K-12 Center at ETS

Invitational Research Symposium on
Technology Enhanced Assessments

Using Technology to Assess Hard-to-Measure Constructs in the Common Core State Standards and to Expand Accessibility

Kathleen Scalise

May 7–8, 2012

K-12 Center at ETS

Invitational Research Symposium on Technology Enhanced Assessments

Using Technology to Assess Hard-to-Measure Constructs in the Common Core State Standards and to Expand Accessibility

Kathleen Scalise

University of Oregon

Executive Summary

The new U.S. Common Core State Standards (CCSS) are in many aspects incredibly challenging for assessment. The educational standards have brought some hard-to-measure areas of assessment, or constructs, to the fore, and also placed them in new contexts for assessment. While it is an exciting moment in the United States as states eagerly look forward to advancing instructional practices in the classroom, the five assessment consortia funded to develop measures around the CCSS are seeking solutions for best practices. Many areas that they are charged with addressing have never been assessed in state summative practices before, so challenges arise. This paper shares examples from three assessment developers on how hard-to-measure constructs are being addressed.

In this paper, technology enhanced item types (TEIs) are examined specifically for the purposes of summative assessments. First, hard-to-measure constructs are defined, and a TEI framework is introduced, along with an accessibility tool. Next, some ground rules are established for the key aspects of measurement that all assessments must address. Finally, the three demonstrations from developers are considered from the hard-to-measure vantage point. The hope of the ETS Technology Enhanced Assessment (TEA) symposium session from which this paper is derived is to provide some direction for states struggling with the tension caused by the desire to incorporate innovative item types into their online assessments. The goals are tapping hard-to-measure constructs while maximizing accessibility and also maintaining measurement standards of high quality evidence.

Hard-to-Measure Constructs: What They Are

The states are not alone in trying to understand hard-to-measure constructs. Teachers and educators unpacking the Common Core State Standards (CCSS) often find themselves struggling to define what makes a hard-to-measure construct. They know it when they see it. But what aspects does it possess? Several attributes can make some types of skills and knowledge challenging to assess summatively. These include the following circumstances:

1. The test measures a trait that is difficult to define, or as yet remains insufficiently defined.
2. The trait is cross-cutting and must play out in a vast variety of contexts that necessarily must remain ill-defined, in order to serve student learning needs.
3. Knowledge, skills, and attitudes (KSAs) to be addressed involve interactions and dependencies with other types of skills and knowledge not intended to be investigated in the construct, or at that time.
4. Adequate construct coverage is difficult to achieve due to limitations in respondent time available, materials or contexts to be used, scoring resources applied, or accommodations/modifications necessary to serve all students.
5. The construct itself does not have sufficient stability for the grain size of inference intended, which may be an issue of fundamental cognitive science.

None of these five attributes mean that hard-to-measure assessment should not be attempted. Rather, the information desired and the decision making to be accomplished should drive the goals and objectives of measurement.

It can be helpful for both assessment developers and educators to engage in a process of specifically calling out what is hard-to-measure. When unpacking the CCSS, teachers can ask themselves the following question: Which of the traits above make a particular part of the core more challenging than others to assess? Then they can ask a follow-up question: What is the most appropriate solution in each case?

Cognitively complex tasks themselves do not necessarily create a hard-to-measure assessment. For many areas such as in medicine, IT accreditation, and other areas of adult learning, the domain has been sufficiently explored and enough respondent time and appropriate resources set aside to yield desired evidence that is quite cognitively complex. In these situations, important inferences can be made that may be considered productive in accountability practices such as accreditation, program evaluation, and educational funding choices, and/or in instructional decision making and classroom practices that powerfully inform learning gains. However, those working in K-12 education, be they policy makers, school leaders, or classroom teachers and educators, should recognize they are often on new ground, taking on more difficult areas to measure.

Identifying in each specific case where the difficulty is coming from is important. This is because the solution can be radically different, depending on the hardness challenge. For instance, for insufficiently defined constructs as in Circumstance 1 above, detailed fine-grained domain modeling such as that developed through evidence-centered design makes it possible to describe KSAs in a fine-grained way and may help resolve the difficulty. However, for Circumstance 2, where cross-cutting traits necessarily must play out in a vast variety of ill-defined contexts, such as critical thinking, it would be a disservice to students and misdirection to teachers if the domain of application were locked down in a fine-grained way. Rather, students need to exemplify critical thinking practices over a broad range of ever-varying tasks in order to be career- and college-ready. This requires data density, or multiple opportunities for students to show what they know and can do regarding critical thinking, and a willingness on the part of educators to not circumscribe too narrowly. As another example, where KSAs interact and have dependencies (Circumstance 3), providing support such as audio tools for passage reading in math word problems may be helpful.

When information is aggregated over an assessment to generate a score report, there is also a hidden source of difficulty of which many policy makers and educators may be unaware. As Roy Levy (2012) described in his paper for the Technology Enhanced Assessments Symposium, the maturing of measurement models for accumulating evidence, or bringing all the evidence to bear to generate the score report, is an issue. Over 6 decades of research and application, measurement technology for summative assessment has matured to establish well-accepted procedures for important issues, which include calibration and estimation of a student overall score, reliability and precision information, test form creation, linking and equating, adaptive administrations, evaluating assumptions, checking data-model fit, differential functioning, and invariance. According to Levy, new approaches are in their infancy when it comes to their application as measurement models in larger assessment enterprises.

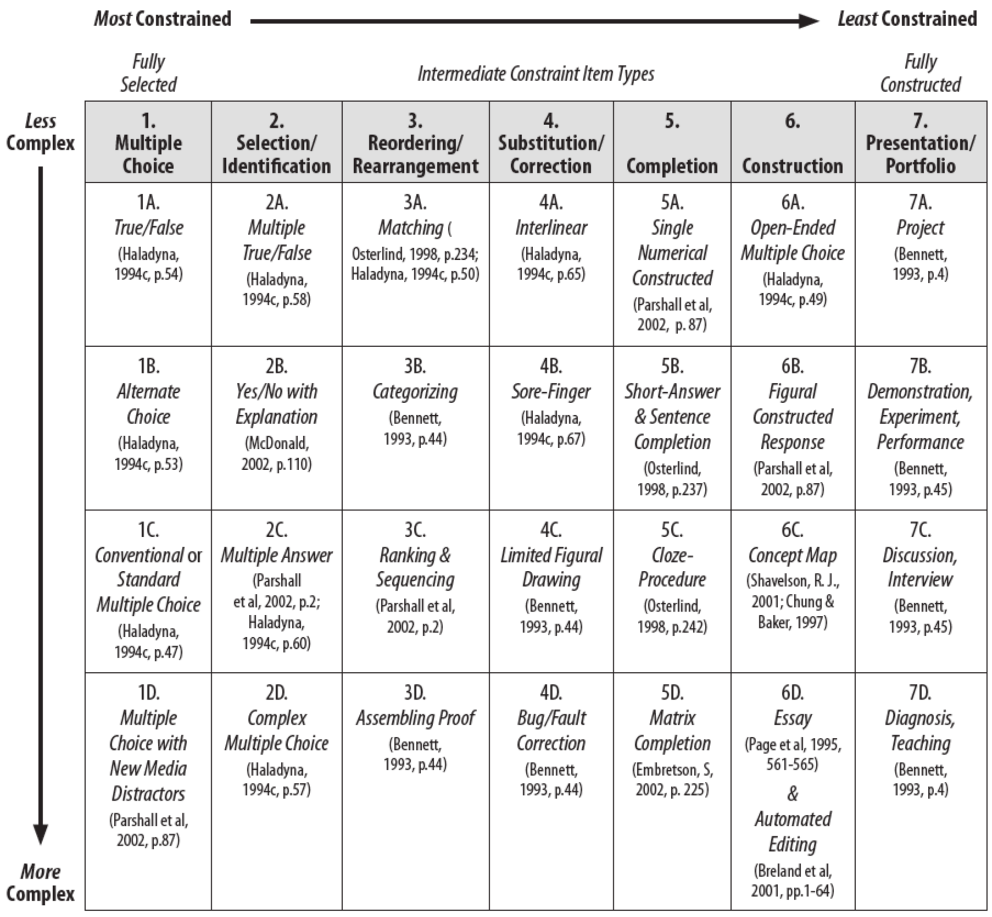
This hidden problem in the measurement model means that current approaches used operationally in large-scale assessment in the United States may not always cover the full need of the new construct challenges. While some additional evidence-aggregation approaches have been used for more than 20 years in other settings, they also have limitations. Ultimately, it may be unlikely that any of these approaches that have been in use for a number of decades are in themselves the next-generation measurement models to go with the next-generation constructs represented by the new standards (Kathleen Scalise, 2012). Rather, borrowing strength across ways to accumulate evidence is likely to be important. A summary of what may characterize next-generation approaches to accumulating evidence (measurement models) includes the following:

- Borrowing strength from each other, for example, one measurement model within another or one model extending another
- Establishing multiple inferential grain sizes
- Sometimes drawing on stronger confirmatory data from improved domain modeling
- Sometimes using richer range of exploratory data to explain more variance
- Functioning with much more data—but also noisier data, and being able to incorporate high quality evidence, in situations where it is deemed to provide utility, across large-scale and classroom-based practices if needed
- Allowing a wide range of constructs, observations, and interpretations for 21st century teaching and learning needs.
- Matching inferential evidence beginning to be provided routinely in other contexts surrounding us, through affordances of the information age.

A Framework for Technology Enhanced Item Types (TEIs) and an Accessibility Tool

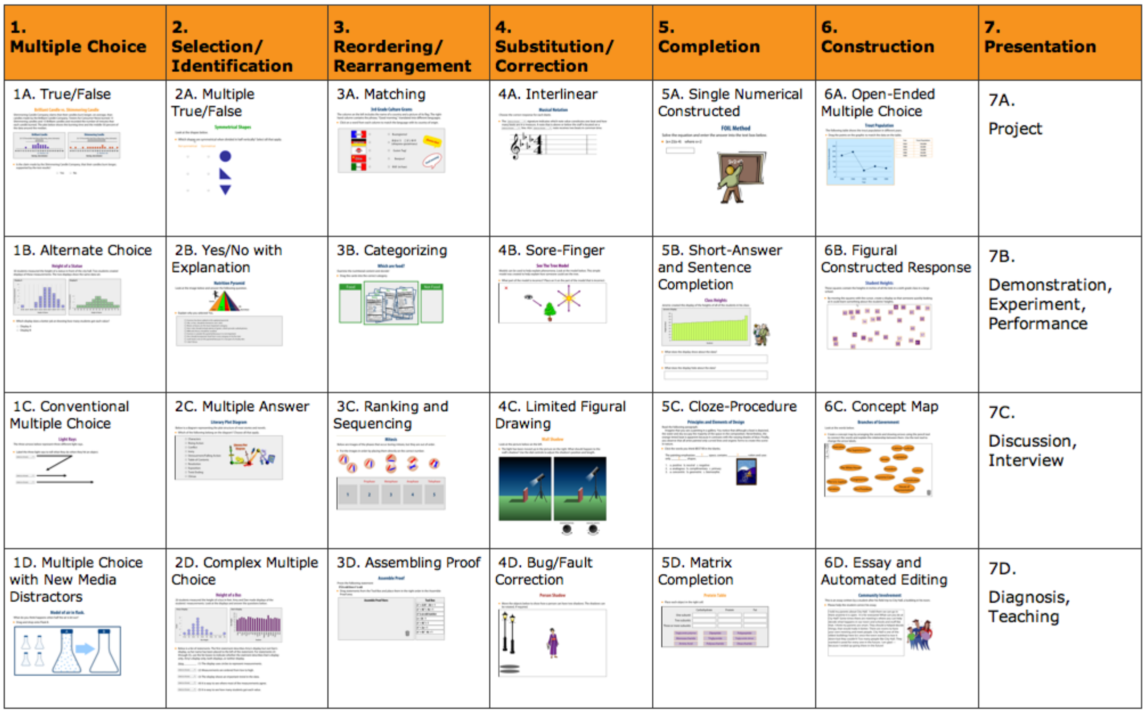
In this section, I briefly bring together a framework for classifying technology enhanced item types (TEIs), along with a tool for accessibility planning. This will help us to examine the demonstrations presented in upcoming sections.

Figures 1 and 2 show 28 technology enhanced item type examples organized into a taxonomy based on the level of constraint in the item/task response format. They were generated by reviewing numerous examples of technology enhanced assessments and classifying tasks that the respondent is asked to completed. The most constrained item types, at left in Column 1, use fully selected response formats. The least constrained item types, at right in Column 7, use fully constructed response formats. In between are intermediate constraint items, which are organized with decreasing degrees of constraint from left to right. There is additional ordering that can be seen within each type, where innovations tend to become increasingly complex from top to bottom when progressing down each column.


Most Constrained
Least Constrained
Fully Selected
Intermediate Constraint Item Types
Fully Constructed
Less Complex
More Complex

1. Multiple Choice	2. Selection/ Identification	3. Reordering/ Rearrangement	4. Substitution/ Correction	5. Completion	6. Construction	7. Presentation/ Portfolio
1A. *True/False* (Haladyna, 1994c, p.54)	2A. *Multiple True/False* (Haladyna, 1994c, p.58)	3A. *Matching* (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. *Interlinear* (Haladyna, 1994c, p.65)	5A. *Single Numerical Constructed* (Parshall et al, 2002, p. 87)	6A. *Open-Ended Multiple Choice* (Haladyna, 1994c, p.49)	7A. *Project* (Bennett, 1993, p.4)
1B. *Alternate Choice* (Haladyna, 1994c, p.53)	2B. *Yes/No with Explanation* (McDonald, 2002, p.110)	3B. *Categorizing* (Bennett, 1993, p.44)	4B. *Sore-Finger* (Haladyna, 1994c, p.67)	5B. *Short-Answer & Sentence Completion* (Osterlind, 1998, p.237)	6B. *Figural Constructed Response* (Parshall et al, 2002, p.87)	7B. *Demonstration, Experiment, Performance* (Bennett, 1993, p.45)
1C. *Conventional* or *Standard Multiple Choice* (Haladyna, 1994c, p.47)	2C. *Multiple Answer* (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. *Ranking & Sequencing* (Parshall et al, 2002, p.2)	4C. *Limited Figural Drawing* (Bennett, 1993, p.44)	5C. *Cloze-Procedure* (Osterlind, 1998, p.242)	6C. *Concept Map* (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. *Discussion, Interview* (Bennett, 1993, p.45)
1D. *Multiple Choice with New Media Distractors* (Parshall et al, 2002, p.87)	2D. *Complex Multiple Choice* (Haladyna, 1994c, p.57)	3D. *Assembling Proof* (Bennett, 1993, p.44)	4D. *Bug/Fault Correction* (Bennett, 1993, p.44)	5D. *Matrix Completion* (Embretson, S, 2002, p. 225)	6D. *Essay* (Page et al, 1995, 561-565) *& Automated Editing* (Breland et al, 2001, pp.1-64)	7D. *Diagnosis, Teaching* (Bennett, 1993, p.4)

Figure 1. Intermediate constraint taxonomy for e-learning assessment questions and tasks (Scalise & Gifford, 2006).



1. Multiple Choice	2. Selection/ Identification	3. Reordering/ Rearrangement	4. Substitution/ Correction	5. Completion	6. Construction	7. Presentation
1A. True/False	2A. Multiple True/False	3A. Matching	4A. Interlinear	5A. Single Numerical Constructed	6A. Open-Ended Multiple Choice	7A. Project
1B. Alternate Choice	2B. Yes/No with Explanation	3B. Categorizing	4B. Sore-Finger	5B. Short-Answer and Sentence Completion	6B. Figural Constructed Response	7B. Demonstration, Experiment, Performance
1C. Conventional Multiple Choice	2C. Multiple Answer	3C. Ranking and Sequencing	4C. Limited Figural Drawing	5C. Cloze-Procedure	6C. Concept Map	7C. Discussion, Interview
1D. Multiple Choice with New Media Distractors	2D. Complex Multiple Choice	3D. Assembling Proof	4D. Bug/Fault Correction	5D. Matrix Completion	6D. Essay and Automated Editing	7D. Diagnosis, Teaching

Figure 2. Interactive examples of the intermediate constraint taxonomy, available at http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html

Figure 1 provides category names and references, while Figure 2 provides worked examples. Figure 2 is available online, where the examples can be seen in their interactive state and are available as open source (http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html).

In the United States, technology enhanced innovations in assessment are tending to move from the left to the right of this taxonomy. By contrast, in other countries such as the United Kingdom, the challenge for technology adoption operates in the opposite direction. This points out how innovation depends on the current context. According to Stuart Elliott, director of the National Research Council Board on Testing and Assessment, what serves as a benchmark of effective innovation is what lets us better measure what we want to assess (Scalise, 2012).

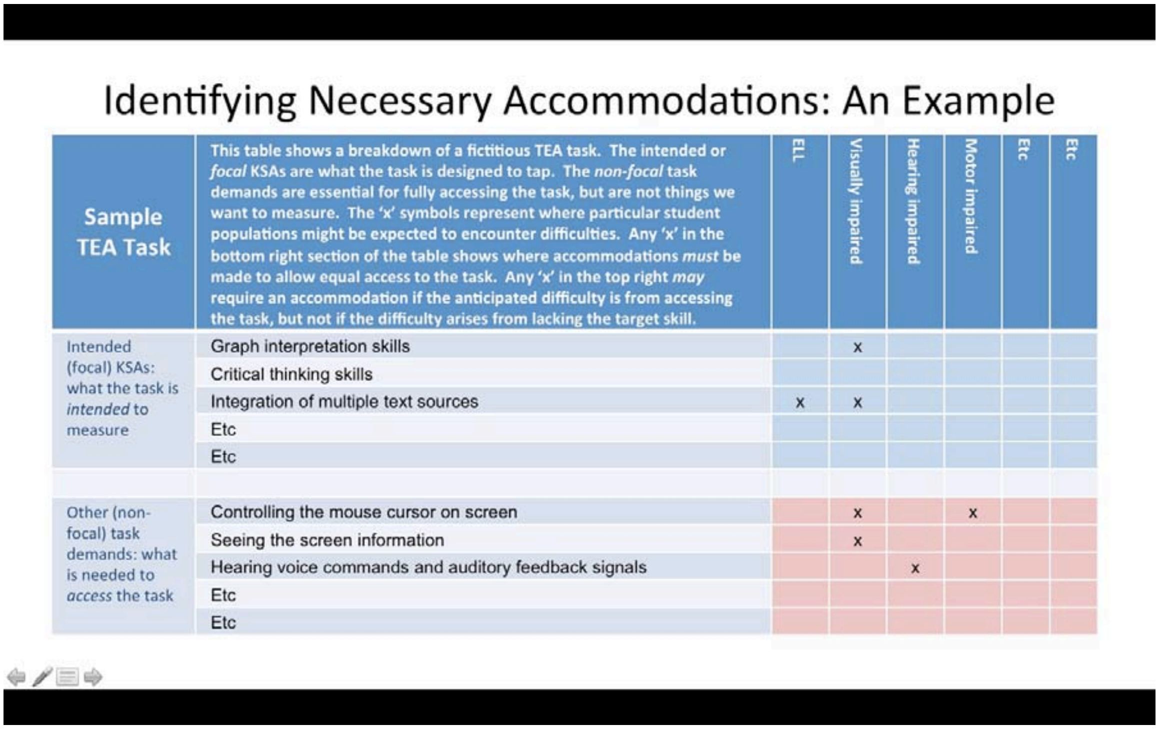
In regard to making assessments accessible to all students, technology enhanced environments offer both new challenges and new opportunities. To take one example, the SRI demonstration paper from the TEA symposium described b how the SRI developers meet the challenge of diversity by employing universal design for learning (UDL) to suggest flexible assessment materials, techniques, and strategies (Haertel et al., 2012, p. 9):

> The flexibility of UDL empowers assessors to meet the varied needs of students and to accurately measure student progress. The UDL framework includes three guiding principles that

address three critical aspects of any learning activity, including its assessment. The first principle, multiple means of representation, addresses the ways in which information is presented. The second principle is multiple means of action and expression. This principle focuses on the ways in which students can interact with content and express what they are learning. Multiple means of engagement is the third principle, addressing the ways in which students are engaged in learning (Rose & Meyer, 2002, 2006; Rose, Meyer, & Hitchcock, 2005).

Each of the developers described below were encouraged to consider their accessibility needs by using the accessibility tool shown in Figures 3 and 4. Originally developed by ETS, the tool is an organizer that calls out specific needs in technology enhanced tasks (Cayton-Hodges et al., 2012). The first portion elicits information from developers on their intended KSAs and what other skills are additionally required. The second considers a range of needs present in the population.

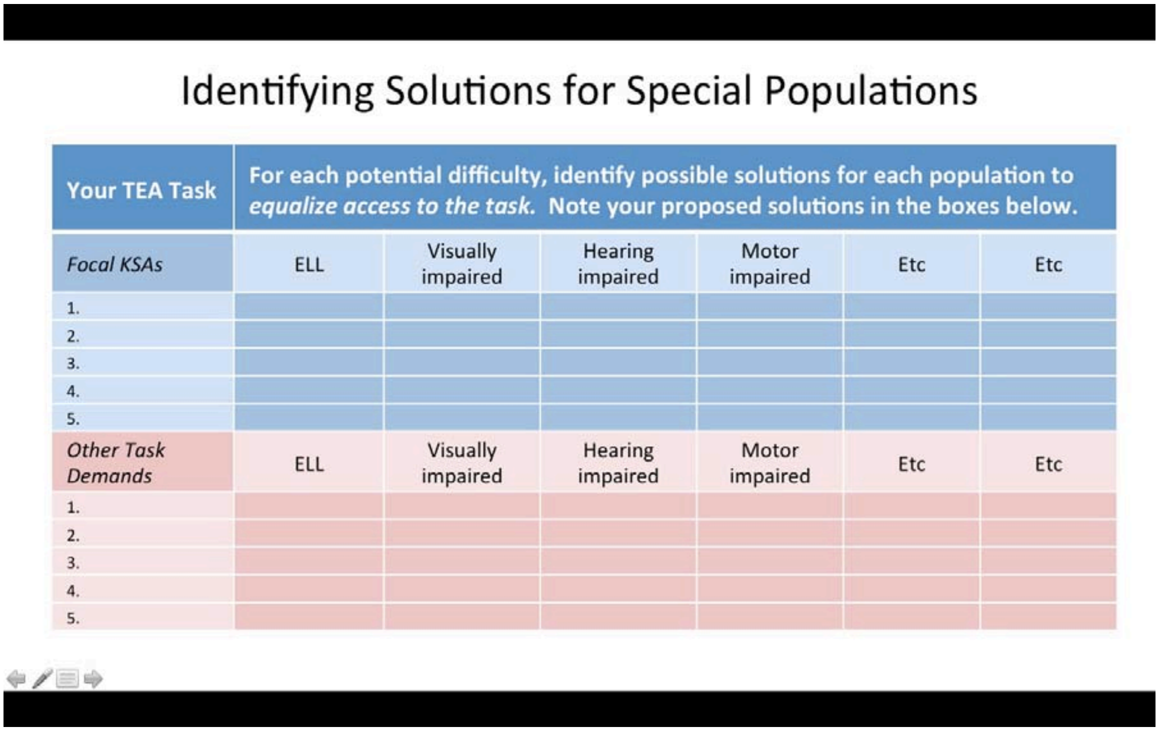
All three examples supply a wealth of knowledge about applying such tools in assessment. Discussed below, the development papers provide important detail on aspects of current best practices.



Identifying Necessary Accommodations: An Example

Sample TEA Task	This table shows a breakdown of a fictitious TEA task. The intended or *focal* KSAs are what the task is designed to tap. The *non-focal* task demands are essential for fully accessing the task, but are not things we want to measure. The 'x' symbols represent where particular student populations might be expected to encounter difficulties. Any 'x' in the bottom right section of the table shows where accommodations *must* be made to allow equal access to the task. Any 'x' in the top right *may* require an accommodation if the anticipated difficulty is from accessing the task, but not if the difficulty arises from lacking the target skill.	ELL	Visually impaired	Hearing impaired	Motor impaired	Etc	Etc
Intended (focal) KSAs: what the task is *intended* to measure	Graph interpretation skills		x				
	Critical thinking skills						
	Integration of multiple text sources	x	x				
	Etc						
	Etc						
Other (non-focal) task demands: what is needed to *access* the task	Controlling the mouse cursor on screen		x		x		
	Seeing the screen information		x				
	Hearing voice commands and auditory feedback signals			x			
	Etc						
	Etc						

Figure 3. First portion of the organizer tool, which elicits information from developers on their intended KSAs and what other skills are additionally required.



Identifying Solutions for Special Populations

Your TEA Task	For each potential difficulty, identify possible solutions for each population to *equalize access to the task.* Note your proposed solutions in the boxes below.					
Focal KSAs	ELL	Visually impaired	Hearing impaired	Motor impaired	Etc	Etc
1.						
2.						
3.						
4.						
5.						
Other Task Demands	ELL	Visually impaired	Hearing impaired	Motor impaired	Etc	Etc
1.						
2.						
3.						
4.						
5.						

Figure 4. The second portion of the organizer tool considers a range of needs present in the population.

Task Surrounds: What They Are and Why They Are Important

In the next sections, hard-to-measure constructs are explored through three demonstrations aligned to the CCSS that have been provided by assessment developers. First, it is important to make the point that many of the more challenging constructs are being explored, as in all three demonstrations here, not through single questions but through a set of activities arranged in what are being called *task surrounds*. A task surround, like all coherent assessments as described by the National Research Council Assessment Triangle (National Research Council, 2001), must consider the construct assessed, the observations employed, and interpretations connecting the two.

Task surrounds are technology affordances – a set of small software programs that work together to create a set of activities, such as for a research or inquiry activity, which can readily be populated with new content. Task surrounds are not item shells in the traditional measurement sense of theoretically calibrated attempts at producing alternate items or forms with similar psychometric properties. The content may or may not have similar psychometric properties, depending on how it has been designed. However, the task surround will allow a new experiment to be created, another essay research topic to be presented, and so forth, readily within the assessment authoring system.

Task surrounds are typically built up modularly from individual task templates classifiable in the Intermediate Constraint Taxonomy. For instance, one U.S. assessment consortium of states describes its TEI templates as a single interaction, response data collected as a result of that interaction, and the logic applied to score the response data (Smarter Balanced, 2012). By becoming an assembly of such TEI templates, the larger task surround can help tap the cognitive complexity of a variety of hard-to-measure constructs. Both surrounds and their embedded templates often can be used across grade levels and content areas, although some adjustments to which templates appear may be desirable as the surround moves between grades and constructs.

Task surrounds built up from such templates look very similar to e-learning activities used online in many classrooms today. Thus they can align well with instructional practices and begin to allow integration between assessment and instruction, which is a major plus where desired. They also introduce interdependence in data collected, if they are part but not all of an assessment instrument, which is another example of the next-generation evidence assembly approaches needed (Type 3), as described previously in the section defining hard-to-measure constructs.

Three Demonstrations: On-the-Ground Experience With Hard-to-Measure Constructs

In this section, the task surrounds designed by SRI, ETS, and CTB will be discussed. Each task surround brings together a different set of TEIs that are intended to build an assessment experience at depth for the respondent. At the same time, different inquiry and research tasks can be presented to the respondent simply by varying the prompts and content materials in the surrounds. Further flexibility can be gained by swapping some of the template pages within the surrounds, to vary the type of research, inquiry or other activities in which the students engage.

The three vendors of assessment instruments were asked to select their strongest example of a technology enhanced assessment tapping a hard-to-measure construct in the new Common Core State Standards. Sue Rigney of the U.S. Department of Education and I worked with the developers and symposium organizer Nancy Doorey to prepare a set of presentations and papers on the demonstrations. Each of three groups was asked to develop a set of items to assess a hard-to-measure construct. SRI demonstrated a construct within the National Academy of Science's recently released Conceptual Framework for New Science Education Standards (National Research Council, 2012), ETS demonstrated for a construct within the CCSS in mathematics, and CTB did so for a construct within the CCSS in English language arts.

The set of items was to be designed to measure the identified construct and to include at least one operational TEI. Papers and presentations explained the extent to which the demonstration met these criteria:

- Measures important hard-to-measure constructs within the CCSS/science framework validly and reliably
- Increases student engagement and motivation
- Improves the precision of measurement
- Signals good instruction
- Is financially feasible to develop and score

Each developer has written a paper describing the assessment development process undertaken to develop the set of items, with special emphasis on the TEI portions. Accessibility pathways are integrally threaded through all three examples. Capturing UDL solutions and building engagement into the task for the respondent is exemplified throughout. A full rendition regarding the development of each project is well worth reading and can be accessed in the three separate papers (Barton & Schultz, 2012; Cayton-Hodges et al., 2012; Haertel et al., 2012).

The papers began with an understanding of best practices in evidence-centered design (ECD) through the SRI example in the science area (Haertel et al., 2012), followed by an ETS example operationalizing the connection to instruction through learning progressions employing technology enhancements for mathematics assessment (Cayton-Hodges et al., 2012). Finally, the McGraw-Hill CTB paper added to both of these conversations by showing the building of a task surround for literature analysis and research in English language arts (Barton & Schultz, 2012). Its example employs a combination of intermediate constraint item types and stretches to the right of the table in Figure 1 with a range of automated scoring approaches.

To provide a clear structure for the reader, each presenter was asked to describe in sequence each of the following major steps in the assessment development process:

1 The CCSS standards/construct to be measured.

2 The primary or focal KSAs to be assessed.

3 The student behaviors or performances/products accepted as evidence of the KSAs.

4 The TEI tasks or stimuli that should elicit those cognitive behaviors and performances.

5 The characteristic features likely to evoke the desired evidence.

6 The options for uses of technology within the item format and rationale for choice made;

7 The aspects of the task that may be varied to improve accessibility for individual students, and how those variable features would be applied.

8 Any constraints on either the presentation of the item or the student response format, or other practical obstacles for implementation.

9 The scoring methodologies and measurement models to be used and level of information that would be reported.

10 The process by which the items would be pilot tested and validated before use in a summative/consequential assessment.

The SRI example (Haertel et al., 2012) described a design methodology for improving the validity of inferences about the performance of students on large-scale science assessment tasks. The authors showed how they integrated two conceptual frameworks, one for ECD and one for UDL. These were instantiated into an online assessment design system.

A primary focus of the SRI demonstration was on the conceptualization of validity as an argument and chain of reasoning. Haertel et al. (2012) described how, through ECD, they developed a coordinated and coherent assessment or assessment system by building an assessment argument across layers of work that begin with the analysis and organization of the conceptual domain to be assessed and culminate in the delivery, scoring and reporting of the assessment results to stakeholders.

SRI showed a Pinball Car Race task as a demonstration of its approach in the TEA symposium presentation. The task, described here, can be seen in Haertel et al. (2012, p. 17):

> The Pinball Car Race is a middle-school science assessment task that was designed to test a student's knowledge of both science content and practices. The science content being assessed is knowledge of forms of energy in the physical sciences. In particular, knowledge of potential and kinetic energy and that objects in motion possess kinetic energy. In the assessment task, students observe the compression of a spring attached to a plunger, the same type of mechanism as those used to put a ball—in play in a pinball machine. The student observes that when the plunger is released, it pushes a toy car forward on a racing track.

Through scenario-based items, the SRI example shows a task surround for scientific inquiry in physical science. The ETS example, Proportional Punch in mathematics, uses an inquiry approach as well—in this case, to explore the recipe for a punch drink, which is adjusted and examined to assessment a developmental understanding of proportional reasoning.

The ETS paper (Cayton-Hodges et al., 2012) highlighted connections between assessment and instruction. The authors examined how a key advantage of TEAs over traditional assessment media is the ability of TEAs to offer a dynamic environment. Cayton-Hodges et al. (2012) also discussed dynamically adjusting a simulation that represents mathematical processes. The task surround examines whether students can extend their thinking from concrete reasoning through abstraction to symbolic reasoning. The authors described how the task surround allows them to capture new kinds of evidence and explained how this evidence includes information about the processes that students undertake to complete a task as well as the outcome of their reasoning.

The ETS demonstration focused on a technology enhanced task from the ETS Cognitively Based Assessment *of, for, and as* Learning (*CBAL™*) research initiative. Competency models and learning progressions form the theoretical foundation of the CBAL approach, so the foundational domain modeling approach is quite different from that in the SRI example. In the CBAL task, Proportional Punch, the ratio between the mix and the water indicates how sweet the punch will be. An interactive tool and sweetness meter simulate making punch and determining its sweetness in the task. Of course, this work builds on considerable efforts by others in mathematics education; citations and references for this effort can be seen throughout Cayton-Hodges et al. (2012).

The ETS paper explored a range of innovative accessibility solutions. Cayton-Hodges et al. (2012) pointed out that technology enhanced items can be particularly challenging for students with fine motor impairments due to the need for manipulation of objects on a screen, for students with issues with memory capacity because often a lot is going on in the screen , and potentially for low vision students or those with hearing needs, as much of the new stimuli can be visual or auditory in new and different ways than those that were used previously in paper-and-pencil assessments. Many of the TEI response modes require vision, and soon they may be requiring more sound.

The McGraw-Hill CTB task surround (Barton & Schultz, 2012), aligned to the CCSS for English language arts, looked quite different from the inquiry task surround demonstrated in the other two projects. The surround arranged a set of five technology enhanced items to assess students through a process of literature research and analysis. The technology enhancements included such features as video, audio, student choice and response flexibility, constrained online search environment, pop-up glossaries, online accommodations, oral response considerations, and automated essay and rule-based scoring. Barton and Schultz (2012) described the development process in terms of the standards, claims, targets, and evidence, and considerations in technology enhancements, accessibility, scoring methodologies, and psychometrics.

During the demonstration, Barton and Schultz (2012) described the challenges of assessing a writing standard with artificial intelligence (AI) scoring in real time. Accessibility is approached somewhat differently than in the tasks by SRI and ETS, by incorporating features for flexibility in the response. McGraw-Hill CTB has some items that are linked to the CCSS that have already been field tested, so Barton and Schultz were able to draw on these results to leverage what they have already come to know and have learned in field trials.

The McGraw-Hill CTB task employs a passage and a media presentation from the History Channel on the same topic. Students engage in both perspectives and start to organize their thoughts about both perspectives that the authors of the passage have. The task poses challenges to students about how they use this evidence to do more research and to describe their thoughts about what they are discovering.

Through the AI components, McGraw-Hill CTB hopes to enable tasks that allow respondents to speak, to write notes, and to put together four or five slides for a PowerPoint presentation. By incorporating both speaking and listening in tasks, McGraw-Hill CTB hopes to generate mini performance tasks through the task surrounds, during which students progress through a number of related tasks. The typical performance task perspective draws on not only the end product but on how the students progress through the work and the processes they use, providing information on students who have evidenced good process development.

Recommendations and Conclusion

All three of the demonstration projects and associated papers (Barton & Schultz, 2012; Cayton-Hodges et al., 2012; Haertel et al., 2012) described here seem to offer good opportunities for reflection and reaction on the current state and future practice of technology enhanced assessments. Attributes of hard-to-measure constructs that may influence adoption of some of the approaches shown are defined in this paper and a TEI framework is introduced, along with an accessibility tool that helps explore design needs in a given context.

Key aspects of measurement that all assessments must address include identifying the goals and objects of measurement, or the constructs; the observations, here through the TEIs and their templates and task surrounds; and the interpretation, which must include scoring of each actual item or observable, as well as the accumulation of this evidence into overall inferences (the measurement model), as a single bit of data is rarely sufficient for high quality measurement information.

A coherent argument needs to be made in each setting for how TEIs go about addressing each of these aspects of measurement. The three demonstrations from developers are considered here through a lens of how hard-to-measure constructs can be approached. It is expected that a myriad of additional laudable solutions are and will be appearing in the coming months and years. The hope is that through TEIs as well as other observation formats, tapping hard-to-measure constructs while maximizing accessibility and also maintaining measurement standards of high quality evidence can be achieved in the United States to help support 21st century needs for teaching and learning.

In order to do so, the following recommendations are made concerning approaches that state may wish to take regarding technology enhanced assessments:

1. High quality evidence needs to be continued to be valued as a high priority in educational assessments.
2. With TEIs, evidence should be interpretable and sufficiently aligned to instruction so that teachers and schools leaders see a connection between the evidence and how to make use of it in practice. This is true not only for instruction, but also for teacher preparation and professional development and for school accountability practices where schools will want to be responsive to evidence results by putting into place effective programmatic reforms.
3. Due to the compelling potential of technology developments, the types and kinds of evidence that can be brought to bear for educational assessment are being revolutionized at this time. States should press for adoption of innovations that provide utility while at the same time respecting the need to continue to establish high evidentiary quality in assessments, including continuing to meet usual U.S. measurement standards of reliability, validity, acceptability, fairness, and access for every student.
4. In order to do this, states can expect that in coming together, fewer unique assessment systems may be needed in the United States (there is no longer as much of a need for 51 customized assessment systems in the states and Washington, DC). However, each system will require more investment of time, thought, and dollar resources. Staffing, development timelines, and budgets should be planned accordingly to support thoughtful, shared development work.
5. This will include better understanding the domain modeling needed for CCSS as well as establishing forums where innovation can be adopted in an ongoing fashion within state assessment ecosystems and then migrated as desirable. Standardizing on any particular era of technology will not be beneficial for U.S. students and should be avoided by U.S. policy makers.
6. States should anticipate a greater interaction between so-called formative and summative evidence settings as technology enhanced assessments move forward. More discussion is needed at the state level on where, how, if and to what extent they would like to see these evidence approaches effectively come together. States should note that curriculum adoptions are likely to change radically the array of information in schools, as electronic resources become more affordable than paper-based.
7. Many of the assessment targets in the CCSS will involve more complexity due to cognitively complex domains and higher order thinking, assessments with more interactions among the

KSAs, dependence in the observations, data density, and noisier data with more construct irrelevant variance. At the same time, this is compensated for by the availability of much more evidence (data density). In order to address the first goal of more complexity, it is likely that capitalizing on the second opportunity of more information will be necessary.

8. Twenty-first century measurement (Wilson et al., 2012) thrives in generating subtle and complex inferences in the presence of abundant, rich data. If the states can effectively capture this dual opportunity, a stronger signal likely can be generated from a less simplistic view of what students need to know and be able to do. This would be a strong realization of CCSS efforts generally and could be expected to help support success in the global economy as well as empower students for a lifetime of learning and citizenship.

References

Barton, K., & Schultz, G. (2012, May). *Using technology to assess hard-to-measure constructs in the CCSS and to expand accessibility: English language arts.* Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Cayton-Hodges, G. A., Marquez, E., van Rijn, P., Keehner, M., Laitusis, C., Zapata-Rivera, D., . . . Hakkinen, M. T. (2012, May). *Technology enhanced assessments in mathematics and beyond: Strengths, challenges, and future directions.* Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Haertel, G. D., Cheng, B. H., Cameto, R., Fujii, R., Sanford, C., Rutstein, D., & Morrison, K. (2012, May). *Design and development of technology enhanced assessment tasks: Integrating evidence-centered design and universal design for learning frameworks to assess hard to measure science constructs and increase student accessibility.* Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Levy, R. (2012, May). *Psychometric advances, opportunities, and challenges for simulation-based assessment.* Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, D.C.: National Academy Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press.

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning.* Alexandria, VA: ASCD.

Rose, D. H., & Meyer, A. (2006). *A practical reader in universal design for learning.* Cambridge, MA: Harvard Educational Press.

Rose, D. H., Meyer, A., & Hitchcock, C. (2005). *The universally designed classroom: Accessible curriculum and digital technologies.* Cambridge, MA: Harvard Education Press.

Scalise, K. (2012). *Technology-enhanced to interactive to immersive assessments: Measurement challenges & and opportunities.* Irvine, CA: National Research Council Board on Testing and Assessment.

Scalise, K., & Gifford, B. R. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Teaching, Learning and Assessment, 4*(6).

Smarter Balanced Assessment Consortium. (2012). *Technology-enhanced items guidelines.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/TechnologyEnhancedItems/TechnologyEnhancedItemGuidelines.pdf

Wilson, M., Bejar, I., Scalise, K., Templin, J., Wiliam, D., & Torres Irribarra, D. (2012). Perspectives on Methodological Issues. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills*. New York, NY: Springer.

K-12 Center
at ETS

Invitational Research Symposium on

Technology Enhanced Assessments

The Center for K–12 Assessment & Performance Management at ETS creates timely events where conversations regarding new assessment challenges can take place, and publishes and disseminates the best thinking and research on the range of measurement issues facing national, state and local decision makers.