidelines are likely to be a valuable point of reference to develop technology standards for accessible assessment systems. Additionally, depending on the specific platform that is developed, it may be important to conform to computer-platform specific (e.g., Microsoft Windows, Apple) accessibility guidelines.

Special education considerations. We should note that the foregoing considerations are particularly relevant in special education, where a large number of teachers of students with hearing impairments and visual impairments are themselves hearing or visually impaired. Outside of the area of special education, the need for accommodations by teachers is likely lower. For example, in a recent year in an ETS computer-based teacher licensure assessment, 1.2 percent of registrations (generally from prospective teachers) requested an accommodation.

How are issues related to Universal Design for Learning (UDL) relevant to standards for accessible use?

Many of the standards mentioned earlier address the implication of Universal Design on Assessment (Pearson & CAST, 2010; Thurlow, Lazarus, Albus, & Hodgson, 2010; Thompson, Thurlow, Quenemoen, & Lehr, 2002; Hansen & Mislevy, 2006). Universal Design for Learning (UDL), however, is broader than assessment and does not currently take sufficient account of the need for scores that have the same meaning, reliability, and comparability across individuals and groups of examinees.

There are other standards that are relevant for accessibility of assessments but are not specific to technology standards or UDL. Among the most important are the joint *Standards for Educational and Psychological Testing* of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Currently, the joint *Standards* (AERA, APA, & NCME, 1999) are under revision and will include substantial changes to the chapter on accessibility (**http://teststandards.net/Revision.htm**). Additional standards include (a) *How ETS Works to Improve Test Accessibility* (ETS, 2010), (b) *Accessibility Principles for Reading Assessments* (Thurlow, et al., 2009; **www.narap.info/publications/reports/NARAPprinciples.pdf**), and (c) the *ETS Standards for Quality and Fairness* (ETS, 2002).

How can technology standards provide for, improve, or enhance Section 504 and 508 of the Rehabilitation Act compliance for assessment technology?

Technology standards such those mentioned above are critical for improving or enhancing compliance to Section 504 and 508 of the Rehabilitation Act for assessment technology. However, it should also be noted that while technology and technology standards play an essential role in establishing accessible assessments, human judgment also is essential for creating accessible assessment content, delivering it, and confirming that the intended benefits are obtained for students and others with disabilities.

For example, while one can profitably employ automated tools so that text descriptions of graphics are available to the test taker who is blind and needs the text description read aloud, human judgment is necessary to confirm that the text description thus provided will support rather than undermine the validity of the assessment results. Therefore, it is important to make adequate provision for human expertise to create and evaluate technology-based assessments for both accessibility and validity, when developing and applying technology standards.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Architectural and Transportation Barriers Compliance Board. (2000). *Electronic and information technology accessibility standards (Section 508)*. Retrieved October 4, 2002, from www.access-board.gov/sec508/standards.htm

Caldwell, B., Cooper, M., Reid, L. G., Vanderheiden, G. (Eds.). (2008). *Web content accessibility guidelines* (W3C recommendation). Retrieved January 6, 2008, from www.w3.org/TR/WCAG20/

Chisolm, W., Vanderheiden, G., & Jacobs, I. (Eds.). (1999). *Web content accessibility guidelines* (W3C recommendation). Retrieved May 5, 1999, from www.w3.org/TR/WCAG10/

Educational Testing Service (2002). *ETS standards for quality and fairness.* Princeton, New Jersey: Author. Available online here: www.ets.org/Media/About_ETS/pdf/standards.pdf

ETS (2010). *How ETS Works to Improve Test Accessibility*. Princeton, New Jersey: Author.

Hansen, E. G., & Mislevy, R. J. (2006). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. Howell (Eds.), *Online assessment and measurement: Foundations and challenges* (pp. 214-262). Hershey, PA: Idea Group Publishing, Inc.

Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., & Lehr, C. A., (2002). *Access to computer-based testing for students with disabilities* (Synthesis Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 6, 2004, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis45.html

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Pearson & CAST. (2010). *Universal design for computer-based testing (UD-CBT) guidelines*. Retrieved January 11, 2011, from Pearson's website: http://education.pearsonassessments.com/NR/rdonlyres/191D5EF9-49A3-4ACD-89E0-C4FE545326C0/0/TMRS_RR_UDCBTGuidelinesrevB.pdf?WT.mc_id=TMRS_Universally_Designed_Computer_Based_Testing

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis Report 78). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments*. Minneapolis, MN: National Accessible Reading Assessment Projects. Available online here: www.narap.info/publications/reports/narapprinciples.htm

3.2.29 English Learners

How do technology standards ensure that assessment platforms support the assessment, reporting of results, and other capabilities related to the assessment of English learners?

In the context of testing students in grades K–12, English learners (ELs) currently are assessed in two arenas: English language proficiency (ELP) and knowledge of academic content. At present, ELP is assessed in order to classify or reclassify students into or out of EL status, and it occurs in a one-on-one testing situation with limited use of technology (Wolf, Farnsworth, & Herman, 2008). Platforms that can accurately deliver assessments (including audio) and faithfully capture responses from students (both oral and written) could increase the validity of ELP assessments. In addition, technology can facilitate the use of adaptive testing of ELP, which could lead to faster and better estimations of a student's ELP level.

With regard to assessing content knowledge, the main obstacle to valid assessments for ELs is the use of inaccessible language (Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh, 2009). The use of technology in assessment platforms should not introduce or add construct-irrelevant variance in the form of linguistic barriers to understanding. Test items and directions should be clearly written and understandable, irrespective of whether the mode of administration is via paper-and-pencil or by other means.

The use of technology in assessment platforms does have the potential for more valid and accessible content assessments by allowing for different testing accommodations for ELs. Those accommodations that provide direct linguistic support are considered to be most useful for ELs (Rivera, Collum, Willner, & Sia, 2006). For example, the ability to provide pop-up glossaries or side-by-side translated test versions would be an effective use of technology in assessing ELs. However, please note that the state of the science regarding the effectiveness and validity of many testing accommodations is quite limited at this point. The most recently published meta-analysis of testing accommodations for ELs (Kieffer, Lesaux, Rivera, & Francis, 2009) found that only one of the seven accommodations studied, the use of English dictionaries or glossaries, had a statistically significant (although small) effect on the performance of ELs. Please also note that the number of experimental studies completed to date of testing accommodations for ELs is still quite small, as a total of only 11 studies were cited in the Kieffer, Lesaux, Rivera & Francis meta-analysis.

It is likely that many of the same technology standards and guidelines that are useful for students with disabilities (please see **Section 3.2.28, Accessibility**) also will be useful for English learners (e.g., simple-as-feasible language requirements, appropriate use of glossaries, etc.). However, care also should be taken to identify needs that might be different between the two groups.

References

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79,* 1168-1201.

Pitoniak, M., Young, J. W., Martiniello, M., King, T., Buteux, A., and Ginsburgh, M. (2009). Guidelines for the assessment of English language learners. Princeton, NJ: Educational Testing Service.

Rivera, C., Collum, E., Willner, L. S., & Sia, J. K. (2006). An analysis of state assessment policies regarding the accommodation of English language learners (pp. 1-174). In Rivera, C., & Collum, E. (Eds.), *State assessment policy and practice for English language learners: A national perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment, 13*(3), 80-107.