# Next-Generation Assessment Interoperability Standards

A white paper from Pearson

Bob Dolan
Ellen Strain-Seymour
Ashman Deokar
Wayne Ostler

November 2010 (Revision B)

PEARSON

*Using assessment and research to promote learning*

# Contents

## Executive Summary

The Race to the Top (RTTT) Assessment Program and state consortia adopting the Common Core State Standards (CCSS) have identified interoperability as essential to ensure the feasibility and affordability of next-generation assessments. In addition, the general nature of large-scale assessment is likely evolving from the 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 enhanced item sophistication and interactivity 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 furthermore ensure 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 next-generation assessment interoperability standards are recommended:

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 to support greater accessibility and a range of devices.

Support Wide Range of Content - Assessment interoperability standards must support a wide range of content for greater interactivity, increased accessibility, more sophisticated assessments, and easier item 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, stakeholders are recommended to 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.

<u>Leverage the Cooperation and Support of Oversight Organizations</u> - 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.

<u>Allow Stakeholders to Determine the Best Interoperability Solutions</u> - 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.

## Purpose

The intent of this document is to elevate awareness and understanding of the importance of interoperability standards in primary and secondary educational assessment and to begin addressing the necessary evolution of these standards to support next-generation assessments. The technology environment in which assessment solutions are operating is evolving rapidly. To keep the competitive and innovative landscapes alive, robust and highly deployable interoperability standards are imperative to enable assessment program consumers to choose the best solutions to advance education outcomes. To facilitate initial dialog around this process, this document provides (1) an understanding of the importance of interoperability standards for assessment, (2) an analysis of the current state of assessment interoperability standards, and (3) recommended requirements and strategies for defining assessment interoperability standards capable of supporting next-generation assessments.

## The Need for Standards

Simply stated, *standards* are what allow various systems, such as mechanical components, electronics, and software, to work together. For example, the United States has a standard configuration for connecting an electric appliance via a plug to an electric power supply via an outlet. While the standard establishes some clearly visible and physical configurations, less visible attributes of the standard include voltage/amperage, materials used, etc. This standard allows a variety of electric appliances to work from the same source of electricity without undue concern for harm to the user or the appliance, when properly used. While standards are necessary, they do have limitations:

- **Standards may not be universal.** Additional devices, processes, or mechanisms may be necessary to achieve interoperability when multiple standards are in use or vary by location or industry. For instance, an adaptor is necessary to convert between standards and operate some U.S. electric appliances in other countries.
- **Multiple standards may be required.** When one standard does not account for all needs and circumstances, one strategy to account for more needs is to introduce another standard. For instance, the United States introduced another standard for electric appliances and outlets to accommodate appliances, such as electric clothes dryers, that require higher DC power outlets and cannot be accommodated by standard AC outlets.
- **Standards must evolve over time.** When a standard becomes insufficient because of new developments, the standard may need to change, be replaced or be supplemented by another standard. Such evolution—such as outlets' support for a third prong for grounding—is often necessary and usually desirable, even though the transition between the old standard and the updated standard is inconvenient.

## Assessment Interoperability Standards

Assessment interoperability standards are necessary for independent testing components to work together. For example, two parties building test items in accordance with an assessment interoperability standard could exchange these items with minimal manual intervention. As another example, if the standard is sufficiently comprehensive and sophisticated, then two systems should be able to exchange entire item banks and associated test definitions in order to deliver, score, and report with similar—if not identical—outcomes. An even more sophisticated example would be when one student's performance results from a test correlated to a set of curriculum standards is usable by a different system to direct the student to targeted instructional materials correlated to those same curriculum standards. Without assessment

interoperability standards, such collaboration between separate players and systems would at best require manual intervention and at worst be unrealizable.

**Existing Assessment Interoperability Standards**

Today a few well-recognized industry standards define assessments or use assessment data as part of their standards. These are:

**QTI**

The IMS Global Learning Consortium[1] developed the Question and Test Interoperability (QTI) specification[2] as an Extensible Markup Language (XML)[3]-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. QTI's initial adoption was primarily focused on online adaptive testing in the licensure and certification markets. Appendix A provides additional information on QTI.

**SIF®**

The Schools Interoperability Framework (SIF) Association[4] has developed standards for data format and transport of data among educational systems. The specification covers a wide range of data for students, staff, facilities, academic records, etc. Assessment objects were added to the SIF specification to support assessment results reporting. While SIF's initial market penetration was in K-12, that scope is expanding beyond K-12 into pre-K, higher education, workforce, and international markets as its usage and requirements expand. Appendix B provides additional information on SIF.

**SCORM®**

The Sharable Content Object Reference Model (SCORM ®)[5] was developed to foster creation of reusable learning content as "instructional objects" within a common technical framework for computer-based and Web-based learning. SCORM describes that technical framework by providing a harmonized set of guidelines, specifications, and standards based on the work of several distinct e-learning specifications and standards bodies. To date, SCORM has not included a robust representation of assessment content, emphasizing instead a "black box" approach to assessment interoperability.

---

[1] http://www.imsglobal.org
[2] http://www.imsglobal.org/question/
[3] http://www.w3.org/standards/xml/; background information on XML and its relevance to assessment interoperability standards is provided in Appendix D: XML and its Relevance to Interoperability Standards
[4] http://www.sifinfo.org/
[5] http://www.adlnet.gov/Technologies/Pages/Default.aspx

**PESC**

> The Postsecondary Electronic Standards Council (PESC)[6], established in 1997, promotes interoperability and data exchange standards among information systems. While focused primarily on the post-secondary education market, there are many connections to K-12. Like SIF, PESC's entry into the assessment domain is focused on results reporting rather than on a broad assessment interoperability standard.

These standards vary in their scope, comprehensiveness, and rate of industry adoption, with considerable overlap among them. For instance, both SCORM and SIF address assessment within the broader context of education-related data sets (learning content and student information, respectively), making them particularly appropriate when considering an integrated view of assessment, learning, and the tracking of student progress. On the other hand, QTI's more complete coverage of item content and test structure (depending on the QTI version) comes much closer to delivering on the goal of sharing test and item content across systems.

## The Domain of Assessment Interoperability Standards

In general, a comprehensive assessment interoperability standard that will support the full life cycle of building, delivering, scoring, and reporting assessments must address the following components:

▪ Assessment item content – The basic starting point with any assessment interoperability standard is the ability to define and exchange item content. Item content is typically structured based on the type of item represented. For example, multiple-choice elements may contain a stem and a set of answer options, while a short-answer item may contain just the stem. Item content can include elements of text, graphics, or multimedia and can be designed for administration by various media, such as paper and digital technologies. For more advanced item types, content may include interactive, or executable, objects.

▪ Assessment assets – Additional content elements separate from items often appear as components of an assessment. Common examples include a reading comprehension passage or a formula sheet used on a math or physics test. Separating these content elements from item content maximizes reusability and minimizes redundant content mappings.

▪ Assessment item accommodations – Extensions to the item content would define appropriate accommodation attributes associated with the item to support accessible presentation and response options for students, including those with disabilities and who are English-language learners. Examples of such accommodations include alternate tagging of graphics and linguistic simplifications of low-frequency vocabulary words. Such content-related elements of these accommodations—those aspects that would vary from item to item—would be a part of the item specification and could reference other assets such as an audio recording, Braille, a high-contrast version of an image, or video captioning data. In addition, allowable modes of response could be specified on a per-item basis and a per-student basis, such as whether use of a scribe or voice recognition software would be appropriate.

▪ Assessment item metadata – State assessment programs track innumerable data points around items to properly build test forms, field test items, and maintain item banks. Such metadata include curriculum standards alignment, performance statistics, usage/exposure data, psychometric properties, friend/enemy/sibling relationships with other items, and suitability for delivery on particular devices, to

---

[6] http://www.pesc.org/

name a few. The sharing of common learning standards across states requires particular attention given the imperative to track both commonalities and differences in individual states' implementation from the standpoint of metadata, reporting categories, aggregation rules, etc.

▪ Test forms – In addition to describing individual items and assets, parameters for assembling items to build a deliverable test form are necessary. To build an adaptive test, the entire item bank for a subject and grade may be compiled into a single form. Forms may also specify test structures (such as sections of the test), specify branching and navigation rules, and identify additional content assets that must be included in the test form. Common accommodations that are typically made available at the test level, such as the availability of zoom capability or text/background color options, could also be specified as part of the test form specification.

▪ Scoring – Scoring rules for each item should be included in the standard. The scoring rules would vary based on the type of item. For standard multiple-choice items, an answer key suffices, but some items may require lists of values, ranges, or a specified scoring approach or logic. Some types of items will require a rubric and may involve human evaluation. In addition to item-level scoring information, scoring will also require information about groups of items (such as objectives or subtests), raw-to-scale score rules, normative tables or rules, and attempted-ness rules to determine a final score.

▪ Reporting – Reporting can include traditional score reporting as well as exchanging results with various information systems. Sufficient detail about the assessment results must be provided to make the information meaningful and actionable. While the information provided by the scoring system is probably sufficient for individual score results, additional information could be provided concerning longitudinal data, growth, or aggregated data analysis.

▪ Feedback – While the existing industry standards provide methods to link assessment content to learning standards, additional content or links to content may be required to deliver the types of feedback needed as part of a formative assessment system, for example. Students and instructors may need diagnostic level information for assessment items or subtests that can help them pinpoint areas of misunderstanding or improvement. It may also be desirable to provide content or links to information for each answer option or other answer options to provide detailed diagnostics about, for example, potential conceptual misunderstanding or lack of background knowledge.

These assessment interoperability components represent the minimal domain that must be covered for standards to support existing and near-future assessment systems in a robust and comprehensive fashion.

## Improving Interoperability Standards to Support Next-Generation Assessments

The Race to the Top (RTTT) Assessment Program[7] and state consortia adopting the Common Core State Standards (CCSS)[8] have appropriately identified interoperability as essential to make next-generation assessments feasible and affordable. In doing so, they have pushed for technical requirements of assessment interoperability standards to move beyond current limits. The Partnership for Assessment of Readiness for College and Careers (PARCC)[9] and the SMARTER Balanced Assessment Consortium (SBAC)[10] have stated their commitment to open standards and to a high degree of interoperability across

---

[7] http://www2.ed.gov/programs/racetothetop-assessment/index.html
[8] http://www.corestandards.org/
[9] http://www.fldoe.org/parcc/
[10] http://www.k12.wa.us/smarter/

systems so that teachers could use templates for creating items across a range of item types and items could be delivered and scored in a variety of systems suited to the varied assessment venues: end-of-the-year/summative, benchmark/interim, through-course, on-demand, and curriculum-embedded. Beyond the core requirements of an assessment program, optional components would be available to states, districts, and schools, with the knowledge that those components could effectively participate in a system of content/data exchange based on standards, as items, tests, scoring routines, rubrics, performance data, and curriculum standards alignment data are passed among interoperable systems.

According to RTTT funding guidelines, not only will next-generation assessments function within a world of greater cross-system interoperability, but the nature of individual assessment items will change from the multiple-choice items common in large-scale assessments to a blend of item types to improve measurement of learning. The PARCC consortium states that its assessments will be computer-based to leverage fast turnaround and innovation possible through new technologies. Computer-enhanced items and performance tasks will "elicit complex demonstrations of learning and measure the full range of knowledge and skills necessary to succeed in college and 21st century careers."[11] Similarly, SBAC embraces computerized adaptive testing (CAT) and complex computer-based simulations to test higher-order cognitive abilities. New item types will be required to assess the full range of skills and abilities represented by the CCSS. In both consortia, learning and assessment technologies will be more closely connected and leveraged to greater degrees to improve the use of assessment to inform instruction and to track and attain educational goals.

A critical question is whether our current assessment standards can handle the challenges of greater expectations of interoperability across systems (including student information and instructional systems as instruction and assessment are more tightly integrated) and enhanced item sophistication and interactivity to measure performance against the CCSS. With the following statement, the RTTT funding guidelines already suggest that some re-evaluation of existing standards may be necessary:

> *We encourage grantees under this competition to work during the grant period with the Department and the entities that set interoperability standards to extend those standards in order to make them more functional for assessment materials.[12]*

Extension of standards has been a key strategy used by adopters of current assessment standards to address gaps between current and desired functionality. However, such extensions have not always been achieved through work with the entities that govern the interoperability standards. Instead, current interoperability standards are often supplemented through proprietary extensions involving a limited set of parties. While such custom extensions have allowed vendors to work around any perceived limitations of current standards to meet students' and educators' needs in strategic ways, they compromise interoperability across systems. Information is lost when all systems have not implemented or are not aware of the custom extensions.

Clearly, custom or proprietary extensions will not provide the necessary solution: they may deliver enhanced item complexity and interactivity but at the cost of reduced interoperability. Conversely, greater interoperability can be achieved through very basic standards that simplify to the lowest common denominators that all parties can agree upon; however, in this case, interoperability would be achieved only through loss of sophistication and power of our assessment strategies. Therefore, collaboration between vendors and assessment specification governance boards is required to arrive at coherent and

---

[11] http://www.fldoe.org/parcc/pdf/prosum.pdf
[12] http://www2.ed.gov/programs/racetothetop-assessment/executive-summary-042010.pdf (page 4)

comprehensive assessment standards that deliver greater diagnostic and measurement precision with total interoperability. Additionally, to ensure that that simultaneous goals of innovation and standardization are not at odds with one another within the evolution of assessment standards, agile ways to expand the standard over time to embrace innovation must be established. Extensions must be easy to propose with reasonable turnaround time and provisional acceptances to allow innovation to thrive as CCSS assessments are established and new research proceeds.

Based upon these new requirements, the next section provides a set of recommendations for defining next-generation assessment interoperability standards.

## Next-Generation Assessment Interoperability Recommendations

Following is a set of recommendations for next-generation assessment interoperability standards.

### Recommendation 1: 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.

The availability of intriguing new devices for the classroom—including but not limited to handheld and tablet devices—combined with the CCSS consortia's interest in the integration of technology into learning and assessment calls for standards that minimize assumptions about existing technologies and allow for evolution of technology platforms. More than pointing out any one technology "winner," recent debates over Adobe® Flash®, Apple® iPad, and HTML5 have taught us that the technology landscape is shaped by a variety of volatile factors: business strategy, larger-than-life personalities, media buzz, unexpected breakthroughs, and fast technology adoption in some cases—as well as a continued reliance on existing platforms and available equipment in other cases. The success of future assessments supported by RTTT and the CCSS is too important to gamble on the success of any one delivery technology. Next-generation content interoperability standards must be robust and flexible enough for deployment on yesterday's platforms as well as tomorrow's devices. Fortunately, an XML-based standard that stops short of mandating specific implementation or formatting strategies achieves that robustness and flexibility (see Appendix D: XML and its Relevance to Interoperability Standards for additional information on XML).

### Recommendation 2: Separate Content and Formatting

The separation of content and formatting within an assessment standard is critical to support greater accessibility and a range of devices.

With a successful interoperability strategy in place and a model for easily authoring and sharing items across states and across classrooms, an item may travel great distances by jumping across systems, living a long life as a secure item, and entering retirement in the form of a released practice item that appears on student handheld devices or in a classroom-based quiz engine. How that item should be styled and formatted will likely differ according to local conventions and the terms of its use. It could be optimized for appearance on an iPad but appear very differently when delivered in a large-print format. An individual student may even be able to control certain aspects of its appearance, such as font size and contrast, that are adjusted to suit that student's ideal viewing scenario within an accommodated test.

Exactly how these issues are managed is the domain of various stakeholders, such as vendors' system architects, usability specialists, and accessibility researchers, but it is the content standard itself, as handled by each delivery system, that assures content integrity across various system-based incarnations. The item that arrives cemented into a singular format is an item that cannot take advantage of dynamic layouts, advanced accessibility strategies, or device-specific formatting optimization. If formatting specifications also dictate use of a certain technology, the standard's flexibility and robustness amidst future technology changes begin to disappear. The sample schema referenced in Appendix C: Assessment Content Examples provides examples of how separation of content and formatting can be accomplished.

**Recommendation 3: Support Wide Range of Content**

<u>Assessment interoperability standards must support a wide range of content for greater interactivity, increased accessibility, more sophisticated assessments, and easier item authoring.</u>

Supporting a wider range of content refers not only to the definition of new item types, but also more seemingly mundane but important aspects of an assessment item. For instance, with the absence of a dedicated tag, endnote content would need to be represented through formatting cues alone. On the other hand, with a standard that does recognize endnotes as a content type, authoring systems can become more intelligent with auto-numbering, and systems can optimize endnote display for the delivery environment. Whereas an endnote at the bottom of a printed page can be highly convenient, students may have more difficulty scrolling down to an endnote on a small screen; rendering each on-screen endnote as a pop-up link from the endnote reference number may provide a better user experience.

Supporting a wider range of content also leads to greater accessibility. For instance, if an image is to be displayed with a title, a permission line, and a caption, those elements should be explicated as a part of the specification. Not only would this allow for the display of these elements to be optimized for the delivery environment and styled to match a given state's style guide requirements, but it avoids the embedding of these elements within the image itself. Once these text elements are embedded, they become more difficult to edit, impossible to map to varying styles/formats, and inaccessible to text-to-speech technologies used to support students with print disabilities (except through embedding a separate version of that content as alternate text that must be synchronized as item edits are made). Additionally, a wider range of content-specific accommodations (such as text equivalents) could be included with any given item as part of a broader interoperability standard.

Ease of authoring is also enabled through a standard that supports a wide range of content. For example, if an interactive item type involves graphing points or lines on a line graph, then the content standard should include all parameters (such as title, axis labels, etc.) for defining the line graph, specify interactivity (such as maximum number of points that can be added, whether lines should connect or remain distinct like in a scatterplot), and determine the correctness of a response (such as the resulting line must have a positive slope, consist of the following points, or adhere to the following equation). The systems consuming this XML would then render the line graph, provide the specified item interactivity, and score the response. The cost for producing such an item decreases with this standardization and system support in place. Similarly, teachers would be able to create this kind of item through simple editing environments that generate the required XML. Although the original intention of such tags might be to support a graphing item type, a desirable side effect is that the creation of XML-rendered line graphs—even static ones—becomes possible, through yet another content tag that designates the line graph as non-interactive, for more conventional items without the cost of graphics production.

The most critical impetus behind expanding or reinventing current standards is to support the evolution of novel item types. These new item types could resemble performance-based assessment, include rich

simulations, mirror classroom activities, and measure some of the CCSS constructs such as critical thinking that are otherwise difficult to assess. Such item types become affordable through a number of means: the sharing of research and development costs through state consortia and federal funding; the sharing of the items themselves across state members of a consortium; authoring templates facilitated through an underlying XML standard for the item types; and standardized delivery rather than one-by-one item programming. The sample schema referenced in Appendix C: Assessment Content Examples provides examples of how standardized representation of a wide range of item types can be accomplished.

**Recommendation 4: Address Assessment Structure, Accommodations, and Additional Standards**

<u>Assessment interoperability standards must address assessment structure and test-level accommodations, and integrate well with other standards and other instructional resources.</u>

Complete assessment interoperability standards must address the exchange of item-level content and support the exchange of entire test definitions, whether those are highly structured (such as paper-based forms) or dynamic, logic-driven assessments (such as computerized adaptive testing). A fully elaborated test definition that included all item structures, resources, tools, navigation and layout controls, etc. could drastically increase the interoperability of assessments across test creation, delivery, scoring, and reporting systems. Sophisticated systems for tracking and improving educational outcomes would not isolate assessments but rather make use of them within learning systems, data analysis environments, and intervention tools. To accomplish this integration, assessment standards must operate within or alongside other standards used within learning management systems and student information systems. In addition, discipline-specific content standards, such as MathML[13], should be embraced for the powerful capability that they provide for equation rendering and intelligent automated scoring.

## Recommended Next Steps

To support the evolution of assessment interoperability standards for next-generation assessments, stakeholders are recommended to consider the following **concurrent** steps:

**Step 1: Engage Existing Standards Boards**

Many of the current successes in assessment interoperability have been achieved through isolated, proprietary extensions that include only stakeholders directly involved, with limited or no participation by the larger community or centralized standards boards. Without centralized agreement and ratification, the opportunity for widespread and consistent use of these extensions—however technically sound—is compromised. Moving forward, it is critical for standards evolution to occur through active involvement via standards boards, most notably IMS Global and the SIF Association. Such evolution of next-generation interoperability standards may occur through extensions to existing versions, definitions of incremental updates, and/or the creation of radically redefined versions. In all cases, the evolution process ideally would be made by a wide range of stakeholders through centralized governing bodies. Although the need for custom extensions may still be necessary in certain circumstances, an open and transparent approach toward standards evolution will minimize their occurrence and optimize their eventual inclusion as ratified modifications or extensions.

---

[13] http://www.w3c.org/Math/

**Step 2: Leverage the Cooperation and Support of Oversight Organizations**

Industry organizations should be invited to participate from the beginning to define next-generation assessment interoperability standards. Two such organizations are the Council of Chief State School Officers (CCSSO)[14] and the Association of Test Publishers (ATP)[15]. Both of these organizations have served the assessment community for a long time; CCSSO is one of two organizations that led the CCSS initiative. By involving these organizations as neutral and knowledgeable third parties, the hope is to avoid potential disputes between competing interests. Such organizations can also serve as partners in distributing information and soliciting input and feedback across a wide set of stakeholders.

**Step 3: 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 standards 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. It is too early to tell whether the best solution will be, for example, extensions to one of the QTI versions currently in use, a radically new version of QTI or SIF, or—perhaps less likely at this time—another standard altogether.

---

[14] http://ccsso.org/
[15] http://www.testpublishers.org/

## Appendix A: QTI as an Assessment Interoperability Standard

The Question and Test Interoperability (QTI) standard, developed by the IMS Global Learning Consortium, is a widely adopted standard for representing assessment content online. The most widely used version of QTI is v1.2, which was finalized in 2002 with an addendum (v1.2.1) in 2003. In addition to conventional item types such as multiple- choice, this standard supports a number of interactive item types that allow a student to sequence, drag, or select words and graphics. QTI couples Extensible Markup Language (XML)[16] with HyperText Markup Language (HTML)[17] (in the form of XHTML[18]) to combine item content and formatting. This strategy is well suited for standard formatting techniques used by browser-based delivery systems; however, this has had unintended side effects in terms of item portability. As described in Appendix D, the standard imperatives behind the use of XML for representing test content are the separation of content and formatting as well as interoperability. QTI's use of XHTML allows content and formatting data to intermingle rather than remaining segregated. Traditionally, it is this segregation that allows for concise, precisely defined content management. Mixing content and formatting data increases the difficulty in mapping multiple formats to the same content set when delivering test items across multiple media, such as print and online. In terms of interoperability, QTI has made significant inroads, but QTI's acceptance of HTML within the QTI standard means that vendors often find it difficult to guarantee absolute interoperability and consistency of appearance across delivery systems.

With the continued reliance on print test delivery systems in primary and secondary education and customer interest in a guarantee of similar appearance across delivery platforms, QTI has not reached widespread adoption in this market. Test publishers with a significant portion of their business focused on print tests have shied away from an HTML-based format and consequently failed to gravitate toward a single content interoperability standard.

While earlier versions of QTI focused primarily on interoperability needs in the context of learning management systems, the latest version, v2.1, was intended to address broader assessment needs, as well as address shortcomings in the assessment structures and correct errors in prior versions. This version also intended to address a significantly greater degree of integration with other specifications, some of which did not exist during the production of v1.0, including IMS Content Packaging v1.2, IEEE Learning Object Metadata, IMS Learning Design, IMS Simple Sequencing, and other standards such as XHTML.

Although released for public evaluation as a precursor to IMS approval in 2008, IMS withdrew v2.1 in early 2009, citing inadequate feedback. The availability of this unendorsed version is currently limited to IMS members, which has stymied continued debate and conversation around this most recent and still unapproved version of the QTI standard. Some vendors have been unwilling to take on the risk of developing systems around v2.1, particularly as discussion of a possible v2.2 has included the statement that v2.2 may not evolve linearly from v2.1. However, IMS has been involved in recent activity addressing student assessment accessibility needs; this effort, funded by the U.S. Department of Education and involving eight states, will result in the Accessible Portable Item Profile (APIP) v1.0 scheduled for release in late 2010.

---

[16] http://www.w3.org/standards/xml/; background information on XML and its relevance to assessment interoperability standards is provided in Appendix D: XML and its Relevance to Interoperability Standards
[17] HyperText Markup Language, the publishing language of the World Wide Web (http://www.w3.org/html/wg/)
[18] Extensible HyperText Markup Language (http://www.w3.org/TR/xhtml1/)

## Appendix B: SIF as an Assessment Interoperability Standard

The Schools Interoperability Framework (SIF) provides standards for data format and transport of data among educational systems. The specification covers a wide range of data for students, staff, facilities, academic records, etc. While SIF's initial market penetration was in K-12, that scope is expanding beyond K-12 into pre-K, higher education, work force, and international markets as its usage and requirements expand. Australia and the United Kingdom have both produced versions of the SIF specification that are adapted to their data needs.

Assessment objects were added to the SIF specification to facilitate the reporting and exchange of assessment results among systems. To date, the SIF assessment specification has not been widely adopted because the required transport frameworks for moving assessment results from vendors' assessment systems to school information systems have not been developed. Many of the current SIF users have been requesting greater capability from the assessment portion of the specification to meet the ever-changing demands that assessments place on their educational programs. The standard today is limited to assessment results reporting, and no commercialized data transport frameworks exist to facilitate exchange of assessment results. SIF does not include significant item content formatting or test construction objects in its current state.

When looking at SIF's suitability as an item/test content interoperability standard, the SIF Association's openness to expanding the standard is encouraging and its larger structure for student data is appealing. The basic assessment information structures that are available in SIF are structured well to map to the online and paper administration styles common in K-12 and that will be required by the CCSS. For example, test forms are of critical importance to large-scale test administration environments in K-12 and are a clearly defined structure in SIF. Test forms are loosely defined in QTI and leave much about the specification to people's interpretations. The SIF assessment specification provides the basic building blocks for a robust assessment interoperability specification. With those basic building blocks in place, SIF has great potential to support the necessary enhancements outlined in this document and to achieve the level of interoperability needed for RTTT-funded assessment programs.

## Appendix C: Assessment Content Examples

This appendix provides real-world examples of how an assessment interoperability standard can accommodate a wide range of item types and item interactivity and represent items in a portable, systematic, and flexible fashion across multiple systems. Since this document is intended for non-technical as well as technical audiences, sample XML snippets and screenshots of the prototype items are provided in lieu of a formal schema, which would require some training and familiarity to understand.

The philosophy behind this prototype schema is to allow consumption of even complex assessment items by a number of vendors' systems, regardless of the platform. In other words, standards based upon such schema would not dictate details of the technical implementation; the item XML would describe content and interactivity that, with appropriate system support, could be rendered on various devices regardless of programming language or platform. Formatting—font size, typeface, line weights, etc.—could be controlled item templates or styles rather than being dictated in the item XML. Methods for optimizing display for the particular devices' screen or paper sizes will undoubtedly vary. Different style specifications could also be mapped to the item based on states' preferences and based on student needs, such as large-print displays as an accommodation for low-vision students.

Stakeholders would collaborate to define any number of elements within such an open-standard schema, the exact names of elements and attributes, ways to enable additional functionality, ways to bring this XML into greater alignment with existing standards, etc. The critical issue is that stakeholders agree about the desirability of such functionality within items aligned to the CCSS and an appropriate schema (or schemas) so that work can begin on various fronts, such as creating authoring tools, building scoring algorithms, and architecting the delivery systems that would host these items.

The full XML schema used in these examples will be made public in the near future.

### Polygon Graphing

One sample set of XML that has been prepared for evaluation controls display and interactivity for graphing lines, functions, and polygons on a coordinate grid. For purposes of aligning to CCSS around data graphing and geometry, this XML would facilitate:

- the display of static line graphs, scatterplots, graphed functions, and polygons rendered on a coordinate grid
- interactive items in which a student graphs a range of function types or systems of inequalities
- interactive items in which a student plots a scatterplot or line graph
- interactive items in which a student creates one or more polygons

Trapezoid PQRS can be formed by rotating trapezoid ABCD shown below 90° in a clockwise direction using the point (0, -2) as the center of rotation. Use the coordinate grid below to plot trapezoid PQRS and label its vertices.
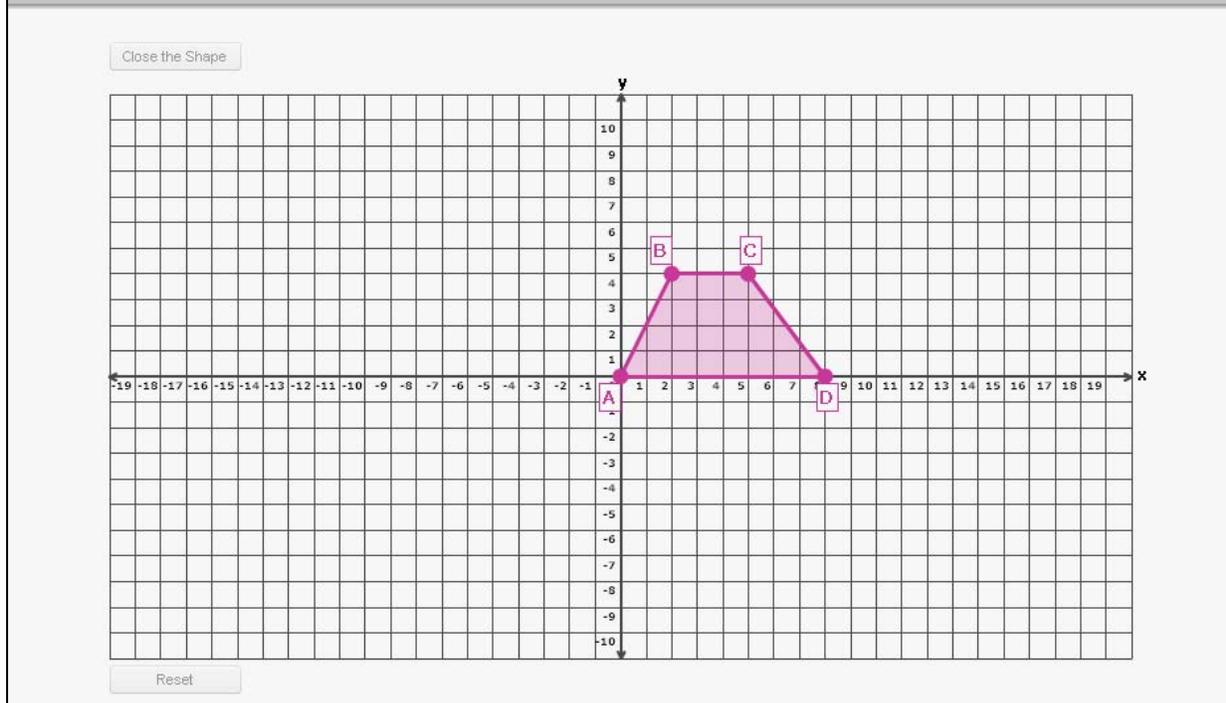


In this example, an XML structure (<coordinateGraph>) and associated attributes would inform the item presentation engine how to build and control the graph display. For example, the number of lines to display on the coordinate grid both vertically (vLines) and horizontally (hLines) would be provided:

```
<coordinateGraph vLines=20 hLines=11/>
```

vLines = 20 indicates that 20 lines to the left and to the right of the Y axis are drawn. With XML structures like these, the presentation engine can determine how to display the coordinate graph independently for each item.

To control how the axes (xAxis and yAxis) are labeled, additional XML tags and attributes are provided with starting values (<startPoint>) and increments (<increment>). In the example provided, all lines are labeled starting at 0 and increase by 1 for each line label. The following XML example snippet illustrates how this would be controlled:

```
<xAxis startPoint=0 increment=1/>
<yAxis startPoint=0 increment=1/>
```

Having the graph display properly on a screen or paper would require a complete set of XML tags and attributes to give the presentation engine enough information to construct the graph.

Once the graph has been created, more informational XML tags and attributes would determine how students would make choices and interact with the graph to construct their answers. A student could be

asked to construct all aspects of the graph or to simply plot points on a pre-constructed grid. For editable graph attributes, students would be given an on-screen control to edit that element. For example, students would either select from a range of choices in an available pull-down menu or fill in a text field to edit graph titles and axis labels.

In addition to enabling students to change some of the graph's characteristics, students will likely be asked to plot points, lines, functions, or shapes on a grid. The programmed logic for connecting plotted points with line segments will vary based on the type of graph (<graphType>) being constructed. For example, for a line graph the dots may simply be connected in order of ascending x-values as they are placed on the grid, but for a polygon shape the dots may be connected in the order they are placed on the graph. For a scatterplot, the plotted points may not be connected at all. In the case of function or equation graphing, the graph type will control the mathematics behind the line/curve of the shape being constructed.

Other XML tags and attributes may control how many points students can place (<maxPointsPerShape>) on the graph or how many shapes (<maxShapes>) can be plotted. For the graphing item example used earlier, these values may appear as follows:

```
<graph graphType="polygonGraphing"/>
<pointRules maxPointPerShape=4 maxShapes=1/>
```

Another control that would increase the accessibility and usability of this type of item is the "snap-to-grid" option (<snapTo>). This instructs the software to precisely locate the points on the graph when students are using a less precise pointing interface such as touch screen or touch pad or when disabilities or challenges make accurate movements difficult for the student. By adding the snap-to-grid option, the points plotted will be automatically placed on the closest grid line intersection:

```
<pointRules maxPointPerShape=4 maxShapes=1 snapTo="true"/>
```

Lastly, the correct answer (<correctAnswer>) could be specified in various ways: perimeter, area, number of sides, shape (such as rectangle, square, parallelogram), or by exact points. For this example, exact point position (<pointPos>) and labeling (<label>) may be required to receive full credit. A sample XML structure for this would look like:

```
<correctAnswer scoringLogic="exactPoints">
 <pointPos x="2" y="-2" label="P"/>
 <pointPos x="6" y="-4" label="Q"/>
 <pointPos x="6" y="-7" label="R"/>
 <pointPos x="2" y="-10" label="S"/>
</correctAnswer>
```

### Voice Recording

To assess speaking and listening skills within English language arts, computer-based item types that allow students to listen to and record audio will become increasingly important. Audio recording can be used in a variety of ways, including (1) to assess reading fluency at lower grade levels, (2) to measure the speaking skills of English language learners, (3) to provide students with options in how they capture their response (spoken, typed, or hand-written), and (4) to test Common Core State Standards that specifically address speaking, and in combination with other tools that allow students to assemble a response using images, video, written language, and recorded audio.



The user interfaces and technologies for recording audio may differ between test delivery systems and across devices, but the parties involved in building CCSS assessments will likely agree upon the types of controls and options available within an audio recording environment. For instance, each item might have a time limit (<maxTime>) specified as well as a maximum number of attempts (<maxAttempts>) allowed for students to delete and start over. Sample XML for this item may be as follows:

```
<audioRecording maxTime=60 maxAttempts=5>
```

It may also be necessary to "turn off" some controls for items intended for younger students; however, for older students more controls may be enabled. For example, controls indicating that the recording is being made (<progressBar>) and proper volume levels (<audioLevels>) are maintained throughout may be desired for some items but not others. The following XML may be used by this item to determine which audio controls are available:

```
<controls record="true" stop="true" play="true" delete="true" progressBar="false" audioLevels="false">
```

## Virtual Labs and Simulations

The CCSS's emphasis on higher-order thinking, problem-solving, and mathematical modeling, in combination with the state consortia's pursuit of performance tasks, sophisticated item types, and online simulations suggest ever-widening boundaries for interactivity within item types. Some work has already begun in the area of virtual laboratories and simulations within computer-based science assessments, but such work may expand with CCSS-aligned math items that pursue real-world applications of math in problem-solving situations. Additionally, future expansion of Common Core State Standards to include science would encourage further exploration of virtual labs.

To date, simulations have been expensive to produce and frequently involve complex scoring logic. Even with state consortia sharing the cost of research and item development, the level of effort involved in developing such items could be a limiting factor without some standardization. One route to standardization is the use of XML to express the underlying logic of a simulation or lab. For instance, if we assume that a simulation involves a range of options that can be selected by students and that those options have an effect on the object of study, an XML schema can be devised to support this logic structure. A complex, mathematically driven simulation may involve a set of data inputs and a set of data outputs that are determined by a set of equations or formulas. Students would choose the inputs by accessing user interface controls such as sliders or pull-down menus. After a trigger button is applied, display objects would show the students the results of the simulation. These data inputs and outputs can be recorded in a table. In some cases, the table may be the primary way that the outcome is displayed, but in other cases, standardized or item-specific display objects (such as a calculator, a thermometer, a digital scale, a number line, a function machine, a pair of dice) may be used to present the data outputs.

Students may be scored on their use of the simulation, on items that accompany these simulations, or on a combination of the two (such as an open-response item being evaluated in light of the outcomes of the simulation). The most complex of these scoring scenarios—student response data (such as a line graph) evaluated against the simulation outcomes sent along as an accompanying data set or image—might require human scorers or the use of artificial intelligence during automated scoring.

The following example assumes that every option students select has an effect: an animation is triggered, descriptive text is generated, one of three success criteria is set to true, certain subsequent options are made available or unavailable, and a new end state is reached. In this case, three unknown materials have been combined in a beaker. Students must successfully separate them using a magnet, water, a hot plate, and a filter. Students may be assessed on successful completion of the task, completion of the task in the fewest number of steps, or on items that assess what students are able to deduce about the materials once successfully separating them.



The XML for this item tracks each potential outcome for each possible state. For instance, heating the materials with the hot plate (studentSelection="hotplate") in their initial state (<state id = "initial">) has no change in outcome; the end state is the same as the initial state, as shown in the following snippet:

```
<state id = "initial">
        <outcome studentSelection="hotPlate">
                <animation>heat_state1</animation>
                <displayText>Mixture heated. No effect.</displayText>
                <endState>initial</endState>
        </outcome>
</state>
```

However, using the magnet (studentSection="magnet") on the materials in their initial state triggers one of the success criteria and a new end state because iron filings have been removed from the materials in the beaker.

```
<state id = "initial">
        <outcome studentSelection="magnet">
                <animation>magnet_state1</animation>
                <displayText>Material sticks to magnet. Substance 1 successfully
                        separated.</displayText>
                <success>Substance 1</endState>
                <endState>salt_sand</endState>
        </outcome>
</state>
```

Although no two simulations are identical, many may share similar logical structures for expressing inputs and outputs. With a library of display objects that can change with data inputs (such as a thermometer that takes a numerical input and uses it to display a temperature or a parabola that changes focus, vertex, and line of symmetry based on inputs) and a series of XML structures that can be used to express the relationship between inputs and outputs, simulations can become more economical to produce. Some animations, used as standardized display objects, can be data driven to show, for instance, a beaker being filled from one volume level to another. Ideally, teachers would even use such objects or create their own for classroom assessment or instructional use.

## Appendix D: XML and its Relevance to Interoperability Standards

As a prerequisite to a technical understanding of assessment interoperability standards, a brief explanation of Extensible Markup Language (XML)[19] is warranted since any interoperability scheme will be created atop XML. XML is a system used for tagging electronic file-based content so it can be consumed and interpreted by technological systems. Similar to HyperText Markup Language (HTML)[20] that is widely used to format web pages, XML goes further by allowing for tagging of aspects of the content beyond specification of formatting alone. That is, XML tags provide access to the content so that systems can make content-relevant decisions, search, or apply business rules. For example, XML tags can define additional data about the content such as who is the <author> or who owns the <copyright>. XML tags can control interactivity and provide easy access to item authors who want to modify interactivity but have no access to programming languages. Conversely, HTML tags contain formatting instructions such as <B> for bold and <U> for underline, but remain unaware of the exact nature of the content beyond a limited set of tags such as <header>.

Businesses using XML may either develop their own proprietary system of tags—known as a schema or document type definition—to best capture internal business data and naming conventions, or they may use an existing industry standard or some combination of the two.

The two typical business drivers for using XML are:
1) Improved content management through the separation of content and formatting
2) Interoperability, or the exchange of data across different providers' systems when using the same XML schema

Content that has been tagged via XML can be used for standard data-driven operations (such as find all content authored by a particular person) akin to functionality often achieved through a database. The tags can also be used to inform formatting processes (such as represent content tagged as <bookTitle> in a particular font style such as <italic>).The separation of content and formatting is important because that allows for the same content to be formatted in a variety of different ways to support multiple modal delivery methods (print, online, large-print, tablet, handheld).

---

[19] http://www.w3.org/standards/xml/
[20] http://www.w3.org/html/wg/