Assessment Technology Standards
Request for Information
United States Department of Education

TECHNICAL ORIGINAL
Due: January 17, 2011
Steve Midgley
U.S. Department of Education
400 Maryland Avenue, SW
Room 7E202
Washington, DC 20202-0001
Dear Mr. Midgley

On December 20, 2010, the US Department of Education issued the Assessment Technology Standards Request for Information (RFI). The primary purpose of this RFI is to explore existing, in-process, or planned open technology standards, specifications, and technology products that support the management, delivery, and exchange of assessment results. Information collected during this process will be used by the Department to help determine the interoperability standards for assessments and related work developed under the Race to the Top Assessment (RTTA) program.

The following pages include advice, technical information, and input from Pearson, for each question raised in the RFI, as to how the Department can select the best available technology standards for the RTTA program.

The RTTA program and state consortia adopting the Common Core State Standards (CCSS) have identified interoperability as essential in helping to maintain the feasibility and affordability of next-generation assessments. In addition, the general nature of large-scale assessment is evolving from current multiple-choice dominated tests to those involving a variety of item types to improve measurement of learning. Together, these factors will push technical requirements of assessment interoperability standards beyond current capabilities. A critical question is whether our current assessment standards can handle greater expectations of interoperability across systems (including student information and instructional systems as instruction and assessment are more tightly integrated) and the enhanced sophistication and interactivity required to measure performance against the CCSS.

To date, existing standards such as the Question and Test Interoperability standard (QTI) from IMS Global Learning Consortium, and the Schools Interoperability Framework (SIF®) from the SIF Association have met the expanding needs of assessment content representation only through significant ad-hoc extension, thus compromising widespread interoperability and compatibility.
Now an unprecedented need exists for collaboration between vendors and assessment specification governance boards to determine coherent and comprehensive assessment standards. Such standards must support the full life cycle of building, delivering, scoring, and reporting assessments and must address assessment item content, assessment assets, assessment item accommodations, assessment item metadata, test forms, scoring, reporting, and feedback. These standards must further verify that simultaneous goals of innovation and standardization are not at odds with one another, and that the process of extension does not constrict future growth.

Based upon these requirements, the following considerations for assessment interoperability standards are recommended, and are discussed in detail in Pearson’s response.

- **Minimize Technology Constraints** - Assessment interoperability standards must maximize the potential to deliver on a range of devices and should minimize constraints that come from connections with technology delivery platforms that are subject to inevitable shift in the near term or obsolescence in the long term.

- **Separate Content and Formatting** - The separation of content and formatting within an assessment standard is critical. As described in Section 3.2.8, there should be two distinct levels of an assessment interoperability standard: one that speaks to the content interoperability (content exchange) and another for assessment interoperability (the ability to call upon an assessment from another system such as a Learning Management System which includes the ability to receive results).

- **Support Wide Range of Content** - Assessment interoperability standards must support a wide range of content required for the CCSS as needed for greater interactivity, increased accessibility, more sophisticated assessments, and easier test question authoring.

- **Address Assessment Structure, Accommodations, and Additional Standards** - Assessment interoperability standards must address assessment structure and test-level accommodations, and integrate well with other standards and other instructional resources.

To support the evolution of assessment interoperability standards for next-generation assessments, we recommend stakeholders consider the following concurrent actions:

- **Engage Existing Standards Boards** - Standards evolution must occur through active involvement via standards boards, most notably IMS Global and the SIF Association, whether they occur through extensions to existing versions, definition of incremental updates, and/or creation of radically redefined versions.

- **Leverage the Cooperation and Support of Oversight Organizations** - Industry organizations such as the Council of Chief State School Officers (CCSSO) and the Association of Test Publishers (ATP) should be invited to help achieve consensus among stakeholders. As neutral and knowledgeable third parties, such organizations could help avoid potential disputes between competing interests and serve as partners in distributing information and soliciting input and feedback across a wide set of stakeholders. CCSSO and other organizations play a similar role as part of the technical workgroup for the Common Education Data Standards (CEDS) Consortium, which is facilitated by the US Department of Education, National Center for Education Statistics.
CEDS creates definitions of—and formats for—a subset of key K-12 data elements and key K-12-to-postsecondary transition data elements.

- **Allow Stakeholders to Determine the Best Interoperability Solutions** - Future requests for proposals from funding agencies, states, state consortia, or other testing bodies should not overprescribe the interoperability standard(s) to be used in the funded proposal. Instead, applicants should be given the freedom to determine, through transparent and collaborative means, the open standard or standards that will best accomplish what is requested.

If you have any questions or would like to discuss our response further, please contact me at 319-339-6407 or by email at jon.s.twing@pearson.com or my colleague Shilpi Niyogi, Executive Vice President, National Services, at 202-378-2128 or by email at shilpi.niyogi@pearson.com.

Sincerely,

[Signature]

**Jon S. Twing, PhD**
Executive Vice President and Chief Measurement Officer
Assessment & Information Group of Pearson

D: (319) 339-6407
E: jon.s.twing@pearson.com
## Table of Contents

Current Landscape ................................................................. 3

Timelines .................................................................................... 5

Process ....................................................................................... 8

Intellectual Property ................................................................. 10

Existing Intellectual Property ..................................................... 11

Customizing ............................................................................. 12

Conformance and Testing .......................................................... 15

Best Practices ........................................................................... 16

Interoperable Assessment Instruments ......................................... 19

Assessment Protection ............................................................... 24

Security and Access ................................................................. 26

Results Validity ......................................................................... 28

Results Capture ........................................................................ 29

Results Privacy ......................................................................... 30

Anonymization ......................................................................... 32

Scoring and Analysis of Results .................................................. 33
3.2.1 Current Landscape

**Requirement**

3.2.1 Current Landscape. What are the dominant or significant assessment technology standards and platforms (including technologies and approaches for assessment management, delivery, reporting, or other assessment interoperability capabilities)? What is the approximate market penetration of the major, widely adopted solutions? To what degree is there significant regional, educational sub-sector, or international diversity or commonality regarding the adoption of various technology standards and capabilities, if any?

**Response**

A variety of interoperability standards are currently used to represent student data, test content/structure, and assessment results. The most commonly used standards include the Sharable Content Object Reference Model (SCORM), the Schools Interoperability Framework (SIF), the Question and Test Interoperability (QTI) specification, and the standards produced by the Postsecondary Electronic Standards Council (PESC). These standards have different emphases, such as SIF’s and PESC’s focus on student data and assessment reporting, in contrast to SCORM’s and QTI’s focus on educational content. They also vary by the assessment sub-sector they primarily support. For instance, SIF emphasizes K–12 education while PESC covers high school and college.

Market penetration of the standards varies by country. Several countries, such as Korea, have pursued greater compliance with QTI than the United States has. The rate of standards adoption has also varied with the degree to which online technologies are leveraged within a particular education subsector. For instance, text to support formative assessments delivered through online learning management systems are far more likely to be developed in adherence to technology standards than high-stakes assessments that are delivered in both print and online forms. Compliance with interoperability standards is also more common in higher education and in licensure/certification exams than in K–12education, in accordance with the degree to which computers tend to be commonplace in those work or education environments.

While organizations, including PESC and IMS Global Learning Consortium, maintain several sets of technology standards and coordinate among those standards within their organization to avoid overlap or conflict, such coordination is not consistent across organizations. For instance, the SIF Association and PESC both maintain assessment reporting standards that apply to high school education. Similarly, work groups within these organizations may overlap such as with the ePortfolio work groups found in both IMS and PESC. When multiple standards exist for the same domain, vendors may choose differently regarding which standard to comply with, thereby diluting the potential for interoperability across products.

The range of available specifications with occasional but limited coordination between standards governance organizations has led to some degree of fragmentation. Widespread
interoperability is also constrained by limited adoption and by adoption accompanied by proprietary extensions that allow for only a portion of a system’s data or content to be used by another system.

In analyzing barriers to widespread interoperability, issues with the standards themselves need to be separated from other factors that have prevented vendors from adopting industry standards for interoperability. To date, proprietary solutions have flourished with an emphasis on differentiating vendor-specific features, products, processes, and systems, rather than on standardizing data models across vendors. With interoperability playing a larger role in the assessment business model, compliance to standards will follow. However, the quality of the results for the US education system will ultimately depend on the quality of the interoperability specifications, and their ability to evolve along with innovations in how assessments are developed and administered and advances in the underlying technologies utilized. The near-term decisions in this regard are difficult and involve a trade-off between timing and quality: interoperability standards can be reached sooner by relying on existing standards while more robust and future-proofed standards could be developed with additional time investment in these early stages of RTTA development.

There are four main challenges when considering wide-spread adoption of assessment interoperability standards to support RTTA’s needs:

1. **Assessment content and media.** A range of item types and assessment modalities (e.g., performance tasks, simulations, serious gaming, portfolios) must be supported by assessment standards, if interoperability of next-generation assessment items is to be realized.

2. **New devices and technologies.** Assessment interoperability standards should not be predisposed towards a single technology platform at a time when the type of devices used in the classroom are diversifying and the content delivery technologies supported by those devices are undergoing radical evolution.

3. **Increased integration of assessment and learning.** The future vision for a globally competitive education system within the United States includes a more meaningful integration of assessment administration and results into the classroom. For technology to effectively contribute to this vision, data must traverse various systems: student information systems, score reporting, data warehousing systems, gradebook software, online testing platforms, assessment item banks, formative assessment tools, personalized learning systems, and parent/teacher/administrator portals. They can then operate as part of a coordinated system through data exchange rather than remaining isolated and self-contained. Multiple standards cannot be fragmented, overlapping, or contradictory; they must be coordinated and comprehensive.
4. **Flexibility and ease of implementation.** To maintain the flexibility provided by system-agnostic and highly portable items, tests, student information, and assessment results, much greater scrutiny will be placed on interoperability standards as vendors strive to provide the same features that their proprietary systems currently make available to clients. If standards are easy to implement and allow for a range of features, compliance can occur at a deeper level within systems. The temptation to extend beyond the standards in non-collaborative ways in order to deliver innovative functionality decreases.

### 3.2.2 Timelines

**Requirement**

3.2.2 Timelines. Approximately how long would it take for technology standards setting and adoption processes to obtain a technology standard that meets many or all of the features or requirements described in this RFI? What are the significant factors that would affect the length of that timeline, and how can the impact of those factors be mitigated? More specifically, would the acquisition of existing intellectual property (IP), reduction or simplification of specific requirements, or other strategies reduce the time required to develop these technology standards and processes?

**Response**

There are many variables that influence the length of time it would take to develop and adopt technology standards. The two main cycles to consider in the planning timeline are: 1) the development of the standard itself and 2) the adoption and implementation of the standard by solutions providers (or vendors) and other stakeholders.

The first cycle, developing the assessment standards, can be a relatively long process. Standards development involves: 1) drafting a proposed standard or alterations to an existing standard, 2) garnering consensus among the standard’s group representatives internal to the standards body and among the participating members, 3) the public review and comment period, and 4) the final steps for moving the draft standard forward to be an approved standard. With any standard, these common processes and practices take time.

The second cycle, solution providers adopting and implementing the standards, is also affected by many variables. Considerations that could affect the schedule and would need to be evaluated include:

1. **Is this a new standard and how effective is the standard for the current solution design?** Implementation of a new standard would be a learning and discovery process by the organization implementing the standard. Depending on how clearly the standard is defined, availability of documentation and sample implementations, and availability of industry experts for consultation, this process can be accelerated or slowed. In addition, how the concepts, constructs, and data structures align with the current provider’s solutions will impact implementation timelines.
2. **Does the solution currently use the standard?** If the solution has implemented the standard in question in prior releases, then this becomes a change to the current solution (as opposed to a potential new design). Depending on the level of change to the standard, impact would be variable.

3. **What is the current backlog and customer demand cycle?** If the solution is currently in operation with an existing customer base, often it is difficult to prioritize standards implementation activities ahead of current customer demands or support activities. Generally, it is impossible to delay resolving backlogs while redesign is occurring—meaning that features not implemented in the current version may not be addressed during a redesign. Therefore implementation of the standard may be a multi-step (release) process over a period of time.

4. **How “deeply” is the standard implemented?** If the standard only governs as an import/export method for data and/or content exchange, then the impact to the system design may be greatly reduced. However, this strategy relies entirely on transformations between the standard and the proprietary implementation within the solution itself. This strategy can result in some information loss during the exchange, although it does buffer the system against frequent standard changes, allows a system to interoperate with multiple standards, and provides a strategy for supporting extensions without changes to the standard. However, if the desire is to have systems natively using the same standards without need for manual or automated conversion efforts, then system redesigns are necessary and will require time to achieve. Different levels of compliance—using a standard natively versus supporting a standard or multiple standards for exchange—may provide flexibility in how product developers respond to new standards. However, transparency around products’ levels of compliance should be maintained, since customers need to be aware of when data conversions may cause loss, error, or increased system overhead. Additionally, if the standard requires the use of particular technologies rather than just governing the data models, then longer product development timelines may be necessary if current systems do not use the mandated technologies.

5. **What resources and funding are available to implement the standards?** Implementation of a standard or ongoing maintenance to the standard is not free. Solution providers must plan for, budget, and allocate resources to make the necessary changes to their products to implement the standard. If the product is very mature and widely adopted, this may be a more difficult proposition than for a new, emerging, product line.
6. How do the customer requirements align with the current standard's capabilities? Providers must also be able to handle situations where a customer asks for a new capability, data element, or behavior that is not currently supported by the standard. The options are to 1) reject the request, 2) defer the request and wait for the standard to evolve, or 3) implement a custom “deviation” from the standard to satisfy the request. Depending on the customer’s level of priority, often option 3 is chosen, ideally with such extensions eventually feeding into the extension of the standard as described in option 2. However, this creates potential standards compliance issues as product capabilities that extend beyond what is supported by the standard begin to be used more widely by customers eager for new functionality. Alternatively, systems become more complex as features are enabled only for the clients who have opted to extend beyond the standard.

7. Is the standard extensible for addressing missing or proprietary capabilities, and what is the approval process for extensions? The option to extend current standards depends on solution providers’ ability to handle deviations to the standards. Because most standards organizations recognize this reality, standards often offer extension points for variability. While this allows for providers to continue moving forward while supporting the standard, this practice results in 1) the custom extensions becoming immediate interoperability concerns between providers and 2) the custom extensions often become a collecting ground for any variability that the standard may not currently address and fall off the “radar” for future standards enhancements.

To help understand the timelines for standards development, we can consider the release schedules for current assessment standards.

**IMS QTI Release Timeline Summary**
- Version 0.5 March 1999
- Version 1.0 May 2000
- Version 1.01 August 2000
- Version 1.1 March 2001
- Version 1.2 January 2002
- Version 1.2.1 March 2003
- Version 2.0 January 2005
- Version 2.1 June 2006 (draft)

**SIF Release Timeline Summary**
- Version 1.1 February 2003
- Version 1.5r1 October 2004
- Version 2.0 October 2006
- Version 2.0r1 June 2007
- Version 2.1 September 2007
- Version 2.2 March 2008
- Version 2.3 February 2009
- Version 2.4 June 2010

In general, it was possible for the standards organizations to make incremental changes (dot releases) fairly quickly—generally in less than a year and in some cases within months. The
challenge is making major revisions to the standards. This tends to be a longer cycle—approximately 2 years for both IMS QTI and SIF to go from version 1.x to version 2.0.

These timelines only reflect the standards development cycles and not the cycle for implementation by the solutions providers. Accurate data on that cycle is not widely (or publically) available.

It is not uncommon for solution providers to adopt new standards before they have been approved by the standards body. For example, HTML 5 has not yet been an approved standard and is still evolving; however, many browsers can support HTML 5. This can create great risk as browser display or behaviors may not be consistent across vendor solutions, which may be acceptable for Web surfing but certainly much less desirable in an assessment environment. However, such premature implementations are a risk when the durations between draft release and approval are protracted

3.2.3 Process

Requirement

3.2.3 Process. What process or processes are appropriate for the adoption, modification, or design of the most effective technology standard in a manner that would answer many or all of the questions in this RFI? We are interested in learning the extent to which the uses of one or another process would affect the timeline required to develop the technology standards.

Response

Standards development processes are defined by each of the standards governing bodies. Each will have policies, procedures, and bylaws that define the processes for drafting, reviewing, and approving standards.

Standards development processes can be similar to the development processes for any technology. For the development of a technology standard it is important that requirements (or use cases) are understood and documented, a thorough design is considered, detailed specifications (or XML structures) are generated, standards are implemented in software systems, and testing processes are developed to verify that the specifications are being met by those solutions. All must be encompassed in a solid development plan and schedule that is visible to everyone so that shared priorities can be established.

In addition to what otherwise might look like a traditional software development life cycle or the Systems Development Life Cycle (SDLC), a standards development process has many more stakeholders, checks and balances, reviews, and constraints at play than a single software solution, which can complicate the process. Stakeholders for assessment standards include the standards bodies, standards working groups, vendors (solution providers), policy makers, and end users at all levels of the educational ecosystem. Engaging all stakeholders throughout the process is imperative to maintain the greatest level of success.
As is true with traditional SDLC, starting with clearly defined and articulated requirements that establish purpose, scope, functional and non-functional characteristics, and usage behaviors is also required for standards development and is the most significant indicator of future success. This step would be particularly critical at this juncture in light of the fact that some priorities may have shifted since standards bodies began their original work. If a common understanding around requirements and evaluation criteria cannot be attained, then confusion and variability will certainly follow, ultimately reducing the impact of the standard on adoption and interoperability. It is also important that all stakeholders, and most importantly end users, are included in the development of the use cases. End user input would result in standards that are more highly likely to hit the target. Standards being defined and pushed down from a higher authority are less likely to succeed. However, external guidance from governance organizations such as CCSSO, NGA, and USED may provide useful for establishing priorities, breaking stalemates, and encouraging greater coordination across standards organizations.

The subsequent design and specification stages are also critical, but for different reasons. Because the standard must be shared and implemented by a large set of solutions providers, many of which may have not participated in the development process, having a well documented and easy to understand specification is critical. Otherwise, the specification may be open to interpretation and therefore result in variability between solutions and a less than optimum interoperability solution. An excellent way to convey meaning is to provide a robust set of examples that exercise all elements of the standard so the knowledge in the specification can be converted into understanding of how the standard is applied.

The requirements that are outlined in the questions of this RFI are extensive and are not completely covered by any assessment standards today. For the most effective development and deployment processes possible, the standards development efforts can (and do) adopt techniques common in agile software development processes (e.g., iterative and incremental development, cross-functional requirements development teams, etc.). Agile practices provide the ability to breakdown large requirement backlogs into smaller “chunks,” prioritize the backlog, and incrementally develop the standards. The ability to breakdown, prioritize, and incrementally develop and refine the standards will result in incremental successes that will generate excitement and participation by a wider audience. Agile practices are structured to handle a rapidly changing landscape as each iteration is short and new requirements can be injected and prioritized on shorter cycles. Agile practices also enable for frequent customer review and feedback increasing the likelihood that development is moving in the right direction. Initial efforts could define overall scope and points of connection between different domains (assessment results, student information, and assessment content). Priorities for different areas of interoperability could be established so that faster wins could come from establishing standards in the more critical areas first.
Some of the requirements outlined in this RFI may be “breaking” requirements that have significant impact on the current standards (see discussion on breaking requirements in Section 3.2.8 Interoperable Assessment Instruments). All must be balanced with delivering useful increments so that solutions providers are not overburdened with a continuous flow of changes that may be met with resistance. Finding the balance between meaningful forward progress and impact on solutions providers will need to be achieved.

3.2.4 Intellectual Property

**Requirement**

3.2.4 Intellectual Property. What are the potential benefits and costs to the Federal Government, States, and other end users of different IP restrictions or permissions that could be applied to technology standards and specifications? Which types of licensed or open IP (e.g., all rights reserved, MIT Open License, or Gnu Public License) should be considered as a government technology standard? How should openness relating to the IP of technology standards be defined and categorized (e.g., Open Source Initiative-compatible license, free to use but not modify, non-commercial use only, or proprietary)

**Response**

The use and implementation of common technology standards, specifications, and guidelines will allow many different software and content vendors to provide service for education stakeholders without the interoperability concerns or constraints that prominently exist when implementing a multi-vendor solution or vendor transition. Incorporating an intellectual property licensing requirement in such a technology standard will create a mechanism of legal enforceability and enhance compliance among solution providers. It is also important to understand that solution providers will deploy the open-source licensed standards in assessments through proprietary delivery and reporting systems. Therefore, the interaction between open source content and technology standards and proprietary assessment delivery systems should be established in the standards to facilitate clarity.

In order to facilitate interoperability standards that meet the needs of education stakeholders, such standards must be licensed pursuant to an open source license. More than 50 open source licenses have been approved by the Open Source Initiative, including Apache License, 2.0, Berkley Software Distribution (BSD) License, Eclipse Public License, GNU General Public License, MIT License, and W3C License.

Some open source licenses permit users to create derivative works, including proprietary derivations. In many cases, these licenses allow users to create the derivative work products without any constraints or expectations regarding redistribution. This allows developers to obtain trade secret protection for the derivative software and permits them to license the software for a fee.
Other open source licenses require derivative works to be licensed and/or redistributed under the same terms as the source. If an organization makes a modification to software under a GNU General Public License (GPL) type license and makes the software publicly available (for free or for a fee), then the derivative and source software must also be made available to the recipients under the same license terms. This type of license is often referred to as a copyleft license.

The General Public License (GPL) and Lesser General Public License (LGPL), commonly referred to as a Reciprocal License, encourage the growth of open source software, regardless if a derivate work is used and/or modified for proprietary purposes.

The purpose of establishing the content and assessment interoperability standards defined in the RFI will only be realized if such standards require the relevant assessment technology to be licensed pursuant to a copyleft open source license, such as the GNU GPL. Allowing solution developers to create derivative versions of the assessment standards which remain proprietary would prevent education stakeholders from realizing the benefits of a continued assessment platform evolution and interoperability between platform providers.

### 3.2.4.1 Existing Intellectual Property

#### Requirement

3.2.4.1 Existing Intellectual Property. What are the IP licenses and policies of existing assessment technology standards, specifications, and development and maintenance policies? Are the documents, processes, and procedures related to these IP licenses and policies publicly available, and how could the Department obtain them?

#### Response

Standards and specifications are generally developed and maintained by a standards organization consisting of members working towards a common goal. The standards organization defines the policies and procedures used to enhance and enforce the standards and/or specifications that it represents (see discussion in Section 3.2.3 Process). The standards organization establishes the applicable Intellectual Property license for the standard, and requires redistribution of the information pursuant to the terms of such licensure. Even though the standards may be freely shared amongst authorized users, the standards organization or governance system must maintain compliance with the defined policies and procedures.

For example, the IMS Global Learning Consortium requires users to register with IMS before utilizing the QTI Specification. The terms of the license allow the registered user to reproduce and distribute the QTI Specification pursuant to a perpetual, irrevocable, royalty-free, non-exclusive, non-sublicensable license. Furthermore, the registered user is granted a royalty-free license to develop, market, import, and distribute products which implement the required elements of the specifications and simultaneously provide attribution to the QTI Specification.
In the event the registered user breaches the terms of the license agreement, their license to utilize the specification may be terminated.

The SIF Association, a non-profit corporation, is responsible for developing, maintaining, and interpreting the SIF standards and specifications. Technology vendors that seek to certify an SIF product must register with the SIF Certification Authority, pay a fee, and undergo conformance testing. The solution developer’s product information is provided to the SIF certification authority pursuant to a confidentiality agreement and therefore will remain proprietary and protected by trade secret, as applicable. The terms of the SIF certification agreement require the vendor to affirm the product’s compliance with the mandatory standards and specifications. Vendors are also subject to a Trade Mark License Agreement if they wish to use the SIF trademark in connection with a Certified Product.

Our recommendation is that a structure similar to Underwriters Laboratories, Inc. (UL) be established for independent product certification against the acceptable standard\(^1\). Such an independent certification will allow vendors to develop their products with established standards without sacrificing innovation or advances in technology. Certification test procedures and acceptance criteria should be publicly available and free to use by all. However, there may be fees associated to get a product certified by an independent certification authority.

### 3.2.5 Customizing

**Requirement**

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?

**Response**

Assessment tools consist of numerous components that utilize many different technologies. Generally speaking, even when the components adhere to defined specifications and standards, the technologies used to create the tools—hardware platforms, operating

---

\(^1\) [http://imsglobal.org/question/qtiv1p2p1/qtiv1p2p1speclicense.html](http://imsglobal.org/question/qtiv1p2p1/qtiv1p2p1speclicense.html).


\(^1\) The UL product certification program is accredited by the U.S. Occupational Safety & Health Administration (OSHA), the American National Standards Institute (ANSI), and the Standards Council of Canada (SCC).
systems, programming languages, database models, etc.—can still be chosen by the producer, as long as the standard does not dictate a specific platform or technology. Thus, at one level customization is still quite possible within the constraints of most standards.

There are exceptions to this, however where the technology is limited by the platform(s) and/or devices the assessment tools are expected to operate with or on. An example of this exception would be a native iOS application that is limited to the supported technologies of those devices.

Specific implementations or deployments (hosting) of technology may also cause compatibility or connectivity issues depending on the level of integration. Loosely coupled, or arms-length, integration typically causes fewer compatibility issues. For example, data integration layers are typically very good candidates for loosely coupled integration. However, tightly coupled integration, such as at the user interface level, can cause technical compatibility issues between components if built by different solution providers using different technologies; an example of this would be in the lack of “single sign on” capability associated with most systems. If the goal is to achieve the most seamless online user experience possible, then often this can become an issue.

Technology is continuing to advance and the ability to leverage and integrate technologies is becoming much more straightforward and flexible. For example, the widespread use of web and cloud-deployed services is breaking down many integration or interoperability issues that have existed in the past. As such, diverse extant capabilities can be leveraged and integrated in customizable ways to create complete solutions.

In addition to the aforementioned customizations of the technologies and solutions used to implement systems developed according to standards, customizations and extensions of the standards themselves are, and must be, possible. In order to be useful, technology specifications and standards must continue to evolve and change as technology evolves and changes. As such, the approach to technology standards should view them as dynamic; given their purpose to drive interoperability, might seem like a contradiction. However, technology standards should allow for multiple integration and deployment strategies. A technology standard that relies on (or requires), for example, a single integration protocol or technology stack will most likely not achieve its adoption targets. Furthermore, open technology standards should not include specific intellectual property or proprietary information that restricts access or usage (adoption); while some technology standards cannot be completely “open” and may require membership or licensing to implement, those should be kept as limited as possible.

Virtually all standards – whether technology implementation standards or data models for content and information – provide means by which they can be customized or extended. For example, it is generally straightforward within a standard’s data representation model to allow for the addition of a metadata field to describe an additional object property. The challenge in allowing for customization of standards is more one of process. Customization of standards, while necessary to support ongoing innovation, can break interoperability if allowed to go
unchecked and without active collaborative planning and decision-making. Standards can—and must by their very nature—provide a minimal set of conformance to verify backward compatibility. It is the process by which “forward” compatibility is verified that requires the most effort. There is an industry-generic tension—one that goes well beyond software and data representation—to the conflicting issues of standardization and innovation. As such, standards bodies must provide for active, inclusive dialogue between stakeholders. Furthermore, standards bodies must support a culture of innovation to allow for decentralized experimentation and evaluation that later inform centralized and formalized evolution of the standard itself.

This culture of innovation will directly benefit learners whose diverse needs can often be met through additional content and formatting/display controls. For instance, customization to support disabled students and English language learners might involve the use of translations, alternative text for images, and linguistic simplifications for words and phrases. Such additions to item data could be additions that occur outside of a standard or they can be made interoperable through inclusion as part of the standard, at the point that sufficient research around best practices has emerged. Other learner-focused customizations may happen at the level of item display or formatting and tend to be system-level features. When content and formatting are adequately separated by an interoperability standard, systems can support learner-centric formatting/display changes (or device-specific changes to item formats) without changes to the standard.

In many ways, the industry is in its infancy in terms of understanding the best ways to support all learners within online assessments. Currently, many accommodations happen outside of the online testing platform and risk a lack of standardization and increased effort from test proctors and administrators. Some of these accommodations could be translated into a form that could be supported by online testing platforms without requiring assistance from a test administrator (e.g., reading aloud an item to a test-taker). However, adequate research has not yet been conducted to understand fully the best way to design and use such accommodations and which policies should be employed to best maintain fairness for high-stakes summative testing. In the meantime, a proper balance must be struck between allowing such research and innovation to flourish within this area before forcing compliance with a single method, on the one hand, and encouraging best practices through compliance with common standards, on the other.

During this period of expanded research and innovation around diverse learner support, extensions to standards are inevitable. To avoid such extensions from unnecessarily compromising interoperability, solution providers should document such extensions using a predefined format that will also serve as input to the standards governance bodies.
3.2.6 Conformance and Testing

**Requirement**

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?

**Response**

Most standards organizations provide methods for verifying (and possibly certifying) that a particular solution conforms or adheres to a given standard and version. For example, both SIF\(^2\) and IMS\(^3\) provide guidelines for conformance to their standards and provide lists of vendors whose products have met those conformance guidelines. These guides define the expectations and process necessary to determine level of conformance by the standards bodies.

Standards bodies such as SIF\(^4\) require fees from the solution providers seeking acknowledgment or certification from the standards body. These fees are the costs incurred by the solution providers to obtain acknowledgement or certification and do not include the costs incurred by the solution providers to prepare the application, implement the test cases, or execute the tests. These costs are variable depending on level of change and prior experience with the process.

In some cases specific test cases are identified and must be executed to test conformance. In addition, automation or test harnesses may be provided for verification of the results.

It is expected with any assessment standards that specifications for testing and verification of adherence can be developed as demonstrated by the current practices in place for SIF and IMS. However, the process rigor, test-case suite coverage, and level of automation provided by the standards bodies can be greatly enhanced. If a standards body “seal of approval” can translate into true plug-and-play interoperability, then these efforts will need to be undertaken. This is not an insignificant undertaking given the complexity of assessments’ content, delivery, scoring, and reporting solutions. While fees may cover the costs of developing verification tools, certification fees should not be so exorbitant as to narrow the field of solution providers. Additionally, while the formal certification may involve a pay-to-play model,

---


\(^3\) [http://www.imsglobal.org/question/qtiv2p1p2d2/imsqti_confv2p1p2d2.html](http://www.imsglobal.org/question/qtiv2p1p2d2/imsqti_confv2p1p2d2.html).

\(^4\) [http://certification.sifinfo.org/docs/SIF_Fee_Schedule.html](http://certification.sifinfo.org/docs/SIF_Fee_Schedule.html).
some verification tools should be publicly available to allow providers to iteratively refine their tools for compliance.

3.2.7 Best Practices

**Requirement**

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?

**Response**

As we have detailed elsewhere in our response, there are a limited number of organizations who have developed standards that address interoperability of assessments. The IMS Global Learning Consortium has developed the Question and Test Interoperability (QTI) specification as an Extensible Markup Language (XML)-based, standardized format for representing test content and student results. This common format is intended to facilitate sharing among authoring systems, repositories, delivery systems, and learning management systems. QTI is the most widely used and adopted assessment interoperability standard in the assessment industry. Several other sets of standards have connections to assessment including the Schools Interoperability Framework (SIF), the Sharable Content Object Reference Model (SCORM), and the Postsecondary Electronic Standards Council (PESC).

These standards vary in their scope, comprehensiveness, and rate of industry adoption, with considerable overlap among them. QTI provides the most complete coverage of item content and test structure (depending on the QTI version) and at this point comes closest to delivering on the goal of sharing test and item content across systems. But even QTI is limited and insufficient to serve as the comprehensive assessment interoperability technology standard, especially in terms of the innovations solicited under the Race to the Top Assessment (RTTA) program. For example, innovative, “computer enhanced” or “technology enhanced” item types intended to support deeper assessment of critical thinking skills are not well supported by the QTI standard.

Assuming further development work will be needed to achieve appropriate interoperability standards that can be used for assessments and related work developed under the RTTA, it is useful to look more generally at best practices related to standards development. For example, the American National Standards Institute (ANSI: http://www.ansi.org/) coordinates, facilitates, and promotes the development of voluntary consensus standards that are relied upon by industry, government agencies, and consumers across the United States and around the world. They have published recommendations for efficient standards development organized in four areas: operations, procedures/processes, forms/documentation, and human
resources (staff and volunteers). Although ANSI standards primarily apply to products and industries, many of their recommendations would serve as best practices to follow in the development of assessment interoperability technology standards.

A second source of relevant best practices for standards development can be seen in the recent activities of JISC CETIS, the Centre for Educational Technology and Interoperability Standards (http://jisc.cetis.ac.uk/). CETIS provides advice to the UK Higher and Post-16 Education sectors on educational technology and standards and sponsored two conferences in 2010 on the development of formal or informal interoperability specifications and standards. The first conference was in January 2010 and emphasized policy and process and the relationship between formal and informal standards-development models. The following summary recommendations resulting from the conference deliberations were provided in a white paper by Yuan, Wilson, Cooper, and Campbell (2010).

- Raise awareness, especially among policymakers, of the diversity of the standards system. Recognize, understand, and work with bodies which differ across a range of dimensions - e.g. legal status, respect, trust, openness and business models.
- Identify solutions for patent, ownership, and licensing issues to enable organizations to adopt, ratify, profile or create derivative works from specifications developed by other bodies.
- Improve transparency across the system and increase effective co-ordination between different bodies through more effective dissemination.
- Understand the drivers and motivations of stakeholders in the domain, manage conflicting expectations, and increase adoption through the involvement of more stakeholders.
- Learn from the culture and lightweight processes of informal specification communities, and provide support for adoption, community engagement and advocacy from incubation to adoption and beyond. Improve the quality of specifications and standards through early implementation and evaluation.
- Use funding support from organizations and governments to make specification documents freely available and release them in such a way that they can be incorporated into application profiles and adapted to meet new requirements.
- Verify all completed standards documents and updates are persistently identified and available for reference by anyone.

Along these lines, we recommend leveraging the cooperation and support of stakeholders to raise awareness, especially among policymakers, of the diversity of the standards system.

---


6 http://wiki.cetis.ac.uk/images/3/38/Fis_whitepaper_final_.pdf
Industry organizations such as the Council of Chief State School Officers (CCSSO) and the Association of Test Publishers (ATP) should be invited to help achieve consensus among stakeholders. As neutral and knowledgeable third parties, such organizations could help avoid potential disputes between competing interests and serve as partners in distributing information and soliciting input and feedback across a wide set of stakeholders.

The second conference in September 2010 addressed technical approaches to creating standards, that is, how standards should be modeled and documented. A number of position papers were developed prior to the conference and discussions from the conference were summarized into themes that included conformance requirements, feasibility for implementation, structure and semantics, building in extensibility and reuse, and gathering requirements. Some of the commentary was very relevant to assessment interoperability technology standards. For example, one participant cautioned that a clear value of the standard or specification to the relevant stakeholder organizations is essential for useful interoperability standards:

At CETIS’ second conference in September 2010, participants addressed technical approaches to creating standards, that is, how standards should be modeled and documented. A number of position papers were developed prior to the conference and discussions from the conference were summarized into themes that included conformance requirements, feasibility for implementation, structure and semantics, building in extensibility and reuse, and gathering requirements.

In general, the CETIS Website seems to be a very valuable forum and tool to promote and support interoperability in education in the UK. A formal effort to develop assessment interoperability standards in the US would benefit from a similar outlet through which innovation and ideas from the assessment and technology communities could be supported. Such an outlet could help to shape the evolution of the more global and formal standards in a more inclusive manner.

Still another set of K-12 education technology standards that provide a set of best practice examples is the National Instructional Materials Accessibility Standard (NIMAS). NIMAS is a technical standard used by instructional materials publishers to produce XML-based source files for use in creating multiple specialized formats—including HTML, digital audio books, and Braille—for students with print disabilities. NIMAS is based on the DAISY Consortium’s American National Standards Institute/National Information Standards Organization (ANSI)/National Information Standards Organization (NISO) 239.86 file format standard. The NIMAS standard was developed by the National Center of Accessing the General Curriculum (NCAC) at Center for Applied Special Technology (CAST) under support from the Office of Special Education Programs (OSEP) at the U.S. Department of Education, pursuant to sections 612(a)(23)(A) and 674(e)(4) of the Individuals with Disabilities Education Act (IDEA). NIMAS source files must be provided by publishers for elementary and secondary instructional materials published on or after 7/19/06, such as textbooks, as specified under

[7](http://aim.cast.org/learn/policy/federal/what_is_nimas)
Parts B and D of the IDEA, and the Chafee Amendment of 1996 of the Copyright Act. As a result, school districts and education agencies are able to provide students with print disabilities access to instructional content previously unavailable or not conveniently available. It is important to note that since testing materials are excluded from the Chafee Amendment, NIMAS is not applicable to assessments.

In addition to serving as a best practice as a useful and usable standard, the process by which NIMAS was developed serves as an example of broad industry and consumer representation.

A notable feature of NIMAS that is relevant to assessment standards is the explicit separation of content from presentation. The NIMAS Technical Assistance Center has developed a best practices resource with exemplars and a style guide⁸.

### 3.2.8 Interoperable Assessment Instruments

**Requirement**

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?

**Response**

Use of a markup language as part of an assessment specification in the strategy will maximize interoperability. Using XML as the base markup language for an assessment standard (which is the case for both QTI and SIF today) provides the greatest level of flexibility for defining data, metadata, relationships/structures, content, and behaviors necessary for any successful assessment standard. The questions posed above reveal some of the real possibilities that can exist if standards are developed appropriately. In order to better understand the possibilities, two “levels” of assessment interoperability need to be considered: content interoperability and assessment delivery interoperability. Each level is described below, as well as some key dependencies that must be addressed.

**Content Interoperability**

The first level of interoperability is at the content layer. For assessments, this would include the items themselves as well as any test or form data, metadata, or structures necessary to define a deliverable set of items that compose a test (either a linear or an adaptive test).

---

This level of interoperability enables two or more item banking solutions (from different vendors) to exchange item and test content, and their definitions. Those items can then be (re)used by the test delivery, scoring and reporting systems in a variety of ways. This level of interoperability will be extremely valuable to states or consortia who own the assessment content and wish to license various vendors’ delivery platforms to deliver, score, and report the assessments. One scenario may be that a consortium is developing content for multiple states and each state can choose the delivery platform they wish to use. Alternatively, the consortium might provide the high-stakes delivery platform for all states but allow each state (or district) to choose a platform to support benchmark or formative assessment that reuses released (previously administered) high-stakes items after initial exposure. A well adopted assessment content interoperability standard will enable the exchange of content for different uses.

At this level, it is also critical to recognize that to achieve a high degree of content reuse that the content markup should be, as much as possible, void of presentation information. This will enable multiple delivery platforms (such as PC, mobile devices, etc.) as well as multiple delivery media (such as paper and online) to utilize the same content standard. Many presentation elements are platform or media specific and therefore should be applied separately to the content specification at some point in the process prior to the delivery of the assessment—implying that layout and style information is also maintained for each presentation format and may also be an area worthy of standards work to be prioritized alongside other areas.

In summary, a comprehensive content interoperability standard can enable exchange of content across multiple delivery platforms.

**Assessment Interoperability**

The second level of interoperability that should be considered is the assessment (or delivery system) level of interoperability. This strategy enables assessment to be “packaged” by the solution provider’s delivery systems and those delivery systems can be activated or called by another system, such as a Learning Management System (LMS). In this scenario, it is likely that the delivery system would also “score” the assessment and return the results from the assessment to the calling system.

This level of interoperability would immediately recognize that not all delivery systems are the same. Even in a highly standardized content world, there will always be competition between the solution providers to have the better test delivery platform or delivery platforms specialized for specific use (such as a math delivery platform). Content interoperability itself is insufficient to allow customers to choose the best delivery platform for their needs.

The Common Cartridge and Learning Tool Interoperability (LTI) standards are of this type for learning content and have had great successes. These standards support the packaging and launch of multiple learning content delivery systems from a single LMS. This type of standard can be extended more fully to include assessments.
This second level of interoperability can best be described as an “assessment service”. In other words, the system that has identified the need for an assessment to be delivered (e.g.: LMS) can call upon the assessment service to deliver the identified assessment and get the results back. This approach will also be very useful in the area of results aggregation and reporting discussed later in this document. In the summative assessment world, there are some exciting possibilities to eliminate the data collection requirements associated with pre-test administration activities for student registration or pre-ID if this approach is taken. However it shifts the responsibility for providing the combined results and test taker data to the local (or calling) system which may not be palatable to state policy makers.

The following diagram illustrates the two levels of interoperability described above. Level 1 or content interoperability is illustrated in the orange arrows and level 2 or assessment or delivery interoperability is illustrated in the red arrows.

Two levels of Interoperability. The diagram illustrates two levels of interoperability for content and assessment.

The two levels of interoperability are intrinsically related. The content level defines what the assessment is “measuring” (i.e., what learning standards the items aligned with) and how the scores are determined (i.e., what determines proficiency for an assessment). The assessment level provides the outcomes of the assessment (in terms of proficiency on learning standards) to the LMS or reporting system. If the content level and assessment level standards do not define these structures in the same way, then a translation is required which
introduces the possibility for information loss. A markup language approach can easily support both levels of interoperability but it would be critical for the two levels to closely collaborate if there are different standards groups at play.

Another key advantage to a markup language approach is that backwards compatibility is better enabled. Markup languages can be extended by adding new information (or tags) without breaking currently implemented solutions. Older versions of solutions can continue to operate; however, any new features or functions added with the new tags would be ignored by the older solutions until they are upgraded. It is important to understand that backwards compatibility will have its limits. “Breaking” changes can occur that invalidate prior versions’ structures and therefore may cause older solutions to also break.

This key point of backwards compatibility can hinder the advancement of a standard—as witnessed by the longer turn cycles on specification releases discussed earlier. If a needed change “breaks” the backwards compatibility chain, it becomes a very onerous process to implement the new standard. This can also impact the longevity of the standard and will often result in completely new standards (and new solutions) emerging and overtaking the old. This process has occurred with many technology standards and has been very impactful on the industry with new industry leaders emerging with each new generation of the standard (e.g., integrated circuits, storage devices, music, videos, etc.).

Prioritizing the Approach

As discussed, there can and should be two distinct levels of an assessment interoperability standard: one that speaks to content interoperability (or content exchange) and one that speaks to assessment interoperability (or the ability to call upon an assessment from another system such as an LMS). Additionally, the interoperability should afford users the opportunity to get results or outcomes from the assessment useful for analysis, reporting, and decision making. It is important to emphasize the distinction between these two levels of interoperability for two reasons. First, if a graduated method of reaching interoperability is pursued, it may be desirable to pursue one type of interoperability in advance of the other. Secondly, different rules for extensions might be applied to content interoperability versus assessment interoperability.

The area of learning content—distinct from assessment content—has benefited from a self-contained module approach, such as facilitated by the IMS Learning Tools Interoperability (LTI) specification. This allows learning content to take any number of forms but still be referenced as a package by different delivery systems, such as LMSs. “Self-contained” refers to the fact that the delivery system need not know what types of interactivity or content is inside in order to make it available to a learner. A similar strategy could be applied to assessment content, with self-contained, self-scoring items. While a strategy of self-contained items (similar to the LTI approach) provides minimal constraints to content developers and allows innovation, a few drawbacks also exist.
A significant difference between learning content and assessment content involves the issue of content ownership. Typically learning content is owned by the same vendor that produces the technology. This is a different model from state or consortium owned assessment content that could ideally be moved easily between vendors and between systems. The self-contained module approach would not allow portability between authoring systems and test delivery systems, for instance.

The self-contained model may not require conformance to a content schema but it does typically prescribe certain technologies to complete the “handshake” between the delivery system and the self-contained object. If a goal is to increase flexibility around delivering items to different devices, this technology prescription may prove to be constraining. Similarly, a self-contained object designed for display on a large monitor may not work well on a smaller display device. System-based accessibility measures and tools (e.g., a highlighter tool or a ruler) would not necessarily extend to the item. Since all scoring would happen inside the self-contained item, limited access would be available for rescoring purposes. File sizes for such items may also be larger since system-level asset sharing would not be enabled for these items.

This is not to suggest that this self-contained approach is without merit, only that it must be used in appropriate venues. For instance, some assessment items may be tightly coupled with the learning content in a formative setting. In these types of learning environments, a comprehensive mechanism for incorporating self-contained modules may already be in place and could be easily extended to apply to assessment content (as some vendors have already done). In such an environment, total portability across systems, optimized displays for different devices, availability of tools, re-scoring, and highly standardized accommodations may not be as large a concern as in summative and high-stakes assessments. Supporting this type of approach within formative assessments may also encourage innovation that could eventually move into other types of assessment, with interoperability standards expanded over time to include sophisticated item types proven successful in other venues. Given these considerations, the following section outlines a path the industry might take moving forward.

**Content Interoperability**

- Converge on a common standard (not multiple standards).
- Focus initial standards development efforts on the content that can be highly standardized such as the more traditional item types and more simple interactive items.
- Conduct a review of the Common Core State Standards and use an assessment engineering approach to arrive at some expanded item types that can be expressed through content standards. Extend the standard to include these item types in order to infuse current standards-based assessments with new capability and potential that is carefully tuned to the needs of the Common Core State Standards.
- Allow the innovative item space to continue to grow and evolve outside of the standard’s control, using the self-contained item approach within the formative assessment standards.
After standard innovative item constructs start to emerge and commonality is recognized, extend content standards to include these new structures.

Apply the current standard to any new content development efforts when the standard can be applied.

Migrate existing content to the current standard formats when vendor exchange occurs and when the standard can be applied.

**Assessment Interoperability**

- Converge on a common standard (not multiple standards).
- Continue to work on assessment tool interoperability standards.
- Continue to develop standards for assessment results (or outcomes).
- Continue to develop standards for assessment results automated data exchange.
- Continue to develop standards for assessment results analysis and reporting.

While this approach does constrain the ability for assessment content (specifically innovative content) to flow freely between vendor’s assessment delivery solutions, it continues to advance assessment standards in areas that will increase interoperability where it is most needed—in the classroom and formative areas. Enabling vendors to package assessment content tightly with their delivery systems (including learning content systems) will allow each to innovate at rapid pace. The potential for new and innovative licensing agreements for accessing a specific vendor’s assessment packages (which is playing out in the learning content space today) may remove real or perceived financial barriers with this approach.

### 3.2.9 Assessment Protection

**Requirement**

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?

**Response**

Validity of inferences drawn from scores is not independent of the purposes or uses of those scores and their resulting interpretations. As such, one cannot list the ways such inferences and, consequently, the validity of the scores themselves may be threatened without taking into account the purposes and context of the assessment. For example, if test questions are
exposed multiple times but the test taker is not aware of this exposure (that is they do not remember, copy/cheat, and or practice such test questions) then there is no reason to believe that such exposure might yield invalid inferences about student attainment of the construct being measured. However, if the context changes and these same questions are high stakes questions then the motivation to remember, copy/cheat, or practice exposed items becomes much greater and the threat to valid interpretations of results in such a context are indeed threatened if not impeached.

At least three factors need to be considered when contemplating how technology standards might help protect the inferences resulting from multiple uses of assessment questions: purpose, context, and degree of exposure. If the purpose of the assessment is to make “high stakes” decisions about individuals (i.e., decisions that have individual student consequences), then the need to control exposure of this content might be greater than when the purpose of the assessment is a self evaluation to provide formative feedback to the test taker and/or to make aggregate program evaluation decisions. High-stakes purposes may include licensure and certification assessments, college placement or college entrance assessments, and graduation assessments. Therefore, technology standards should enable the users of the assessments to delineate in a standardized and interoperable manner, the purpose of the assessment. This could be as simple as allowing common or consistent classification categories (similar to what the US Department of Education requires for breakout group analysis under No Child Left Behind: Hispanic/Latino, American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian or Other Pacific Islander, White or Multi-Racial) for the nature of the stakes or purpose of the assessment. This could be as simple as High Stakes vs. Not High Stakes more complicated like Summative, Benchmark, Interim and Formative or even more elaborate such as Summative Content-Referenced, Summative Norm-Referenced, Summative Skills-Based, etc.

Next, technology standards should allow for some standardized and interoperable definition of the context of the assessment. For example, the context in which students can retake an assessment almost immediately, with memorization or previous exposure to the test questions very likely, should dictate different exposure controls than would a context in which students will see previously exposed content, but content that has not been exposed to them. Technology standards should allow for the collection and classification of such differing contexts. Other contexts might include the number of test forms, groups of items, item pools that are required, and whether item or test exposure will take place within the same testing window, school, tested subject and tested grade.

Finally, contexts regarding preparation and/or review time between exposures might also be part of the technology standards. At a very minimum, the technology standards should allow for the collection and control regarding the amount of exposure test questions, groups of questions or item pools receive. For example, users of assessments might very well want control over how often and how frequently test questions become available for presentation to test users. Some assessment users will want to expose some questions only once. Others will want to expose them multiple but yet a finite number of times. Some may require a specific amount of elapsed time between exposures, for example if a testing asset is exposed...
this week it cannot be seen again until next week. The technology standard should, in an interoperable and standardized way allows users to set and track these exposure parameters.

Exposure context controls are quite common in computer-based and computer adaptive testing. Coupled with large pools of assessment content and a well-defined purpose for inferences resulting from the assessment, these controls minimize the likelihood of cheating or anomalous assessment results (“tampering”).

Because of the different purposes, stakes and context surrounding assessments, it is difficult to find examples of publically available pools of assessment assets that could be used to generate assessments that are likely to yield valid results. In addition, many current large-scale, high-stakes assessments release the questions to the public following use. While any one release might be too exposed to serve as a practical pool for future assessments of the same purpose, the number of these items exposed across multiple administrations could grow so large that sampling from this exposed pool may not constitute much of a threat to the validity of the resulting scores.

3.2.10 Security and Access

**Requirement**

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)?

**Response**

Core security and controlled access are a few of the key advantages online assessment administration has over other forms of evaluation. Arguably, technology standards can be used to help maintain testing security in three areas: enrollment, environment, and reporting/data. In the area of enrollment technology, standards can help to verify that only the correct people are afforded access to the testing environment. Secure access can be created via the usual or traditional means of security, i.e., encryption of information during communications processes, or caching encrypted information away from examinees, proctors and test administrators. It can also be achieved via other less-traditional uses of technology standards for assessments such as certifying via image/character recognition or verification of IP addresses via registered e-mail exchanges or even biomechanical verification, which has been the subject of much research lately. Such devices or procedures for controlling the enrollment will only work if technology standards are clearly defined, standardized, and interoperable so various delivery platforms, enrollment data base access and information exchange do not rely upon priority software and/or hardware investments.

Technology standards can also help secure the testing experience itself by controlling for the testing environment. Seemingly innocuous aspects of the testing environment like the need to
lock down the desktop and restrict access to hard drives, the internet or other software (preventing the capture of testing information depending upon the stakes but also preventing the illicit use of support materials or simple cheating) will also help preserve the integrity of the testing experience. All of the factors ultimately enhance the validity of the resulting inferences made from the testing information itself.

For example, the use of “USB keys” which contain the separated item pools and testing presentation formats has been used by the Organisation for Economic Cooperation and Development (OECD) for the Programme for International Student Assessment (PISA) assessment internationally for some time. Such a procedure controls who sees what content and when. It also standardizes the testing experience by taking control of local machine operating systems regardless of technology infrastructure and internet accessibility, and it places the management of the general testing endeavor in the hands of test proctors or administrators.

Again, depending upon the purpose of the assessment, technology standards need to take into account a range of test uses and/or purposes. For example, the technology standards may need to anticipate multiple levels of assessment security. The lowest level could be essentially non-secure test questions. These questions could be a listing of items previously released to the public by states, schools, or vendors and anyone can access the listing at any time for any purpose. In this case, the technology standards would likely focus on how to exchange this content such that it can “play” on a host of various delivery platforms, operating systems and hardware configurations.

Another more secure pool of content could be content collected in “digital libraries” in a manner similar to that described in the Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER-Balanced consortia plans for RTTA. This content could have previous test use data/information, be linked to an underlying score scale and have measures such as Lexiles™ (for text complexity) associated with testing content. If so, the investment in this content to collect this information might be such that simply putting it out for everyone to use at any time may not be prudent. In this scenario, technology standards could be used to restrict access to this content via the enrollment process and/or in the presentation of the content either to the student or via restricted presentation of the statistics and collected information. For example, the technology standards should probably facilitate how Lexiles™ would be used but would not allow manipulation or changing of the Lexile™ values themselves.

Finally, technology standards have the potential to expedite and enhance the value of information resulting from assessment administration. Information will be presented consistently and accurately at the test-question level, the individual student-candidate level, as well as other levels of abstraction (e.g., relative to a norm pool; relative to a college passing standard; or relative to an aggregation level such as a classroom, campus, or school district). Technology standards can standardize how the test information is stored such that exchange of information across various testing entities can be automated, reducing labor
hours associated with data exchange as well as reducing errors in the exchange and use of such data.

Having nothing more than a clearly defined “old fashioned” record layout, data dictionary, listing of variables, as well as automated or semi-automated query and information/report extraction (e.g., exporting to a spreadsheet, PDF, or School Information System) would be an improvement over many current procedures. Such technology standards could then also standardize protocols for who receives access to what data and in what fashion, resulting in a user experience that is comparable regardless of need but created in an interoperable manner such that expensive technical support, training, and human intervention is avoided. Single sign-on protocols may be one way to help maintain this ease of information or results exchange, but standardized data disposition and interoperable data and information exchange protocols might suffice without the need for the more ambitious and complicated effort of single sign on.

3.2.11 Results Validity

**Requirement**

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?

**Response**

Because assessments are contextualized by a purpose and desired use, there is no way to verify that the inferences made from the results of an assessment (i.e., the very definition of the term validity) will be valid or even reliable (i.e., repeatable). Nonetheless, there are clear steps in the establishment of technology standards that can enhance the likelihood that inferences will be valid.

Interoperable technology standards can help maintain that test questions constructed for and tried out in one context are legitimately applied in another context so that change in context is controlled for or removed as a potential threat to validity. For example, a multistep, innovative, online, inquiry-based science item that is highly interactive and was originally developed for delivery on current technology via the internet will not measure the same thing as when presented on a third or fourth generation older device supported by a slow LAN without internet connectivity. The inferences resulting from comparing the two different sets of results will likely be quite disparate. In other words, the reliability and validity of the assessment results may only be applicable to environments similar to those in which evidence was originally gathered. Therefore, technology standards can describe which environments are required for valid inferences.
An analogous question is when does screen size become so small such that inferences resulting from the assessment are called into question? Presumably there would be little threat to the validity of assessment results if some test takers used computers in a lab equipped with 21-inch monitors while others used 19-inch monitors. But will the same inferences result if one student uses the large monitor while another uses a laptop with say, a 14-inch monitor? What about a Netbook with a 10-inch monitor? What about an iPad, Smartphone, or Blackberry? While technology standards may not be able to answer these questions, standards could define the minimum screen size requirements so that content can be exchanged and used in an interoperable fashion without threats to the inferences (validity) of the resulting scores.

3.2.12 Results Capture

Requirement

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?

Response

Arguably, capturing a fuller, more actionable view—a 360 degree view—of student performance by linking information and data of all types (individual learner information, assessment results, assessment content, and other collateral data typically contained in school information systems or transcript centers) is the essence of the value proposition for interoperable technology standards. It is unlikely that we can craft and/or afford the time and money that would be required to make direct changes to disparate and proprietary systems in order to create a ubiquitous user experience immediately. Rather, a more prudent way to gain compatibility is to make existing and future systems follow the same set of standardized and interoperable information exchange protocols so they can share both inputs (i.e., content and enrollments) and outputs (i.e., student results, assessment data) in order that this information can be shared across all aspects of the information lifecycle.

This interoperability is not “pie in the sky” future planning but is being addressed by various organizations now. For example, IMS (http://www.imsglobal.org/question/) has been championing the Question & Test Interoperability (QTI) standard that attempts to define exchange standards for test questions. The SIF Association is dedicated to interoperable data exchange (http://www.sifinfo.org/us/index.asp).

Each of these organizations and their technology standards are pushing the envelope toward interoperability, but each organization is essentially dealing only with one dimension or one piece of the interoperability puzzle. QTI has been focused on the test question interoperability, while SIF is focused on data and information interoperability. Neither of these
A comprehensive set of technology standards such as those envisioned in this RFI could bring these separate efforts together such that a coherent and comprehensive assessment lifecycle set of standards for interoperability could be defined. If so, the existing technologies and delivery platforms could build future releases using these comprehensive interoperability standards such that anyone building to them could exchange content, information and assessment results without the burden and restricts that currently exist. The Department of Education would be the idea choice to guide the cooperation of various entities in engaging in a collaborative effort that will result in interoperable data, assessment, and information exchange without threat to proprietary or entrepreneurial investment and motivation. In other words, vendors (both for profit and not-for-profit) should embrace such comprehensive interoperable exchange standards because these standards will make their content, services, information and technology more widely accessible to an even broader pool of potential users. As previously mentioned, industry organizations such as CCSSO and ATP should be invited to help achieve consensus among stakeholders.

3.2.13 Results Privacy

**Requirement**

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?

**Response**

One could conceptualize “results privacy” as stated in the requirements here as comprising at least three distinct aspects of security. First, there is the pure physical security of information, which involves traditional issues of confirming that only authorized people log on to a secure database, Web site, or information/assessment exchange. Second, there is the disclosure of personal or private information that might happen inadvertently even if system logon and recognition is secure and accurate. Third, there is the need to secure derivatives or derived information resulting from analysis of seemingly private or secure information. All of these concerns exist for the scenario of single service provider as well as when multiple service providers are involved. There are certainly other concerns as well, but we will focus on these three as representing the majority of such concerns when considering technology standards for interoperability.
Physical security may be the most well researched and currently controlled for aspect of information and data exchange. Protocols for secure logon, unique identifiers, security questions, IP address verification, encrypted transfer of secure information, secure transfer protocols (such as HTTPS) are in use and should provide some guidance for the types of technology standards needed to secure a testing environment.

There is no reason to believe that assessment information (enrollment, test content, or results) should be treated any differently than bank accounts, social security numbers, passports and online commerce when it comes to privacy. While efforts to maintain physical security for such information are underway and improving continuously, these efforts are limited in that they focus on one use and one set of data that presumably is not shared or exchanged across multiple entities for multiple purposes. For example, it is unlikely that your bank will share your personal overdraft history with another bank routinely and publically whereas a student’s performance on a test question might very well be shared with others either explicitly or inadvertently (e.g., the student’s essay response and score could be included on the parent report and sent to one or more higher education institutions).

This second type of security, which addresses privacy, can also be addressed with clearly defined technology standards for interoperability and data exchange. For example, the National Transcript Center (http://www.transcriptcenter.com/ntcprofile.php) has been in the business of secure information exchange relative to high school and college transcripts for some time. Their privacy policy (http://www.transcriptcenter.com/privacypolicy.php) is very comprehensive and built upon the premise of fully disclosing what information is shared and to whom, seeking consent prior to the disclosure of private information and maintaining this privacy via policy when interacting with third parties. Hence, a comprehensive set of technology standards should include or facilitate the collection of consent, the range of potential users of private information and who they will be using this data as well as procedures in place to track if the recipients of this information are indeed following protocol.

Another aspect of results privacy involves maintaining that users or recipients of the information cannot discover or derive private information from an analysis of what would otherwise be non-secure (i.e., data stripped of identifying information). For example, simply masking the names of two schools from one district in which one school is historically a top achiever while the other school is an underperformer is insufficient to protect privacy, because all knowledgeable users of such data will know which school is which, thereby exposing school staff to potential public ridicule. While this seems an obvious error in this trivial example, controlling such use is not so simple when dealing with individual results and dozens, if not hundreds, of variables that are likely to be disaggregated in unknown ways.

For example, assessment results disaggregated by school, subject, grade, ethnic background, and special needs classification might result in a disclosure that one student who is at the lowest level of the distribution or group and who is a minority with a special need did poorly on the assessment. It might not be obvious to an outside observer who this individual is, but it would be painfully obvious to others in that school, subject, and grade who
that individual is likely to be. The technology standards could outline and enforce protocols for disaggregation. For example, the standards might require that disaggregated information be repressed when the total size of the disaggregated group falls below some arbitrary and/or fixed number.

Another area where technology standards are likely to help enhance security is in the area of derived or “mined” information. Google has been in the media recently because they use patterns of Web search behavior to route retailers to users of Google for marketing purposes. Hence, from a privacy perspective, it might be an issue if service providers targeted, for example, remediation opportunities at students who did not perform adequately on assessments, as this would identify these students as poor performers. Yet the very value of mining such assessment data is to identify those who would benefit from various improvement programs so a sort of “Catch 22” would seem to exist. Technology standards could help remove this catch by delineating what information regarding consent must be collected, how information about remedial or other opportunities should be conveyed to students, parents, and teachers and in what venue, as well as accepted and unaccepted practices of data mining and information derivation.

### 3.2.14 Anonymization

**Requirement**

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?

**Response**

What is referred to as “anonymization” as been the standard for researchers using assessment data for quite some time. For example, if a researcher wants to see the relationship between performance on a statewide assessment and the ACT or SAT and the researcher secures cooperation from the state authority, the process is something like the following:

- The state authority (who owns the secure data with individual identifying information contained) shares the “master file” with the owner of the ACT or SAT (in this case, ACT Inc., or the College Board).
- ACT or the College Board then matches the file with their database (presumably using student names, gender, date of birth, unique identifier, state, district, school, and grade).
Once the data is matched such that a new master file exists with both the state assessment data and the ACT or SAT data, the file is “scrubbed” where any personally identifying information is removed (the process this RFI calls “anonymization”) and the clean file would be provided back to the state authority who then provides the clean file to the researcher.

Presumably the researcher could then do anything he or she wanted with or to the file and not be able to identify individuals. As such, the process of anonymization requested seems to be accepted methodology for much data exchange and analysis. Technology standards can help this process by requiring standardized and interoperable data requirements such that fields, field length, data values, data definitions, data dictionaries, formats, media for exchange and exchange protocols are all defined and known so that the actual process of anonymization of the data could be automated.

In addition, and relevant to the previous section (Section 3.2.13 Results Privacy), technology standards could also define protocols for use and/or disclosure of data to further help maintain that no private information is inadvertently disclosed even after the data is “anonymized.” For example, a minimum size for a disaggregated group could be specified so that results, even anonymous results, could be suppressed if the number fell below minimum.

### 3.2.15 Scoring and Analysis

**Requirement**

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?

**Response**

In order for assessments to be “fully” scored so that appropriate analysis can occur, the technology standard must support all data and metadata necessary for a scoring system to complete the scoring process after the test (or item) has been delivered. This would include all item level data and metadata such as answer keys, rubrics, automated scoring algorithms, and weights, plus all of the form or test level scoring data and metadata such as score tables (raw to derived scale conversions), attemptedness rules, and performance levels with their associated cut scores. In addition to item and test level scoring data, the specification must support the ability to group items related to a common learning construct (often referred to as a strand or cluster).

Much of the information required to score the item is determined by the content and psychometric staff building the assessment and should therefore be captured as the content is being developed. In addition, by placing the information with the content, then multiple delivery platforms can be enabled as all the information necessary to deliver and score the assessment lives with the content.
However, this greatly extends the scope of the content interoperability standard level to include more than just item level content and information. It must also include all of the information necessary to structure the assessment, group items, provide score tables, cut scores and performance levels among others. This is why the current, yet well defined, interoperability standards like QTI and SIF are not sufficient.

Conversely, if the content layer is not able to specify all data and metadata necessary to calculate the final score (or outcome), then this information must be injected into the delivery and scoring system separately and independently. This is common practice today in the legacy systems that process paper-based assessments. Separate structures and scoring information (often referred to as “test definitions”) are provided in offline inputs such as Excel spreadsheets or parameter databases. This can cause significant synchronization issues between the content levels and the delivery and scoring system levels, since changes to one immediately impact the other. It also yields considerable quality risk as all the information required for testing is not contained in one place.

In addition, as item types become more sophisticated, specifying all possible outcomes in a standard will be a very difficult proposition. No longer will traditional “keys” or lists of values be sufficient for scoring. Alternate approaches may include imbedding the scoring “logic” within the item itself or specifying external scoring algorithms (by reference) that can be called by the delivery system after the item has been delivered. These approaches may simplify the standard definition; however this may limit the level of interoperability of the content between delivery systems.

It is expected that a well designed assessment content standard could address the scoring needs of most assessments that are delivered today, and if close consideration is given to the future direction of assessment content development, structures can be put in place to support this as they evolve without future breaking changes.
3.2.15.1 Results Aggregation and Reporting

Requirement

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 (see also 3.2.14 above).

Response

Aggregation of results typically requires the intersection of two types of data, the assessment results and the data associated with the test taker (demographic data), organization data (district/school) and in the K–12 sector, potentially the teacher. It is mitigated based on the intended purposes of the reporting.

In addition, there are at least two different use cases for how results are aggregated.

1. **Results reported based on the data at the time of testing.**
   
   This strategy is typically used for state accountability reporting. Results are calculated and aggregated based on the data that was captured when the test was given (or on a specific date) and does not change as the test taker’s data changes over time. As an example, a test result would be aggregated and reported based on the school where the student tested or for the primary teacher for the student at that time regardless of where the student may be today.

2. **Results reported using current data.**
   
   This strategy is more typical for reporting systems used at the local level (as an example: to inform instruction). In this model the data is reported and aggregated based on the current student's data. This model allows a teacher to access last year's results for the students currently in his or her class.

These two use cases may dictate system and database designs specialized for each case. The first use case may be more of a true longitudinal system (or warehouse) that tracks data over periods of time and is able to capture “snapshot” data for historical purposes. The second use case may require an Operational Data Store (or ODS) approach that is able to recognize and react to data changes in near-real time. In addition, the first use case is typically a service provided by states as part of their accountability systems, while the second use case is typically a local service integrated with other local systems. Again, this is affected by the purpose of the results—in this example, for program evaluation in the large-scale high-stakes arena or for instructional improvement in the classroom real-time environment.
In either case, if the standard for defining the reporting outcome of an assessment supports the intersection of both results and demographic data, then the reporting system itself can determine how data changes over time are handled. This could enable the development of analysis and reporting “engines” that are independent of both the delivery system and the use case in which the data is used. An analysis and reporting service could be developed that is given the set of outcomes formatted using the standard for a group of students and the engine would perform aggregation or calculations that can then be returned or used by a reporting system as dashboards, reports or data.

Because the intersection of assessment results and test taker demographic data is necessary to support a reporting system, then a larger design decision would need to be made concerning responsibility to collect and store the demographic information within the assessment delivery systems or externally. Note that this demographic data is not limited to physical attributes of the student or test taker. Rather this might include previous instructional interventions, at risk or remedial classifications, and a host of other information relevant to making instructional improvement decisions based on reported results.

In today’s summative assessment world, the assessment systems typically collect all demographic data necessary to report results (often referred to as a registration or pre-ID process). If we consider a different model of “assessment as a services” model as discussed previously in the Interoperable Assessment Instruments section, then the responsibility for collecting and storing (and taking snapshots) of student information remains in the system that activates the assessment (such as a Learning Management System LMS). In this scenario, the assessment delivery system would only need to collect the test taker unique identifier so that it can be used later to link back to the test taker information.

**Data Anonymization or Data De-Personalization**

Assessment results are often used to communicate information to the public or to be used for research purposes. In either case, protection of personally identifiable information (PII) is of critical importance and is controlled by law.

Publicly consumable information is generally the aggregated summary results and never includes individual test taker details. In order to protect identity, a common practice is to mask or hide results when a summary n-count drops below a certain level (for example, do not report results if less than 10 test takers are included in the summary numbers). This will verify that it is impossible to specifically identify an individual from the summary data. It is not clear if a standard n-count is defined or accepted by all users and is therefore unclear how an assessment standard would support this activity.

Researchers typically need access to the detailed information, just not the test taker identifying information such as identifiers, names, SSNs, etc. However, researchers do want access to certain demographic data elements about the test takers such as gender, race,
ethnicity, and grade level to conduct the research. Any outcome reporting standard should identify which attributes in the data are considered personally identifiable elements. Reporting or data generation system that expose data to researchers would mask or remove those attributes from the data set. Because exposing personally identifiable information can have severe consequences, it may be necessary to have a separate outcome-results standard that removes the PII entirely, eliminating the possibility of a software “glitch” exposing the data unintentionally.

3.2.16 Sequencing

**Requirement**

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?

**Response**

For linear or traditional forms-based assessments, item sequencing is typically performed when the form is constructed by the author, psychometrician, or user of the assessment system. With this type of testing, and assuming the standard supports multiple form definitions for a given assessment, item sequencing is easily accomplished through numerical sequencing of items as a metadata attribute within the form specification (part of the content specification).

For non-linear testing, such as adaptive testing, there are many other attributes that need to be considered for construction of a test, since no real form or infinitely many forms might exist. Because the adaptive test would typically pull or select items from a pool, the algorithm will need access to item statistics, test starting parameters, content balancing parameters, exposure controls, termination parameters, to mention a few. It is expected that all of these parameters can be accommodated in the technology standards for all of the same reuse reasons outlined in Section 3.2.15 Scoring and Analysis of Results of this document.

At the core of an adaptive test is the algorithm which provides the logic for determining how items are selected from the pool of items available as the test taker progresses through the test. It is not expected that the assessment standard would identify all possible algorithms but would provide the necessary item and form parameters to enable the algorithms. The standard would need to be able to identify the algorithm (by name and/or location) so that the delivery system can call upon it at run time. Some algorithms may be proprietary solutions licensed by the owners while others may be available publically. The standard should be designed to accept any algorithm that can adhere to the parameters defined for the item and form within the specification (in other words an application or service interface or API). This is
especially true given the likelihood that additional future algorithms will be developed which may or may not be proprietary.

It is important to note that interoperability standards must support the construction of an assessment that may contain several sections, potentially with both linear and adaptive sections in the same test. As an example, the test may start with a sequential set of practice items where all students see the same items in the same order. Upon completion of this sequential section, the adaptive section can start. Following the adaptive section, the test may then issue another sequential set of items—such as survey items. There are many other use cases of how a multi-section test can be constructed.

Other characteristics of an adaptive test standard that should be considered are groups of items that may reference a shared resource (such as a reading passage). In this case, it is likely desirable to deliver a set of questions about the resource as a unit. This design could also be extended to enable adaptive “testlets” (which are small pre-packaged groups of items delivered as a unit). Again, all of these relationships between items and resources would need to be addressed to verify that adaptive algorithms can function properly without the need to inject external (asynchronous) information.

Item exposure controls are also a key consideration for an adaptive algorithm. Any standard for managing an item pool or launching an adaptive test may need to have access to item exposure parameters such as how often the item has been seen in the past. Some adaptive programs have required that an item not be exposed more than once to an individual test taker if they have multiple opportunities to take the same test.

Field testing items is another key element of the assessment standard and delivery system. Knowing which items are to be included in the field test item selection process and how frequently or evenly (or unevenly) dispersed those items are will all become attributes of the item and adaptive form parameters. Transitioning items from a field test status to an operational status must also be considered.

A key design consideration for the technology standard as well as the solutions implementing the standard is how the adaptive test accesses the pool of items. Directly accessing the “item bank” may not be desirable as the item bank may be evolving and changing over time. For example, items may be undergoing revisions that are not immediately available to the operational test. There are many design patterns that can address this concern both within the standard and/or within the delivery systems themselves. The standard should identify the options for implementation.
3.2.17 Computer-Driven Scoring

**Requirement**

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 bubble-fill scoring?

**Response**

A well-formed assessment technology standard should allow a provider the ability to integrate its delivery solution with computer-driven scoring systems, including automated scoring technologies, or to integrate with human or professional scoring systems. These types of higher-order scoring systems are often required to analyze responses to open-ended items, and require a higher level of sophistication in terms of scoring requirements than do items that constrain the test-takers response and can be usually be machine-scored, such as multiple-choice or drag-and-drop items.

There are at least two elements required for a scoring system to successful; 1) the response data and 2) the scoring rules or engine. In order for the scoring engine to operate correctly, it must understand the structure of the response data. Therefore there is a close relationship between the specification for how response data is formatted and what response data is available and the scoring process itself. Responses can come in a variety of types and formats such as text based (e.g., an essay), MathML for constructed math responses, binary audio or video files for portfolio, language or other types of assessment are a few examples. More sophisticated (or innovative) items may provide interaction data such as points clicked, figures drawn, tables completed, etc. Instructional based assessments may also include information such as time-on-task, attempts, etc. to diagnose understanding or level of engagement.

In addition to specifications for response data, an assessment specification must support traditional constructs that are necessary to support intelligent (machine or human) scoring that must be included in the specification. The scoring rubric is one of the constructs that must be supported. The scoring rubric is one such construct. The scoring rubric identifies the score points that can be attained and describes the characteristics of responses at each score point. A rubric may have multiple “traits” that are also evaluated. In addition to the scoring rubric, often exemplars are provided that represent actual test-taker responses at each score point level.

Much like an adaptive test algorithm, the technology standard must be able to identify the scoring algorithm or system (by name and/or location) so that the delivery and scoring systems can call upon it at run time. Again, the technology standard would facilitate the
exchange of information between the scoring algorithm (with a defined set of parameters that exist within the specification) and the delivery system.

The technology standard must also identify rules for how many times a response should be scored. With computer-based scoring, it is possible that the scoring algorithm cannot determine a score and must then be routed to a human scorer. Other use cases for computer-based scoring require a sampling of responses to be scored by a human as a quality control measure. For professional scoring use cases, a response may be scored differently by two human scorers and then scoring rules are applied to determine the final score or if the score must be adjudicated.

Regardless of whether the responses are scored by a human or machine, a “training” process must be conducted to verify that the human or machine understands the rubric. Sample test taker responses are collected and first evaluated by subject matter experts to find good exemplars. These responses are then used to train the “system” on what to look for by score point. It is not expected that these processes would all be supported by the assessment technology standard. In fact, it is only the “outputs” from these initial manual processes that drive the final scoring processes that need to be captured and instantiated in the technology standard.

### 3.2.18 Formative, Interim and Summative Assessments

**Requirement**

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 non-traditional assessment methods, such as evidence, competency, and observation-based models?

**Response**

What defines an assessment as formative, interim, or summative has less to do with the instrument itself and more to do with the intended use of the instrument and the results it produces. Formative assessment, for example, is a process that involves both assessment and instruction; while results from test instruments can be used to inform this process, no assessment instrument in and of itself can be considered a “formative assessment.”

Nonetheless, there are (or should be) clear differences in the instruments used to support these different types of assessment. Not only will test content vary as a function of intended use, but factors such as the frequency of testing, setting, mode of scoring, and availability of results will impact the media and technologies used for testing. For example, summative and interim assessments are generally administered as isolated, formal events, while instruments used to support the formative assessment process may occur embedded within instruction
and thus are more likely to be represented as part of learning objects. Consider student response systems (aka “clickers”), which may be used as a part of formative assessment. Representation of such items would ideally support the ways in which results may be used on the fly, such as real-time diagnosis through distracter analysis, and even real-time instructional prescription. Current assessment standards may be insufficient to represent the information needed to support these practices, while those used within learning management systems might offer increased flexibility.

The technologies and technology standards used to support these testing instruments vary widely as a function of how they are administered and the media used for administration. At one extreme, most summative assessment still occurs using the traditional technology of paper and pencil. Content nonetheless is typically created and maintained digitally. Items and tests can be generated using customized or commercial tools including Quark, Adobe InDesign, and Microsoft Word, with and without customized templates and scripts. Adobe PDF represents a highly standardized means for consistent printing of tests regardless of how they were created. The test content assembled using these aforementioned tools can be represented in a number of ways, including using standards such as IMS Global Question and Test Interoperability (QTI; though this standard was designed for and is more commonly used to represent computer-based tests), or using proprietary XML-based data representation schemas using generic commercial and open source XML editing platforms such as XML Spy and W3C Amaya.

Online-administered summative assessment adds to these sets of tools and technologies. For example, rather than using Adobe PDF for standardizing the representation of tests, many systems rely on display using free, commercial, or embedded/proprietary web browsers such as Mozilla Firefox, Microsoft Internet Explorer, and Apple Safari. Various standards exist for such administration, including HTML, XHTML, XML, CSS, Flash, using a suite of free and commercial tools including W3C Amaya, Adobe Dreamweaver, and Adobe Flash. As with paper-and-pencil tests, test content can be represented using QTI and other XML-based data representation schemas, as well as using other standards such as SIF and SCORM.

Instructionally embedded assessments (e.g., those used to support formative assessment) add a layer of complexity to administration of tests and also open up the applicability of an additional set of standards for representing content. Again, at one extreme are textbook end-of-chapter questions and teacher-generated paper-based quizzes. The technology and tools here overlap greatly with the paper-and-pencil technologies used for summative and interim assessments. Where instructionally embedded assessments are distinct from summative and interim assessments is when they are made available digitally. A host of standards to support learning all have means of representing assessments: IMS Global Learning Tools Interoperability (LTI) and Common Cartridge, IEEE Learning Technology Standards Committee (LTSC) Learning Object Metadata (LOM), Sharable Content Object Reference Model (SCORM), School Interoperability Framework (SIF), and Centre for Educational Technology and Interoperability Standards (CETIS) Joint Information Systems Committee (JISC). Other standards are also frequently used to represent learning content, such as
DAISY, DocBook, and the National Instructional Materials Accessibility Standard (NIMAS), but these do not provide standardized means for capturing student responses.

Regardless of the type of assessment, accessibility and flexibility are critical to help verify that all students have equal opportunity to demonstrate their construct-relevant knowledge, skills, and abilities. For example, the Accessible Portable Item Profile (APIP) specification has recently been drafted to extend QTI to support provision of accommodations for students with disabilities.

Finally, many non-traditional assessment methods could currently be handled generically through various electronic portfolio schemes and tools, both proprietary and standardized including the IMS Global ePortfolio standard. In addition, various proprietary and open source solutions for representing student- and teacher-generated content are available such as Microsoft Office, OpenOffice, HTML, and Kaltura for video. However, there is an obvious tradeoff in standardization of response and scoring for this added flexibility in capturing student work. Additional work by various stakeholders will be necessary before such representations can be standardized enough to allow for interoperability in item development, banking, administration, scoring, and reporting.

The wide scope of technology standards and interoperability issues for representing assessments across summative, interim, and formative situations illustrates the need for a comprehensive set of lifecycle approaches to support the next generation of assessments and instruction supporting educational reform. This can only be accomplished successfully with broad representation of stakeholders to develop discrete processes for standards evolution to support future innovation while not impeding its growth.

3.2.19 Learning and Training

**Requirement**

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?

**Response**

Formative assessment is the process by which assessment results are used to shape subsequent instruction. Most learning management systems and the standards that support them (e.g., IMS Global; IEEE LOM; SCORM; SIF; CETIS) provide means for representing test and/or item content such that it can be administered to students and results can be stored. The challenge, however, is in linking appropriate instruction as a function of test results and perhaps even more importantly, collecting, maintaining and providing efficacy regarding the linked instruction's ability to improve student learning. Models for accomplishing this vary from simply creating student study plans that contain instructional modules aligned
with items students haven’t yet demonstrated a set level of mastery, to using intelligent
tutoring systems where explicit domain, pedagogy, and student models which inform content
and support selection.

At the very least, assessment items and learning objects must be tagged at the highest level
with content standards that allow them to be associated with one another—but this is the
“Catch 22”, if that tagging is too high (general) or too low (specific) it becomes almost useless
as a means to improve instruction. More sophisticated models require deeper linking
methodologies that rely on pedagogically and/or cognitively oriented ontologies—or other
classification protocols like Latent Symantec Analysis. The IEEE LOM schema\(^9\), for example,
provides metadata fields such as Learning Resource Type, Interactivity Level, Semantic
Density, and Typical Age Range that can support deeper linking of assessment items with
suggestions or prescriptions of instructional interventions. These can support various learning
situations ranging from formal to informal, from institutional to on-the-job training, and from
civilian to military.

Despite the current efforts, the goal of the “Holy Grail” of improved learning is to be able to
provide teachers, students and parents the opportunity to select whatever content is the most
appropriate based on feedback provided during instruction and assessment to maximize the
efficacy of instruction. This “Holy Grail” will require a technology system that is agnostic to
where the content resides, yet is still able via technology standards to “understand” what
content and aspects of learning support are required based on data gleaned from an
assessment or classroom behavior and to be able to offer this to the user seamlessly, without
a large technology footprint and in near real time. Let us hope that our vision for the Holy
Grail not become a lifelong quest but rather, that a set of coherent interoperability and
exchange technology standards can be constructed starting from where we are currently
which will evolve into where we need to be in the relatively near future.

### 3.2.20 Repositories

**Requirement**

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?

**Response**

Many different types of learning management systems and item repositories exist ranging
from very data applications like HyperCard to simply databases like Microsoft Access to
sophisticated authoring systems like Documentum. All of these systems have their strengths
and their weaknesses depending upon use, technology requirements and legacy constraints.
The following paragraphs present a general outline of item repositories.

Item repositories can combine several types of data (note that analogous considerations are made for instructional learning repositories as well):

- **Content representations** that will be viewed by the test-taker when the item is displayed by a test delivery engine: item type, item text, references to graphics, feedback for incorrect answer selection in a tutoring system.

- **Metadata** used as selection criteria for constructing a test: performance statistics, content standards alignment, friend/enemy relationships (items that should or cannot appear in a test together due to cuing), inclusion of ethnic names or images of disabled individuals when certain percentages of such names or images must be included in a test.

- **Item development process data**: prior versions of the item, versioning data (edit date, description of change, and person responsible for making the change), client and reviewer comments, stage in content lifecycle, fact-checking data, image rights permissioning information, internal unique identification numbers for items and assets.

- **Item usage history**: number of uses, dates of test administrations.

Within existing item repositories, the data type that is most likely to conform to existing standards is the content representation. However, proprietary formatting conventions (in the case that a standard does not separate content and format), extended item types, the desire to represent content types not covered by existing standards, and the need to deliver content outside of browsers (in the case that a standard prescribes HTML-based delivery) are all factors that can lead to vendor decisions to extend beyond or altogether abandon existing standards. In addition, the choices regarding what data aspects of the items are considered metadata and/or decisions about what data stays in the repositories or databases can also be a limiting factor to the sharing of these assets across systems.

Item development process data is typically the least likely of these data types to conform to interoperability standards. Item development processes can be vendor, system, or client-specific rather than subject to standardization. Some process differences are determined by the assessment type. For instance, the level of rigor pursued in reviewing an item would logically differ when considering an item written by a teacher for use in an online quiz versus an item used in a high-stakes assessment, which may be fact-checked, sent to a bias committee for review, reviewed by a board of educators, and submitted to usability testing in the case of an interactive item. Thus, item repositories tend to be designed for specific types of assessments. Collaborative item banks for sharing teacher-created items may contain peer-provided rating data but would be unlikely to include performance statistics and bias committee review comments.

Many item repositories differentiate between exchange format, repository format, and delivery formats for items. The exchange format is the XML representation of an item that conforms to interoperability standards. The repository format is the XML-based representation of an item that includes all of the data types described above and which would be specific to the item bank (or to a customer-specific configuration of an item bank). If items are authored outside
of the item repository, such as in a different vendor’s system, then the items are imported in the exchange format and then automatically converted to the repository format. If the repository format has mandatory fields that are not a part of the exchange format, then typically a conversion routine would automatically assign values or a system user would be prompted to enter the missing data. The repository format allows proprietary data to be added as an item moves through the content lifecycle (reviewer comments, versioning data, etc.). If the item needs to be moved to a different system, such as when a state switches vendors and wants to transfer an item bank from one system to another, then the item is converted from the repository format to the exchange format. All proprietary data is stripped off during this conversion process.

Some standards governance boards explicitly differentiate between exchange formats and system-native formats when providing implementation recommendations. For instance, IMS recommends QTI as an exchange format but does not require that a system—whether an item repository or test delivery system—store and utilize QTI natively in order to be considered QTI-compliant. In fact, IMS recommends against native consumption of QTI by a test delivery system whenever parsing performance may be compromised (e.g., delivery on mobile devices or the need to have low system requirements for school computers so that economically disadvantaged schools are not barred from participation). With the introduction of the Accessible Portable Item Profile (APIP) draft, the distinction between exchange format and delivery format becomes even more critical. An APIP-compliant item may include a number of different alternate representations of the item. Thus, the file size for an item (in its exchange format) is significantly larger since it is carrying the data necessary for many different renderings of the item. It may also carry additional files such as text-to-speech assets. These assets and the data for all possible representations of an item would not necessarily be provided to the delivery system for every student, unless a school, state, or district was known to have unlimited bandwidth.

More typically, to optimize performance and minimize bandwidth demands, test delivery systems utilize a delivery format for the item that contains only the information and assets necessary for that administration. For instance, Braille data would not be passed through to the online delivery system on a non-Braille test. A student who does not need a text-to-speech (TTS) accommodation would not have TTS assets bundled into her test, unnecessarily inflating the test file size. Similarly, much of the data that is a part of the repository format (prior versions, client comments, prior usage data, and performance statistics) would be stripped off in this conversion between the repository format and the delivery format.

Arguably, one of the biggest problems with the current state of the art regarding test and learning asset pools or repositories is transitioning from one service provider (vendor) to another. While issues during these transitions are sometimes caused by proprietary systems and sometimes by bad behavior, most often with the difficulty results from the incompatibility of item content, item formats, item process and item data. Presumably, technology standards could address this incompatibility.
3.2.21 Content Lifecycle

Requirement

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

Response

A “plug and play” exchange of assessment content should be able to take place during any moment within the content lifecycle, facilitated by technology standards. The most critical component for item portability is arguably the ability to exchange an item as a snapshot in time. However, an information archive about any individual item’s past may also exist. Some of this information would be related to the content lifecycle: prior versions, reviewer comments, fact-checking sources, and distractor rationales (a justification for the inclusion each incorrect answer choice, e.g., reflects a common misconception or represents a typical calculation error). The most expansive view of interoperability would involve moving this information archive along with the item itself when moving an item between item repositories, for instance.

While this type of portability may ultimately be desirable, it is recommended that types of item/test portability be prioritized to allow the most critical aspects of an interoperability standard to be established quickly and then expanded to less critical areas. In such a prioritization, the portability of content lifecycle data would most likely be lower than other types of portability, but state customers may wish to have access to this wealth of information to make informed choices about an item’s use. Assuming that vendors use similar processes and track similar data throughout the content lifecycle, interoperability standards can be expanded to include such information.

More investigation is necessary on the issue of item versioning, potentially in conjunction with the issue of carrying content lifecycle content with the item. If items are freely exchanged across systems and vendors, no single authoritative source for an item might exist. For instance, if a high-stakes item is released and made available to teachers within a less secure item bank, teachers may edit those items to serve particular needs. At that point, the performance statistics and the standards alignment data associated with that item may no longer be accurate.

Yet, portability of content development information is only one aspect of such a comprehensive set of lifecycle technology standards. Other aspects of the lifecycle, such as test forms or groups of items, reported data, formats, styles, and links to learning objects, must be considered to support comprehensive interoperability. For example, knowledge of the full revision history of a single test question does not help a test creator to determine whether the item should be used in a particular context. An item might be deemed inappropriate for a special needs assessment, for instance, unless performance on that item is related to specific content, skills, or processes needed by the student or if evidence
suggests that the associated skills measured by the item lead to improved performance by the student. In both cases, far more information will be required about the test question than its simple performance history, and only through a set of comprehensive technology standards will such exchangeable content—and responsible item usage in multiple contexts—be realized.

### 3.2.22 Interfaces and Services

**Requirement**

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?

**Response**

Today’s applications need to interoperate not only with browser-based clients but also with other applications. Various approaches to achieving interoperable Web applications are possible. Some of the most common and emerging of these are:

- **Web Service API:** A Web service API offers remote access to an application over a network, such as the Internet. Web-services are the de-facto technology for developing and integrating assessment management systems and applications that are capable of supporting cross-organizational processes. Web-services can be defined as self-describing, interoperable and reusable business components that can be published, dynamically combined and invoked through the Internet, even when they reside behind a firewall. Until recently, Web service APIs were integrated using Simple Object Access protocol-based (SOAP-based) Web services which could be complex to implement.

  There has been an emergence of the Ruby on Rails (ROR) open source web application framework utilizing web services based upon the Representational State Transfer (REST) software architecture style. Such services are particularly fast and simple to implement and consume when using the ROR Active Resource framework.

  Advantages of Web Services include:
  - Support for various representational states of a resource
  - Additional support for non-browser-based HTML clients is simplified by implementing the Model Controller view technique
  - Interoperability between applications and with new clients

  REST is not a protocol; it is an architectural style of large-scale networked software that takes advantage of the technologies and protocols of the World Wide Web. REST describes how distributed data objects, or resources, can be defined and addressed, stressing the easy
exchange of information and scalability. REST architecture describes six constraints applied to the architecture.

SOAP, on the other hand, is a protocol specification for exchanging structured information in the implementation of Web services in computer networks. It relies on XML as its message format and usually relies on other application layer protocols (RPC and HTTP) for message negotiation and transmission. SOAP can form the foundation layer of a Web services protocol stack, providing a basic messaging framework upon which Web services can be built.

To date, there has not been a large scale adoption of REST within the assessment industry but trends suggest it is likely to eventually replace SOAP. Ruby on Rails (also called Rails, or RoR) is an open source Web application framework written in the Ruby language that might be this replacement. Ruby on Rails is designed to help you develop and deploy Web applications faster and easier by making several assumptions when compared to other languages and frameworks. It is intended to be used with an Agile development methodology for rapid development.

- **Service-Oriented Computing (SOC).** SOC has been emerging as a new computing paradigm that utilizes services as fundamental elements for developing loosely coupled distributed applications/solutions. This paradigm has made inroads into mainstream business and enterprise deployments however its adoption in the assessment space has been almost negligible.

- **Agent-oriented platforms.** Agent-orientation is a new software engineering paradigm. An agent is a software module, just like objects and components. Just as with SOC agent oriented platforms are almost nonexistent in the assessment space.

One dilemma confronting the current state of the art around Web services is the question: How does an interested entity discover whether a Web service is already defined that might meet a current need? The same issue comes up in every communication situation. How do I find others to communicate with? In some venues, this need has been answered by a registry in which entities list their capabilities. In some instances, a directory, either maintained centrally or distributed, can answer the "who can I talk to" question. The National Circulation Interchange Protocol (NCIP) has long been discussing a policies directory to meet a similar need. The Australian Interlibrary Lending community has instituted a centralized directory of providers. The current situation in Web services is not very different. Various lists of Web services, categorized or not, exist. These are generally sponsored either by major interested parties such as Microsoft's University Description Discovery and Integration (UDDI) directory listings or by technical groups such as W3C's Web Services Activity.

---

Best Practices for Implementing and Standardizing Web Services

Small Interfaces
In general, simplicity and small interfaces are to be desired in the design of Web services. Small and simple interfaces are easier to document, test, maintain, and understand. Small and simple may also prove to be more robust. Length of time to implementation may be reduced if the service interface is small.

Document Service Interface
There are several different models involved in fully describing or documenting a Web service interface. Depending on the service in question, one, some, or all of the following models may be needed to adequately and unambiguously describe a service interface.

Information Model
The information model of a service includes the type of information that may be exchanged, the format in which that information can be presented, definitions of any terms used, and a consistent interpretation of strings and tokens used in the information exchange. Information models for exchanges within a single information domain may be able to leave some details unexplained. Information models for exchange across information ownership boundaries will need to be more exhaustive in all areas of the model. The information model should also include structure (syntax) and meaning (semantics) descriptions—which centers on the meaning of the information to be exchanged. Information may be intent as well as content. For example, a purchase order combines the description of the items being purchased and the fact that one party intends to purchase those items from another party. When information is exchanged across ownership boundaries, consistent interpretation of that information is essential. Such interpretations might be explicitly described in the information model. Existing industry-standard descriptions might be included in the information model by reference. The point is that the information itself, as well as any tokens representing information, must be interpreted consistently in order for interoperation to occur.

Behavior Model
The behavior model concerns itself with descriptions of actions, responses, and dependencies between actions included in the service. This description will include what behavior is expected of each party including: what (if any) response is required from each party when an action is invoked by the other, what dependencies might be involved (such as a sequence of actions that cannot meaningfully occur in a different order), what error conditions exist, and what to do upon encountering errors. A security-based service is an example needing a behavior model in that certain services (such as access to secure information) can only be supplied after the authentication of the party making the query is successful.

Action Model
The action model lists the actions possible within the service, as well as implied effects of the actions. For example, in a service managing a bank account, beyond knowing how to access the account and how to issue commands (service requests), to be successful one must also
understand that using the service may actually affect the state of the account (for example, withdrawing cash).

**Process Model**

The process model is not fully defined, and the need for it may vary depending on the service. It describes the temporal relationships between properties of actions and events interacting with the service. In other words, any required time-ordering of events and actions would fall under the Process Model.

**Description Limits**

Despite the models and the best intentions of those describing the service, there are limits to descriptions which can often lead to unstated assumptions. Another limit to description is what happens in certain multi-branched responses where private choice may be involved. For example, in the case that there is more than one response, this set of responses has to be converted into a single choice. This is a private choice that must be made by the consumer of the search information.

**Enable HTTP Caching**

A significant factor in the growth of the Web has been the ability for clients and intermediaries to efficiently cache transmitted content. Caches retain copies of content that can be reused by later requests instead of asking the service for it again. Caching can reduce bandwidth consumption and also decrease the latency of requests making the service appear faster and more responsive. Caching of content can also reduce load on the Web service by reducing unnecessary requests.

It is recommended that Web service implementations should make use of Hypertext Transfer Protocol (HTTP) caching mechanisms to integrate into the existing Web infrastructure. Content accessed via HTTP GET (a command to “get” information) is particularly amenable to caching and the service provider can take advantage of a number of standard HTTP headers to assist caching intermediaries.

Most caching systems will not cache HTTP POST requests since this method is designed to effect a change of state on the origin server. If the Web service implementer has a free choice of HTTP methods then GET should be preferred for requests that do not modify the state of the server (idempotent requests). Most query and search services fall into this category, whereas a login or item checkout service typically will not. Enabling query or search responses to be cached can drastically reduce the load on back-end systems such as databases.
Following is a list of best practices around HTTP caching behavior as it applies to Web services.

- **Prefer HTTP GET for Requests That Do Not Modify the State of the Service.** POST requests will not usually be cached by intermediaries so use GET where the request will not modify the state of the service. A particularly efficient mechanism provided by HTTP is conditional GET. This mechanism allows clients to issue a single GET request to the server and receive either a short response indicating that the content has not changed since the last request or a copy of the new content if it has. Conditional GETs are controlled by two pairs of HTTP headers. The first pair is ETag which is sent by the server and If-None-Match which is sent by the client. The content of the ETag header is a string enclosed in quotes that uniquely identifies the content being sent in the response. This string can be stored by clients who may then use it when performing subsequent requests. If the server has generated an E Tag for some content a client may at a later date re-request the content and specify an If-None-Match header containing the value of the ETag previously supplied. If the server determines that the content to be returned would have the same ETag it can return a 304 response code with no content. If the content is different, the server should return a 200 response code and include the new content.

- **Responses Should Include an ETag HTTP Header.** Supplying an ETag response header enables clients and caches to save bandwidth and processing time by using conditional GET requests with the If-None-Match request header.

- **Responses Should Include a Last-Modified HTTP Header.** Supplying a Last-Modified response header enables clients and caches to save bandwidth and processing time by using conditional GET requests with the If-Modified-Since request header.

- **Prefer Conditional GET Headers over the Expires HTTP Header.** To be effective the Expires header requires clocks to be synchronized between client and server. Prefer ETag and Last-Modified headers instead since all comparisons are performed by the server alone.

- **Avoid User Specific Information in URLs.** Using URLs without user identifiers for content that is not user specific increases the chance that a particular request can be satisfied via a cache. Many Web services provide content that is not user specific such as search results or data conversions. This kind of content is ideal for caching by intermediaries such as proxy caches which may be serving entire institutions or Internet service providers. However if user specific information such as session identifiers are included in the URL the efficiency of the caches is drastically reduced. The cache cannot assume that the content accessed by slightly different URLs is the same and therefore must forward many more requests to the origin Web service. If possible, Web services should avoid user information in URLs for content that is suitable for multiple users.
Filter User Input

Because all web-based applications are subject to security threats, all user input to a Web service needs to be intercepted and vetted in some sense before being allowed to initiate the service or consume its results. The user may be a legitimate consumer of the service, or it may be malware or some other security issue threatening the integrity of the operation. Such attention to verifying the integrity of the input will go a long way toward protecting the Web service from unintended use.

Reuse Existing Output Formats

When designing a Web service careful thought must be given to the format of the service requests and responses. These formats must be machine processable but still expressive enough to convey all the information necessary for the service to function correctly. Typically XML is used to structure output formats, although there are many other options depending on the specific situation. Often the needs of a particular service can be met by reusing an existing well known format. It is recommended that Web service designers look for and use an existing format rather than invent a new one specifically for their service.

Document Output Formats

Once developed, Web services need to be described in a way that is understandable to humans and directly accessible to machines. There are several commonly-used methods for expressing/codifying a Web service. A brief description of some of these follows, including some differentiating features of the various methods.

- **DTD (Document Type Definition).** This method of documenting output formats depends on declarations of elements and attributes lists. An element declaration names the allowable set of elements within the document. Part of this is to also specify whether declared elements may be contained within each element. Rules about how such inclusions occur are also stated. Attribute-list declarations name the allowable set of attributes for each declared element, including the type of each attribute value, or an explicit set of valid values. A DTD is separate from the XML document it is meant to "explain" or define. A Document Type declaration in the target document will associate it with its relevant DTD. There is a DTD syntax that must be adhered to in order to produce a valid DTD.

  Among the strengths of the DTD is that it is used widespread in XML tools and that DTD is included in the XML 1.0 standard.

  Among the weakness of the DTD is that DTD does not support some important XML features like namespaces, and that there are some limits to what DTD can express—it is more limited than XML in general

- **XML Schema.** XML Schema, also referred to as XML Schema Definition (XSD), has achieved Recommendation status within the W3C. It is a schema language that also supports validation, collecting information about the document structure during the act of validation. XML Schema is associated with Microsoft support and is particularly apt to support object oriented programming.
Among the strengths of the XML Schema are the Microsoft connection, the easy linkage with object oriented programming, and the W3C Recommendation status.

Among the weaknesses of the XML Schema are the Microsoft connection, which leads to a charge of lack of openness, and some restrictions arising from data type dependencies on other W3C specifications.

- **Resource Description Framework (RDF).** RDF refers to a group of specifications for a metadata model that is often implemented in XML. RDF specifications are maintained by the World Wide Web Consortium (W3C).

- **RDF Schema.** It is an extension of RDF. It can describe groups of related resources and their interrelationships. RDF Schema descriptions are written in RDF.

  Among the strengths of the RDF Schema is the fact that it is a favored method for Semantic Web and knowledge management applications and it is part of a set of specifications maintained by the W3C.

  Among the weaknesses of the RDF Schema are its very association with the Semantic Web and the slow uptake of this method by very simple Web services.

- **Relax NG (Regular Language for XML Next Generation).** RELAX NG is a schema language for XML that specifies a pattern for the structure of the XML document. There is also a non-XML syntax of Relax NG. An OASIS technical committee maintains the specification. Relax NG is also part of the ISO Document Schema Definition Languages (DSDL) standard (ISO/IEC 19757).

  Among the strengths of Relax NG are the simplicity of use and the leverage of nesting Russian-doll structure as well as support for data typing, regular expressions, and namespaces. It also supports interleaving to allow flexibility in specifying elements.

  Among the weaknesses of Relax NG are its difficulty in handling recursive elements in its nested version, and the fact that the W3C XML Schema specifications are better known and more widely adopted, though that is slowly changing. Relax NG also lacks the range of data types supported by XML Schema.

- **DSD (Document Structure Description).** DSD is a schema language for XML, meaning that it is a language used to describe valid XML documents. DSD is an alternative to DTD and to the XML Schema. It is reputed to be an extremely flexible easy to use language. But its differences with and separateness from the W3C set of specifications can be viewed as a barrier to common use.

**Security**

Other National Information Standards Organizations (NISO) groups have worked on security in terms of authentication methods. Of particular appropriateness here is the NISO Metasearch Initiative’s paper on authentication methods. For Web services in general, secure transport in a Web environment (usually meaning https) plus attention to authentication methods will help verify that the Web service is used for its intended purpose by its intended audience while safeguarding any personal information it may use.
Throttling
Web applications can be swamped by too much traffic or demands that are too heavy. Just as user input is inspected and accepted with caution, so too many queries for the service may have to be intercepted, examined, and treated either in separate streams depending on the results to be delivered, or meted out in a measured way so as not to overwhelm the resources of the service.

3.2.23 Internal Transparency and Ease of Use

Requirement
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 commenters to consider the extent to which various technology standards possess native “human readability” and comprehensibility.

Response
Striving to have standards be “human readable” is a noble goal and one which should be emphasized in the development of assessment technology standards. Any efforts to increase the breadth and depth of stakeholder involvement are beneficial in the long run, as they help verify that the pool from which innovation draws is kept large and diverse. This is true both for application of the standards to develop conformant content and for evolving the standards to meet future needs. That said, some standards are inherently complex and efforts to make them “readable” could potentially risk a trend toward oversimplification of the underlying frameworks behind the standards. In these cases, the solution is not to compromise any necessary complexity but rather to rely on accompanying text that explains the standards in easier to understand language, together with additional support in the form of examples, FAQ’s, user forums, training, and technical support. It is important that any such supporting materials and supports be under continual review so as to be kept in sync with the standards themselves as they evolve.

An additional consideration when addressing transparency and ease of use with respect to standards is the “human readability” of content itself. The most prevalent standard used today to exchange, share and store data is XML. A significant number of application programming interfaces (API’s) and XML-based development languages have been developed over the past several years. All browsers contain built-in parsers that can read and process XML and it provides the flexibility necessary to adapt to a broad array of uses. As noted in the requirement, it is readable by a technical user but not by the less sophisticated user. An alternative standard to XML is JSON (JavaScript Object Notation). It was derived from the JavaScript programming language and is language independent with parsers available for a
variety of programming languages and most modern browsers support native JSON encoding/decoding.

In addition to being easier to write and interpret than XML, encoded data are typically smaller in size because of the nature of XML tags. While the usage of XML today far outpaces that of JSON, the technical debate on the advantages and disadvantages of each will continue over the next several years.

Experience regarding the tagging of test items, and the attempts to make such tagging human readable, may provide insight to the US Department of Education’s query. Our first experience with item banks and the need to delineate metadata in a human readable way was in the early 1980’s and has evolved ever since. At first, it was seen as a key requirement that the unique item coding be human readable for at least key aspects such as subject, grade and content classification. Then additional needs to include scoring keys, content clusters, and readability levels extended the size of this “human readable” code. Before long the bottom of a test item review card was filled with this ”human readable” code that was so complicated that no human actually read it. With technology becoming more automated a machine readable code should be used and, as alluded to in previous paragraphs of this section, the technology itself can make it human readable on demand or when required by the user.

### 3.2.24 Discovery and Search

**Requirement**

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?

**Response**

Existing content interoperability standards currently do not strictly specify whether assessment metadata provides “searchable tags” that can be exploited by search engines. However, there are commercial, customized search engines (e.g., netTrekker) that are able to access relevant, valid, and reliable assessment materials by performing matches either within item metadata and/or content, albeit rather crudely.

There is no single technology or standard today that addresses all of the different aspects of assessments upon which a search might be conducted, such as assessment types (e.g., text-based ELA multiple choice, Flash-based drag-and-drop algebraic constructed response), item technical parameters (e.g., p-values, point biserials), results (e.g., samples of student responses), and rights and permissions (e.g., freely available, restricted distribution). Neither is there a standard for types of search engines or search methods to be used; some search engines do direct keyword matching while others match based on semantic features. However search engines do typically provide APIs to their search technology, opening the
possibility for more systematic approach toward tagging of assessments to facilitate discovery and search. One suggested approach would be to profile several different existing standards (e.g., IMS QTI, IEEE Learning Object Metadata, SIF Assessment Objects) to create a meta-taxonomy that define a schema against which assessment assets can be tagged. In this way a comprehensive set of search-relevant assessment features could be defined for broad industry adoption across multiple standards. Such an approach overlaps conceptually with the goal of the Common Education Data Standards (CEDS), as discussed in Section 3.2.25 Metadata.

An additional concern in standardizing discovery and search is the lack of a comprehensive assessment lifecycle perspective; individual standards have evolved with foci on limited aspects of the assessment lifecycle, such as specification of technology-administered items or exchange of data. If a comprehensive set of technology standards is to support consistent searching, it is necessary to define what aspects of the search are currently applicable, which metadata are relevant and should be exposed, and the level of granularity (e.g., item, testlet, and test) at which the search should operate. As such, representation of a broad set of testing perspectives, such as test forms, item pools, presentation considerations, and “mixed-mode” (paper and/or online) assessments, is necessary. Thus a comprehensive lifecycle approach to technology standards regarding interoperability would facilitate standardization of discovery and search.

3.2.25 Metadata

**Requirement**

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 semi-structured data incorporated within assessment technology standards?

**Response**

In general, all technology standards include metadata, and assessment standards are no exception. Metadata can be used for many purposes, such as describing data (data about data), extending meaning, defining structure or relationships, and identifying rules. Markup languages such as XML are excellent approaches for representing metadata in a standard or system.

For assessment standards, metadata can also be used for many purposes. As an example, current assessment standards provide methods to identify learning standards that are “linked” to items and assessments. These existing structures are well suited or could be easily extended to meet the needs of the assessment markets today. The Common Core State Standards may extend these structures if multiple relationships between multiple standards structures are required. In other words, if an item has to be simultaneously linked to multiple
taxonomies such as a state taxonomy or core taxonomy, then these multi-dimensional relationships may require some extensions to current standards. In addition, industry standards for learning standards themselves (such as global and unique identifiers, commonly defined structures and levels across subjects and grades, etc.) are somewhat missing in today’s market.

A recent effort to formalize and organize common data standards by a national consortium is highly germane to the topic of metadata standards. The Common Education Data Standards (CEDS) consortium11 is creating definitions of and formats for a subset of key K-12 data elements and key K-12-to postsecondary transition data elements, with the goal of voluntary adoption and wide acceptance within and across states in both K-12 and postsecondary sectors. Obviously coordination of efforts with this initiative could greatly serve the effort to define metadata standards within assessment technology standards.

The answer to the broader question of what types of metadata are permitted in assessment technology standards is that the standard and its use of markup language should not limit the types of metadata that can be included in the technology standard. The limit should only be what has been defined in the standard for any given release of the standard. In addition, most technology standards allow for custom extensions to accommodate unique requirements (but keep in mind custom extensions potentially reduce the level of interoperability).

Metadata that defines unstructured or semi-structured data elements are typically used to support extending, sorting, searching, queries, etc. For example, an unstructured data attribute in a standard may include a text passage. The metadata about the passage may include attributes such as the author, copyright information, key words dates, language, semantic density, and readability. Another example may be searchable attributes about images or videos that are otherwise not inherently searchable.

Broadly speaking, and at the risk of over generalizing, metadata typically provides meaning and context to information that may otherwise be difficult to understand on its own. In addition, metadata are often used by automated systems or software programs to perform operations on information, such as search, aligning to standards, or matching of content. Therefore, metadata ideally should be defined precisely so that logic and rules (and understanding) can be universally applied and thus broadly consumed.

Perhaps one way to understand how technology standards can help with precisely defining metadata, and solve a burden that has long perplexed assessment experts as well as other educators, is to consider the current paradigm in standards alignment—one for which the CCSS alone are insufficient.. Currently each state has its own set of content standards. Similarly, there are such classification taxonomies that underlie instructional modules.

---

11 The consortium consists of the Council of Chief State School Officers (CCSSO), the State Higher Education Executive Officers (SHEEO), the SIF Association, the Postsecondary Electronic Standards Council, the National Center for Education Statistics (NCES) of the U.S. Department of Education, and the Data Quality Campaign. (http://www.commoneddatastandards.org/).
instructional content, intelligent tutoring systems, etc. Each of these instructional assets will have metadata describing what is being taught, measured, or tailored as a function of its underlying classification taxonomy. Hence, the goal of technology standards to make these systems interoperable is separate from (but dependent upon) having a common set of metadata and/or a common taxonomy. Rather, the goal of technology standards is to find a way in which each various set of metadata could be translated to another (perhaps through a common protocol) without loss of fidelity. Arguably, an example of this could be the PDF format currently used by Adobe for text conversion. Before copyright lawsuits, companies like Microsoft, WordPerfect, Quattro and Lotus routinely allowed for the conversion of proprietary formats from one system to another. This might be an example of the type of interoperability such technology standards should seek to support.

### 3.2.26 Recommendation, Rating, and Review

#### Requirement

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?

#### Response

Currently, professionally developed items are reviewed for clarity, accuracy, format and efficacy at almost every step of the development lifecycle. This takes the form of content reviews, fact checking, formatting and/or typographical reviews, reviews for the accuracy of the correct answer, fairness reviews, sensitivity reviews, anti-bias reviews, public reviews. These reviews include multiple stages of data and information reviews from field testing to pilot testing to data reviews prior to live testing. Almost all of these reviews are facilitated by standards and use technology in various ways. As such, when considering comprehensive technology standards it is important to understand the state of the art of the current system of review and incorporate the best aspects of this system as being supported or enabled by the next generation technology standards.

While it may be debatable if classroom and instructional assessments are “professionally developed” and perhaps the review of such instructionally related content is not as rigorous, the principles of review and the collection and maintenance of data and information associated with the items could be enhanced and maintained if the technology systems that supported them (including the associated technology standards) facilitated such action. As such, these instructionally linked items and learning assets could also benefit from technology standards supporting or enhancing reviews.

Under the goals of the current Common Core State Standards initiatives and/or in support of the Race to the Top applications, one type of review that could be exploited and supported by
a new generation of technology standards would be “crowd sourcing”. Take for example the iTunes help facility established by Apple. Apple incents other users to answer questions and provide technical support by utilizing a technology system that only rewards suggestions that other users find helpful. This is seemingly relatively easy technology to build into the design of a system but would be very difficult to add after the fact. As such, and if crowd sourcing reviews (or review of the reviews or reviewers) is to benefit from comprehensive technology standards, this potential solution will need to be considered and incorporated into the technology standards prior to them being implemented.

Another example where there is clear value in being able to annotate items with ratings, reviews, and/or recommendations, for example to support effective use of distributed item banks by individuals such as teachers, curriculum directors, and parents may be found in the learning object tagging schema, such as the IEEE Learning Object Metadata12, which provide simple extension capabilities that could be used to represent ratings, reviews, and recommendations (See Section 3.2.21 Content Lifecycle for more information. Whether or not existing assessment technology standards specifically provide for tagging of items with such information, this is a feature that certainly can be addressed as part of the metadata issue as discussed in the previous section with little or no additional technical challenge.

3.2.27 Content and Media Diversity

Requirement

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.

Response

An evaluation of Common Core State Standards in combination with an assessment engineering approach could lead to a set of item types unlike those in common usage but which may be more appropriate ways to measure certain constructs. These item types, supported by an enhanced set of interoperability standards, could then be authored, delivered, and scored by any system. For instance, educators might agree that a range of item types using equation editors, graphing tools, and geometry construction tools would be appropriate for assessing certain math standards. Such item types, if constructed following a more generalized interoperability standard, could be used to map out an evolution path for assessment systems to support an ever growing range of item types. These items would presumably measure the salient aspects of the next education reform effort and assess problem solving, critical analysis, higher-order cognitive skills in accordance with the college and career readiness as well as international benchmarking. However, this approach requires anticipating item types and their specified uses early on in the specification design process.

12 http://ieeeltsc.org/
As described in 3.2.8, with a different approach item content does not have to conform to an XML schema but such items would need to conform to a standardized communication protocol so that these items could be summoned by the test delivery system and so that the items could send back student performance data. Any item using this communication protocol would be able to be delivered on any system that adheres to this data exchange protocol. An advantage of such a method is that it encourages innovation and allows a diverse range of item types to exist without needing to extend the current sets of standards. However, some disadvantages also exist as described in 3.2.8 and summarized again here:

- Less transparency into the scoring logic and less flexibility around re-scoring;
- Use of a technology-specific communication protocol that may not generalize to all devices;
- Items potentially unavailable to system-based accommodations (font re-sizing, high contrast filters, and text-to-speech);
- System-provided helper tools—highlighter, ruler, calculator, underliner, e-pencil, note pad, answer-eliminator—unavailable within self-contained items, unless included within self-contained item (with a potentially different user interface);
- Higher average file size due to unavailability of an asset sharing strategy;

No decrease in item production costs (since items tend to get produced in a unique fashion with this approach rather than building the capability into the system once to support all items of a given item type). No decrease in item production costs (since items tend to get produced in a unique fashion with this approach rather than building the capability into the system once to support all items of a given item type).

While simulations could be accommodated through either of these strategies, some other types of assessment are more difficult to conform to interoperability standards. Games and virtual worlds tend to be system-dependent, although the output of assessment results could still conform to standards for assessment reporting.

Mobile devices should be considered when evaluating how easily a standard can support diverse delivery platforms. The most flexible type of standard for representing item content across a varied range of delivery devices is currently an XML-based standard, rather than an XHTML standard. An XML standard that separates content from formatting allows the test delivery platform to format dynamically depending on the device. The XML-based source for the item content would be the same regardless of whether the item is being delivered in print, on a widescreen monitor, or on a mobile device. However, the technology called into play to control the formatting and the particular format that would be chosen could be specific to the device. Vendors who deliver current print and online services often deploy this type of solution, so that different formats can be supported for print and online. This is more difficult to accomplish when the item source intermingles content and formatting.
3.2.28 Accessibility

Requirement

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?

Response

General Technology Standards and Guidelines

Various standards and guidelines exist to improve technology access for individuals with disabilities. Outside of education, the most notable among these are the Web Content Accessibility Guidelines and Section 508 of the Rehabilitation Act.13

Web Content Accessibility Guidelines (WCAG 2.0)14, created by the W3C’s Web Accessibility Initiative (WAI), provide the most comprehensive, widely used guidelines for creating and evaluating accessible Web content. WCAG 2.0 success criteria provide technology-agnostic, testable conditions and thus facilitate determinations of compliance. WCAG is one of a set of three related recommendations from WAI, the other two being Authoring Tool Accessibility Guidelines (ATAG) which addresses Web content authoring tools and User Agent Accessibility Guidelines (UAAG) which addresses Web browsers and media players, including some aspects of assistive technologies.

Section 508 of the Rehabilitation Act (Section 508 Standards)15 was enacted to maintain that all electronic and information technologies developed, procured, maintained, or used by the federal government are accessible to individuals with disabilities. Increasingly its provisions have been adopted by non-governmental agencies, such as universities, school districts, and state education agencies. While not as stringent as WCAG 2.0 with respect to defining Web accessibility, Section 508 Standards are broader and apply to software, telecommunications, video and multimedia, kiosks, and computers.

13 Additional software accessibility guidelines, such as those from Adobe™, Apple™, IBM™ and Microsoft™, are not covered in this document.
14 http://w3c.org/WAI/
Education Technology Standards and Guidelines

Successful use of educational technologies brings specific demands to students with disabilities. For example, the population tends to be younger, resources within schools are often quite limited, and learning itself is a unique situation that places additional burdens on technology’s role in supporting pedagogical needs. Within education the most notable standards and guidelines for accessible technologies are AccessForAll, NIMAS, and the Accessible Digital Media Guidelines. An additional set of accessible educational technology guidelines, Universal Design for Learning, is described below in the section Universal Design Guidelines.

IMS Global Learning Consortium (IMS Global) AccessForAll\(^{16}\) is a learning object metadata specification that enables identifying resources that match a user’s stated preference or needs. It works in conjunction with a sub-schema for the IMS Learning Information Package (LIP) specification that defines a means to specify accessibility preferences and learning accommodations. Thus, learning management systems that utilize the LIP specification have a means for matching accessibility features to individual users. These preferences are designed to support not only learners with disabilities, but any user in a challenging situation, such as in a noisy environment or while multitasking. Accompanying guidelines for use of accessible technology are available as well.

The National Instructional Materials Accessibility Standard (NIMAS)\(^{17}\) is a technical standard used by instructional materials publishers to produce XML-based source files for use in creating multiple specialized formats—including HTML, digital audio books, and Braille—for students with print disabilities. NIMAS is based on the DAISY Consortium's ANSI/NISO 239.86 file format standard. The NIMAS standard was developed by the National Center for Accessing the General Curriculum (NCAC) at the Center for Applied Special Technology (CAST) under support from the Office of Special Education Programs (OSEP) at the U.S. Department of Education, pursuant to sections 612(a)(23)(A) and 674(e)(4) of the Individuals with Disabilities Education Act (IDEA). By law, NIMAS source files must be provided by publishers for elementary and secondary instructional materials such as textbooks published on or after 7/19/06, as specified under Parts B and D of the IDEA, and the Chafee Amendment of 1996 of the Copyright Act. As a result, school districts and state education agencies are able to provide students with print disabilities access to instructional content previously unavailable or not conveniently available. It is important to note that since testing materials are excluded from the Chafee Amendment, NIMAS is not applicable to assessments.

The National Center for Accessible Media (NCAM) Accessible Digital Media Guidelines\(^{18}\), originally developed in 2000 and revised for the second time in 2006, provide comprehensive guidelines to developers of educational software and digital publications. The guidelines place high emphasis on consideration of students’ individual differences due to age,

\(^{16}\)http://www.imsglobal.org/accessibility/
\(^{17}\)National Center on Accessible Instructional Materials; http://aim.cast.org/
computer experience, and nature of disability, as well as preserving pedagogical intention and the benefits of multimodal learning.

**Educational Assessment Technology Standards and Guidelines**

Accessibility of technology-based educational assessments, while in many ways a subset of general educational technology accessibility discussion, brings with it specific concerns, most notably with respect to security, reliability, settings, and preservation of intended constructs. Ideally, accessibility within testing should be approached from a validity perspective; an accessible test—whether for students with or without disabilities—is one that minimizes the introduction of construct-irrelevant factors through the media used for testing.

Accommodations—the general approach toward reducing barriers—are a necessary first step in providing students with supports that minimize influence of construct-irrelevant factors on their scores. However, by their very nature, testing accommodations are post-hoc solutions, and as such risk compromising the validity of interpretation of test scores. Ongoing research is necessary over time to verify that students are provided with the most appropriate user interfaces, including when necessary specific accommodations and/or assistive technologies. Approaches such as defining general user experience guidelines are necessary, as is an understanding that accessibility and usability are part of a continuum.

Within educational assessment, one set of draft standards—the Accessible Portable Item Profile—and one set of guidelines—the Test Access Guidelines—are most notable. An additional set of accessible testing guidelines, The Universal Design for Computer-Based Testing Guidelines, is described below in the section Universal Design Guidelines.

In December 2010, IMS Global released for public comment a draft version of the Accessible Portable Item Profile (APIP) specification developed by a consortium led by the Minnesota State Department of Education through funding from the US Department of Education Enhanced Assessment Grant program. APIP leverages three existing IMS interoperability standards—Question and Testing Interoperability (QTI), Common Cartridge, and AccessForAll—to develop an integrated set of tags, along with descriptions of expected behaviors. These tags can be applied to standardize the interoperability and accessibility of test items and accommodations across various testing platforms.

American Printing House for the Blind (APH) Test Access guidelines provides suggestions to accommodate students with disabilities, particularly blind and low-vision students, during

---


technology-based testing. The document also describes the educational impact of visual disabilities and provides a general overview of accommodations for blind and low vision students.

**Universal Design Guidelines**

While the aforementioned standards and guidelines have all contributed significantly to a trend toward improved access to the general curriculum for students with disabilities, a critical problem remains: unless accessibility is considered from the start, it is difficult to effectively retrofit later. This difficulty manifests as limitations in effectiveness, compromises to original design, and/or high expense. Such is the central tenet of universal design\(^{21}\), which states that flexibility of use should be an initial design consideration. The goal is to minimize assumptions about individuals’ abilities, challenges, and usage patterns that are orthogonal to the task at hand, and hence to remove unnecessary barriers to use and participation.

Successful use of technology during learning requires not only access to materials and methods, but also to the learning potential that the materials embody. To address accessibility of digital learning materials, curricula, standards, and assessments, the concept of universal design has been extended to Universal Design for Learning (UDL)\(^ {22}\) by considering cognitive and pedagogic aspects of learning. As such, it address a more comprehensive set of barriers that students—even those without disabilities—encounter during learning, as well an approach for successfully leveraging the pedagogical potential of digital technologies and new media. High-level guidelines\(^ {23}\) for applying UDL principles to education have been developed but would be a challenge to standardize due to their heavy pedagogical and cognitive consideration.

The Universal Design for Computer-Based Testing (UD-CBT) Guidelines\(^ {24}\) brings together a validity-based approach toward test accessibility with a UDL-based approach to understanding cognitive and pedagogical interface considerations to provide recommendations for technology-based testing. The UD-CBT Guidelines provide a highly technical, detailed approach toward use of technology to reduce construct-irrelevant variance, while minimizing the introduction of new barriers through inappropriate use of technology. The guidelines are organized by the individual components of computer-based test items, such as use of interactive media and constructed response math items, as a function of student processing considerations, namely perceptual, linguistic, cognitive, motoric, executive, and affective. As such, they provide for deep consideration of the accessibility of technology-based testing, and for all students, including those with disabilities.

---

22 National Center on Universal Design for Learning, Center for Applied Special Technology (CAST); [http://www.udlcenter.org/aboutudl/](http://www.udlcenter.org/aboutudl/)
3.2.29 English Learners

Requirement

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? Questions about process and IP for technology standards development begin in the next section.

Response

Because English learners have unique learning needs, technology has great potential to improve and perhaps even accelerate learning by adapting and personalizing needed accommodations to optimize the assessment experience. Two important ways this can be accomplished include:

1. Two learning paths—for English learners, their learning of academic content requires that they first learn enough English so they can meaningfully engage in instruction in English. Therefore, understanding the educational progress of English learners and planning their instruction requires that we accurately assess and interpret results for English proficiency, academic performance, and the relation between their proficiency and performance.

2. Linguistic accommodations—when English learners engage in academic content in English, they often need linguistic accommodations. The accommodations that best meet English learners' needs vary dramatically from not only student to student but for any one student over time. As students progress in their English proficiency, the accommodations they need for learning and demonstrating their academic performance changes. Furthermore, the linguistic accommodations that vary over time for a particular student may differ in substantive ways for the different content areas.

It is important that the use of technology for English learners be accompanied by careful evaluation to confirm that the use of technology allows us to improve our ability to accurately infer information about English learners. For scores to accurately reflect students' English proficiency and academic performance and for students, parents, educators, and administrators to accurately infer student educational progress from these scores, threats to the validity of the inferences must be evaluated and ultimately reduced when technology is used. Threats can come in the form of construct irrelevant variance, or factors that get in the way of accurate measurement of a construct (that is the thing being measured and construct underrepresentation, or when the assessment captures only some of what is intended to be measured (an often heard criticism of assessments under NCLB). An example of construct irrelevant variance, when a fourth grade English learner does not understand the directions for completing a mathematics assessment, the inability for that student to understand how to complete the mathematics assessment is a factor that impedes that student's ability to show her mathematics knowledge and results in construct irrelevant variance manifested as underperformance. As an example of construct underrepresentation,
an English learner who uses a dictionary with definitions of science terms to complete the science test and is not assessed on his or her ability to know those content-relevant science terms may not be assessed on the entire intended measure. Therefore, that student is not being fully assessed on the science content and is likely to overperform.

To prevent or minimize construct irrelevant variance and construct underrepresentation, technology standards that are targeted at the unique needs of English learners are needed. Technology standards and guidelines for accommodating English learners may need to exist at specific levels of individual organizations, such as schools and districts, certification bodies, and testing agencies. While some of the requirements of English learners during testing may be supported by providing solutions designed for students with disabilities for which standards and guidelines do exist (that is, by borrowing some of the lessons learned in these other special populations), English learners have unique learning needs that require presentation and response options not provided for students with disabilities, such as dual language options and on-the-fly translation options.

Furthermore, there are more likely to be interaction effects between content and individual student characteristics such that presentation and response modes and accommodations may vary from item to item for a given student. Certainly to the extent that universal design principles have been applied during item and test design and construction, the flexibility needed to support English learners may be more likely present than, for example, in a test designed only to provide specific accommodations for specific disabilities. Clearly the long term answer is to help ascertain that the special needs of English learners is researched and articulated clearly and directly addressed in technology standards. Note that the current universal design standards developed for special needs students evolved over a very long time and, as such, such a comprehensive set of similar standards for English learners will take some time and investment. Ideally, individual characteristics of the students beyond a binary EL/not EL classification would help in choosing appropriate modes of testing.

An important consideration when addressing technology access for English learners is the likelihood of concurrence of high-incidence disabilities, such as specific learning disabilities as well as other more traditional “at risk” indicators. Therefore, the set of presentation and response modes and accommodations appropriate for such a student would require even more forethought.

Recommended technology standards in support of English learners could be organized in three areas:

1. The assessment of English proficiency
2. The assessment of academic content with linguistic accommodations
3. Methods for combining data from both types of assessments, so decisions about instruction and assessment can be fully informed
Under each of these areas, recommendations for technology standards are made as well as ideas for technology use with English learners.

**Technology Standards for Assessing English Proficiency**

Before English learners demonstrate sufficient English language proficiency to access academic content in English, their educational needs and learning focuses on listening, speaking, reading, and writing in English. Technology standards that could be considered in this regard include the following:

**Recommended Technology Standard 1: Interoperability of Standardized Test administration Procedures and Training**

The test administration procedures and training components need to be designed with interoperability in mind so that they can be exchanged modularly across various vendors’ learning management systems and test delivery platforms.

For students to accurately demonstrate their ability to listen, speak, read and write in English using technology, they will need to understand the directions and will need to be able to use the technology. Therefore, the test directions, instructions, and training components—and any supports specific to English learners—must be transferred along with the assessment items and tasks.

**Recommended Technology Standard 2: Maintenance of Historical Information**

Databases and systems should be developed so that they can incorporate historical information about students that may have come from different systems.

Since student progress in learning English is important, systems that house student information should be equipped to include historical information, so that student progress can be evaluated and documented. For example, student speech samples collected with technology may be what is used over time to evaluate a student’s progress in learning to speak English.

As technology for capturing student’s speech improves over time, the samples for a particular student will likely be generated from different technology systems. The databases and systems that house a student’s data and that will be used to transfer a student’s data across locations will need to be flexible enough to include information from those different systems.
Technology Idea 1: Benefits for Use of Computer Adaptive Testing

There are potential benefits for the assessment of listening, speaking, reading and writing in English using computer adaptive tests (CAT).

English learners will not learn English like other students learn academic content, in a grade-by-grade manner that has a predictive pattern. Instead, English learners learn English at their individual rates. Therefore, technology can facilitate this learning, wherever it is on the proficiency continuum. To improve our ability to most efficiently measure each student’s English proficiency, computer adaptive tests (CATs) can target student’s English proficiency and assess the student at and around that student’s current level without requiring the student to attempt test questions that are too difficult or too easy. The CAT method is advantageous from an efficiency standpoint for the measure of English proficiency for several reasons:

- The range of English proficiency is extensive, therefore, an assessment that spans the full range will inevitably require students to complete test questions that are too easy and those that are too difficult.
- Students can progress from level to level rapidly, so it will be difficult to know in advance of an assessment where that student is likely to perform.
- If the method targets the assessment to the student’s current level and the student’s level changes rapidly, the student will not experience the same questions or stimuli very often, thereby maximizing use of the information obtained and minimizing item exposure risks.

Technology Standards for Assessing Academic Performance with Linguistic Accommodations

When English learners demonstrate sufficient English language proficiency to access academic content in English, their educational needs and learning typically require linguistic accommodations so that the validity of measures of student’s academic performance are not threatened by construct irrelevant variance and construct underrepresentation.

Recommended Technology Standard 3: Linguistic Accommodations Should Not Require Students to Ask for the Accommodation

In many states, students use paper assessments and linguistic accommodations specific to a student’s needs that often require that the student ask for the accommodations. Some English learners might be too shy or self-conscious to ask for the help, so they do not receive needed accommodations. Students who are provided with the option of translating words and phrases that are not essential to the academic content should be offered the translation in such a way that they do not have to rely on a teacher or test administrator to deliver the option.
**Recommended Technology Standard 4: Varying Levels of Support within an Assessment Item, Task, or Activity**

The linguistic accommodations offered to English learners should offer varying levels of support at the item, task, or activity level and not be the same across all parts of an assessment.

The linguistic accommodation that students need depends on a complex set of factors, including when the student came to the United States, how long the student has been in the US, language in which the student learned the academic content, etc. For example, a student may come to a state from a Spanish-speaking country mid-year and learn the initial grade 7 mathematics content in Spanish and the latter grade 7 mathematics content in English. That student might need more linguistic accommodations for items, tasks, or activities assessing the initial grade 7 content and fewer linguistic accommodations for items assessing latter content. By offering accommodations that can vary item by item, standards will offer more flexibility in meeting the linguistic needs of students and will help maximize the validity of the assessments.

An example of how this standard has been applied is found in Texas. There are more than 800,000 English learners in Texas, which is an increase from just under 600,000 English learners in the year 2000. Over 120 languages are spoken as the primary language by these Texas English learners, though the primary language is Spanish for over 90% of Texas English learners. English learners make up approximately 17% of the total students in Texas, which means that about one in six students is an English learner. Pearson, in collaboration with the Texas Education Agency has been aggressive in its use of technology with English learners, as the reading test of English is offered to the over 500,000 English learners online only (paper versions are provided in rare circumstances for special accommodations). As Texas moves to its new testing program, students who take the academic assessments with linguistic accommodations will test online. The linguistic accommodations will offer students the flexibility of varying support item by item. See Figures 1 and 2 as an example of a sample test item. In the item, each English learner has the choice of clicking on a word with a dotted line (see Figure 1). By clicking on the word, English learners can have the word read aloud, translated to her primary language (e.g., Hindi, Spanish, Urdu, and Vietnamese). See Figure 2. The student may also choose to have the item stem, the question, or the responses read aloud. The read-aloud is not for the full item, which would limit the flexibility for which the student has to use the accommodation for only the part of the item needed.
**Figure 1.** Sample mathematics item developed by Pearson to allow flexible linguistic accommodations at the item level

A student designs an investigation of plants in a meadow. The student creates a diagram to represent the plots where the investigation will take place.

What is the total area, in square meters, of Plot 3?

- A 9.00 m²
- B 16.0 m²
- C 36.0 m²
- D 60.0 m²

**Figure 2.** A view of what a Texas English Learner would see when she clicks on the word meadow in the sample item.
Technology Standards for Combining and Reporting Student Data

To enable a better and more comprehensive understanding of the learning and plan instructional interventions for English learners, reports must combine information from the English proficiency assessments and the assessments of academic content. These recommended standards are needed to promote comprehensive, useful, and helpful reporting.

**Recommended Technology Standard 5: Combining Data from Both Assessment Types**

Systems need to facilitate the combination and exchange of data from English proficiency as well as academic performance assessments so that students, parents, and educators can more fully understand a student’s progress.

Since the progress of an English learner in academic content depends on the learning of enough English so they can meaningfully engage in instruction, understanding the educational progress of English learners and planning their instruction requires that reports for these students provide results for English Learners in English proficiency, in academic performance, in the relation between their English proficiency and academic performance. Reports need to combine these sources of information, so systems need to be designed to facilitate this reporting.

**Recommended Technology Standard 6: Reports Accessible in Students’ Primary Language**

Reports for English learners should be provided (where practical) in the student's primary language even after students have learned English.

Regardless of whether a student has acquired sufficient English to interpret assessment reports, his or her parents or guardians who are English learners may still require score reports in their primary language. Therefore, the accessibility of score reports and instructions for interpreting them should also be considered as part of technology standards. In fact, many of the same technology strategies that help students who are English learners can be used to facilitate the ability for their parents to understand assessment information; however, as with their children, parents and guardians who are English learners may have unique needs that must be addressed.

**Technology Idea 2: Statistical Methods for Linking Student Scores on Both Assessment Types**

Adapting common statistical methods such as growth and projection models to enrich the measurement of progress of an English learner’s readiness to meaningfully engage in English academic content will continue to move the assessment of English learners forward.

Students will develop English proficiency at unique rates and will at some point acquire sufficient English to meaningfully engage in assessments of academic content written in
English as long as linguistic accommodations are available to the student. Two critical components to understanding where students are and what instructional interventions are best for a particular student include:

1. **Students’ English Progress**—methods to measure student growth in academic content can be adapted to inform educators about a student’s growth in learning English. For example, growth measures such as growth to proficiency, value tables, projection measures, and growth percentile methods should be evaluated for application in informing the growth or progress of students learning English. The use and reporting of progress measures with English learners has the opportunity to inform parents not only where students are in their English language development, but also what progress they have made and the extent to which that progress puts them on track to engage in academic content in English within a reasonable timeframe.

2. **Accurate identification of Students’ Readiness to Learn and Test in English**—methods currently being implemented to measure college readiness should be evaluated for application for use with English learners to inform about their readiness to engage in academic content in English. For example, statistical methods can be used to provide a probability that students with a specific pattern of scores (e.g., English proficiency scores in listening, speaking, reading, and writing) are ready for academic assessments in English with linguistic accommodations. As an example of this, the Texas Education Agency contracted with Pearson to implement a model that projects student performance from one set of assessments to another. For example, Texas currently uses the Texas Projection Measure to project a student’s English reading score in grade 5 from that student’s reading and mathematics scores in Spanish in grade 3.

Additionally, as measures of text complexity and their requirements driving future education reform, such statistical methods may facilitate the transition of English learners as the move from learning English to using English to learn. For example, these statistical linking and/or projection models can be used to estimate if English learner candidates are ready to engage English text given its documented text complexity.
3.2.30 Transparency

**Requirement**

3.2.30  **Transparency.** How do the organizations that develop assessment technology standards approach development and maintenance activities? Is it common for such work to be performed in an unrestricted or open public forum? Are there examples of organizations conducting technology standards development through private (e.g., membership-driven) activities? Are the final work products produced through standards-development activities made publicly available in a timely manner? If not, when or for how long is it necessary to keep these products private? What circumstances require, justify, or benefit from protecting trade secrets or intellectual property?

**Response**

Pearson acknowledges this requirement and believes that these questions would be best answered directly by the group working on the standards.

3.2.31 Participation

**Requirement**

3.2.31  **Participation.** Does the development of assessment technology standards depend on membership fees from individuals and organizations who wish to contribute to development and maintenance activities? Are there requirements for “balance” within membership across different constituencies? What are the cost and structure of such memberships? Are there viable alternative methods for generating revenue necessary to conduct the work? What are the most realistic and useful ways to generate participation, fund work, and ensure public access to a technology standards-setting process?

**Response**

Pearson acknowledges this requirement and believes that these questions would be best answered directly by the group working on the standards.

3.2.32 Availability

**Requirement**

3.2.32  **Availability.** What are the costs associated with final publication of technology standards, and with all supporting materials for those standards, and can these assessment products be made available at nominal or no cost to users? Do technology standards require restrictions for use or application, including limitations on derivation, resale, or other restrictions? Is it appropriate to obtain patent, copyright, or trademark protections for assessment technology standards? Are the publications for technology standards and materials provided in a machine-readable, well-defined form? Are there restrictions or limitations on any future application of the publications and materials after initial release? Are developer-assistance materials (e.g., Document Type Definitions, test harnesses, code libraries, reference implementations) also made available free under an open-license? In what circumstances should technology standards-setting organizations retain rights or control, or impose restrictions on the
use of publications, derivations, and resale or developer-assistance technologies, as opposed to open-licensing everything? When should materials be made freely available (that is, at no cost to the consumer) while still retaining most or all copyright license rights?

Response

Pearson acknowledges this requirement and believes that these questions would be best answered directly by the group working on the standards.

3.2.33 Derivation

Requirement

3.2.33 Derivation. For technology standards, do copyright licenses for publications and all supporting materials and software licenses for software artifacts permit the unrestricted creation and dissemination of derivative works (a.k.a. “open licensed”)? Do such open licenses contain restrictions that require publication and dissemination of such works in a manner consistent with the openness criteria described by, for example, a GNU Public License (a.k.a. “viral licensed”) or an MIT Public License (a.k.a. “academic licensed”)? Are there policies or license restrictions on derivative works intended to prevent re-packaging, re-sale, or modifications without re-publication for assessment technology standards?

Response

Pearson acknowledges this requirement and believes that these questions would be best answered directly by the group working on the standards.

3.2.34 Licensing Descriptions

Requirement

3.2.34 Licensing Descriptions (for materials contained within the standard, not for the standard’s licensing itself). How do technology standards address licensing terms for assessment resources described within the technology standard? Are there successful technology standards or approaches for describing a wide variety of license types, including traditional per-use licensing, Web-fulfillment, free (but licensed), open (but licensed, including commercial or non-commercial use permitted), and public domain status. Are there other resource licensing issues that should be addressed within a technology standard as a best practice?

Response

Pearson acknowledges this requirement and believes that these questions would be best answered directly by the group working on the standards.