



Conducting Research on Technology-Enhanced Assessment: Lessons Learned from the Field

PAPER PRESENTED AT THE ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
CHICAGO, IL • APRIL 2015

Jessica Masters, Ph.D.

masters.jessica@measuredprogress.org

Lisa Famularo, Ph.D.

famularo.lisa@measuredprogress.org

Kristin King, MPPM

king.kristin@measuredprogress.org

Measured Progress Innovation Lab

Abstract

The Validity of Technology-Enhanced Assessment in Geometry (VTAG) project is evaluating the extent to which technology-enhanced items provide valid measurement of Common Core State Standards in elementary geometry. Using the lessons learned from the ongoing VTAG project, the authors present challenges and recommendations associated with conducting research on technology-enhanced items. The goal of this discussion is to support and guide researchers in conducting studies with the ultimate aim of building the base of research related to technology-enhanced assessment.

Technology-enhanced items have the potential to provide improved measures of student knowledge, create more engaging assessment environments, and reduce the effects of guessing and test-taking skills. For these and other reasons, the national assessment consortia and many state departments of education have developed next-generation assessment systems that use technology-enhanced (TE) items in formative and summative assessments. Despite the forward momentum and rapid adoption of technology-enhanced items (TEIs), there is not broad evidence of the validity of inference made by TEIs and the ability of TEIs to provide improved measurement. Without such research, there is no way to ensure that TEIs can effectively inform, guide, and improve the educational process.

A small number of research efforts have begun to build a base of research about TEIs, but much more is needed. The current paper presents the lessons learned from one such ongoing research effort, the Validity of Technology-Enhanced Assessment in Geometry (VTAG) project. The authors present challenges and recommendations associated with conducting research on TEIs. The authors' goal is to support and guide researchers in conducting investigations on TEIs with the ultimate aim of building the base of research related to technology-enhanced assessment.

The Validity of Technology-Enhanced Items

Assessment is a critical component within the instructional process and instruction should be differentiated based on assessment results (Pellegrino, Chudowsky, & Glaser, 2001). The 2010 National Education Technology (NET) Plan's goal related to assessment is that "Our education system at all levels will leverage the power of technology to measure what matters and use assessment data for continuous improvement (USDE, p. xvii)." Research has long documented the inadequacies of selected-response (SR) items to measure high-level knowledge and understanding (Archbald & Newmann, 1988; Bennett, 1993; Birenbaum & Tatsuoka, 1987; Hickson & Reed, 2009; Lane, 2004; Livingston, 2009; Darling-Hammond & Lieberman, 1992). One solution to the shortcomings of SR items is the use of text-entry or constructed-response (CR) items, which have frequently been used to measure high-order skills and knowledge. In recent years, researchers have leveraged technological advancements to combine the measurement power of CR items with the automated-scoring capability of SR items. One branch of this research has focused on automated text and essay scoring (e.g., Dikli, 2006), while another branch has focused on using technology to allow students to interact with digital content in innovative ways through the development of TEIs. This line of research is consistent with the NET Plan's assessment-related recommendations, which include the development of assessments that provide "new and better ways" to assess students and the expansion of the capacity to design, develop, and validate technology-enhanced assessments that can access constructs difficult to measure with traditional assessments (*ibid.*). For this recommendation to be realized, more research is needed on the validity of inferences made from technology-enhanced assessments in a variety of contexts.

TEIs offer many potential benefits over SR items. The most significant is that TEIs have the potential to provide improved measurement of certain constructs, specifically high-level constructs, because they require students to produce information, rather than simply select information, which is often a more authentic form of measurement (Archbald & Newmann, 1988; Bennett, 1999; Harlen & Crick, 2003; Huff & Sireci, 2001; Jodoin, 2003; McFarlane, Williams, & Bonnett, 2000; Sireci & Zenisky, 2006; Zenisky & Sireci, 2002). A second benefit

is that TEIs reduce the effects of test-taking skills and random guessing (Huff & Sireci, 2001). A third benefit is that TEIs have the potential to provide richer diagnostic information by recording not only the student's final response but also the interaction and thought process that lead to that response (Birenbaum & Tatsuoka, 1987). CR items have always offered these benefits, but TEIs allow these benefits to be leveraged on items administered via computer that can be automatically and instantly scored. A fourth potential benefit of TEIs is a possible reduction of cognitive load from non-relevant constructs, such as the reading load for items designed to measure mathematics or science, and the cognitive load required to keep various item constructs in memory (Mayer & Moreno, 2003). Finally, TEIs tend to be more engaging to students, an important consideration in an era when students frequently feel over-tested (Strain-Seymour, Way, & Dolan, 2009; Dolan, Goodman, Strain-Seymour, Adams, & Sethuraman, 2011).

These potential benefits have led the two federally-funded assessment consortia, the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for the Assessment of Readiness for College and Careers (PARCC), and many state departments of education, to include TEIs in summative and non-summative tests that are part of their next-generation assessment systems. In addition, advances in technology and technology interoperability standards have offered the promise of more efficient item authoring and delivery and assessments with high portability. For example, QTI (Question and Test Interoperability Standard) is an XML-based standard for developing and delivering assessment components (IMS, 2001). QTI has been widely adopted and implemented in a host of assessment and learning systems. QTI can be used to provide information about the content that forms an item, including the directions, prompt, stimuli, response area, response options, etc. Items that are QTI-compliant can be delivered through any QTI-based assessment or learning system. As an open, non-proprietary standard that is employed within many assessment systems, QTI provides a powerful tool to support the efficient development of interoperable items. The availability of such technology has further encouraged the development and use of TEIs.

Despite this forward momentum to develop and use TEIs, there is only a small research base evaluating the validity of TEIs in various contexts within K-12 education. One of the earliest published studies involved cognitive labs with elementary, middle, and high school students to evaluate perceptions of TEIs, the cognitive processes used to respond to TEIs, and the potential for TEIs to better evaluate constructs in both mathematics and English language arts (Dolan, Goodman, Strain-Seymour, Adams, & Sethuraman, 2011). Although the results cannot be broadly generalized because of small sample sizes, the research found preliminary evidence to suggest that TEIs are highly usable and engaging. More importantly, the research found that TEIs can measure constructs that are not easily measured with traditional item types, particularly high-level constructs. The study found that the use of TEIs reduced guessing and allowed students to have more authentic interactions with content. The study also found that TEIs required more time to complete and that this factor was influenced by students' technical proficiencies (*ibid.*).

In another research effort, researchers evaluated and compared the performance of TE, SR, and CR items in the context of fifth grade, eighth grade, and high school science (Wan & Henley, 2012). This study explored TEIs of the figural response type, which includes hot spot identification, drag-and-drop, and reordering. Through item response theory analyses, this study found that TEIs provided the same amount of information as SR items in fifth and eighth grade, and slightly more information in high school. While CR items provided more information than both TE and SR items, those items required human scoring. This study also found that TE and

SR items were equally efficient (e.g., provided the same average amount of information in an average amount of time). The researchers concluded that their statistical analyses supported the use of TEIs in K-12. However, they were careful to note that further psychological testing (e.g., cognitive labs) should be conducted to confirm the results of their statistical analyses. The researchers also advised caution in using TEIs when standard SR items are able to measure a construct. “We reviewed the test forms administered in this study and found that a number of innovative items could be easily replaced by [multiple choice] items without changing the [knowledge, skills, and abilities] measured. This issue is not uncommon in the development of innovative assessments: the face validity of innovative item formats is so appealing that their real potential of providing something beyond what is available using the [multiple choice] format may be overlooked (ibid., p. 74).” This claim is supported by other researchers who warn of innovation solely for innovation’s sake, when traditional items can fully and accurately measure a construct (Haldyna, 1999; Sireci & Zenisky, 2006). The researchers recommended both psychometric and psychological research to determine when TEIs are appropriate.

More recently, a group of researchers explored SR, CR, and TE items in the context of seventh grade mathematics and Algebra I. The researchers found that a test comprised of CR and TE items, while aligned to the Common Core standards, did not have a significantly higher correlation to teacher ratings of student knowledge than a test comprised of SR items. The CR/TE test was reviewed by experts and found to be similar to the SR test in terms of measuring the intent of the standards and the depth of knowledge. The CR/TE test was found to be significantly more reliable and to provide more information than the SR test (Winter, Wood, Lottridge, Hughes, & Walker, 2012). “These results indicate that tests incorporating CR/TE items can measure some mathematics content with less error than tests comprising only [SR] items (ibid, p. 53).” This study provides promising results, but must be generalized cautiously because of the narrow content focus and the blending of CR and TE items on the same test form.

The VTAG Project

The VTAG project contributes to this small but critical base of research related to the validity of TE items. The VTAG project differs from previous efforts and makes new contributions to the research base by using a combination of cognitive labs with small samples and statistical analysis of field test data from larger samples. The VTAG project focuses on a broad content area not previously studied: elementary geometry. The VTAG project addresses three research questions:

RQ1: To what extent are TEIs a valid measurement of geometry standards in the elementary grades?

RQ2: To what extent do TEIs provide an improved measurement compared to SR items?

RQ3: What are the general characteristics of mathematics standards that might be better measured through TEIs?

To date, the VTAG researchers have developed parallel sets of SR and TE items targeting the Common Core Grades 4 and 5 Geometry standards. These items have undergone expert review by an external advisory panel and will next be used in cognitive labs and field testing. Through the process of writing, revising, and digitizing the items, the authors have encountered a number of challenges that they believe are common to any project that includes research related to TEIs. The authors summarize these challenges, describe the approach taken in the VTAG project, and provide considerations and recommendations for other researchers. More research is needed on the validity of TEIs and the goal of this paper is to support other researchers in those efforts.

Challenge: There is no universal definition of TEI.

There is currently no universally accepted definition for what constitutes a technology-enhanced item. By definition, a TEI is an item that uses technology in some way to *enhance* the item. Some items, such as traditional multiple choice items (e.g., Figure 1) are clearly not TE. Even if a traditional multiple choice item is delivered and scored by computer, it is generally accepted that this would not be considered a TEI. Other items, such as an item that asks a student to draw a response (e.g., Figure 2) are generally accepted as TE. However, there are many items for which the categorization is not as clear.

For example, the item in Figure 3a includes a multimedia video paired with a multiple choice item. The items in Figure 3b and Figure 3c use a drag-and-drop interaction. Are any of these items technology-enhanced? This seemingly simple question is actually quite complex and the answer can differ based on a variety of factors and perspectives. How and to what extent does the item rely on technology? How does the student interact with the item to submit his/her response? How is the item scored? How does the use of technology affect the ability to elicit the construct?

The answer to the original question (“Is this item a TEI?”) will vary based on the perspective of the researcher and the nature of the research questions. For example, a technology-focused researcher might consider all of the items in Figure 3 to be TEIs. The first item requires a video that cannot be delivered without technology and the second and third items use an interactive drag-and-drop interface (the QTI interaction called *graphic gap match*). A researcher investigating user experience might consider the second and third item to be TEIs because they require more interaction when submitting a response. An assessment researcher might consider only the third item to be TE because the first two items are essentially multiple choice items where the student selects one out of two or four pre-defined options.

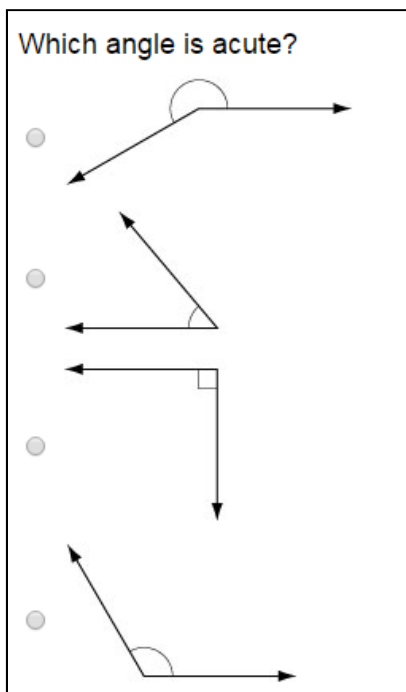


Figure 1: Sample VTAG SR Item

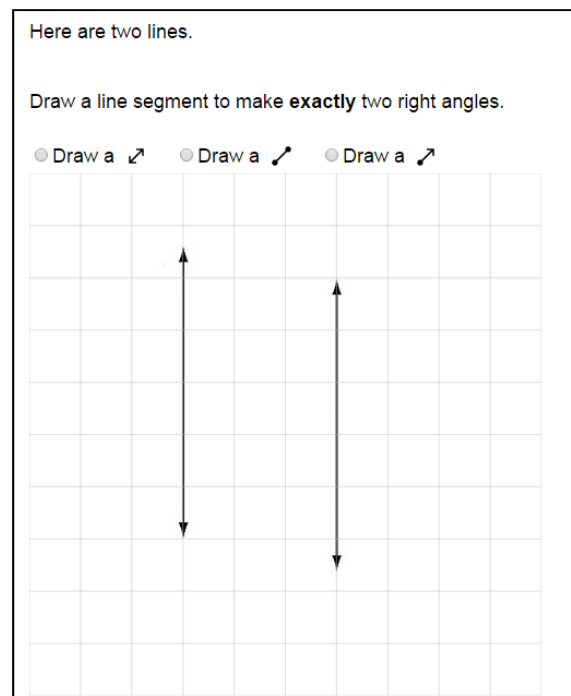
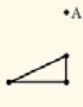


Figure 2: Sample VTAG TEI

Watch the following movie. Decide if it shows the correct way to rotate (turn) the triangle 360 degrees counterclockwise around point A.

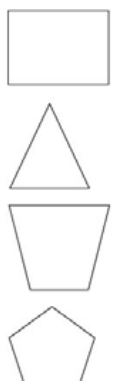
▶



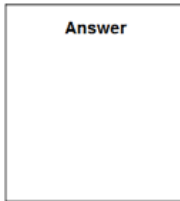
Does the triangle rotate correctly?

(a)

Place the figure that appears to have only right angles in the answer space.




Answer



(b)

Drag each shape into every column to which it belongs. Some shapes may belong in more than one column. Some shapes might not belong in any column.

Polygons	Quadrilaterals	Parallelograms



(c)

Figure 3: Sample Item with Varying Uses of Technology

The current paper focuses primarily on the assessment perspective, as this is the type of research the authors aim to support. The example above uses *level of constraint* to define what constitutes a TEI within this assessment perspective. Scalise and Gifford (2006) defined a taxonomy of items to that can be used to specify an item’s type (Figure 4). This taxonomy organizes items by the *degree of constraint* placed on the student’s options for responding to or interacting with the item. The taxonomy does *not* classify items based on the method or mode of interaction required by the student or the media included in the item. This taxonomy is the framework used in the VTAG project.

	Intermediate Constraint Item Types					Fully Constructed	
	1. Multiple Choice	2. Selection/ Identification	3. Reordering/ Rearrangement	4. Substitution/ Correction	5. Completion	6. Construction	7. Presentation/ Portfolio
Less Complex	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)
More Complex	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 4: Taxonomy of Item Types based on Level of Constraint (Scalise & Gifford, 2006, p. 9)

While the level of constraint framework provides a robust structure to discuss item types, it does not resolve all of the practical challenges of item classification, nor will any framework. For example, whether or not an item is a TEI is a binary outcome. Thus, a threshold must be set within a framework such that item types below that threshold are not TE and item types above the threshold are TE. Even once a threshold is defined, some items will not clearly fall above or below the threshold. Different members of the research team might disagree on such cases near the border of the threshold.

Additional challenges in how to define a TEI relate to text-entry items and automated scoring. There is a wealth of research on the automated scoring of essays and constructed-response items. Should an open- or constructed-response item that is scored with an automated essay scoring engine be considered technology-enhanced? If using the framework of level of constraint, these items would clearly be TE. But the research related to how essays and

constructed-response items measure knowledge and how to score these items is quite distinct from research on other types of items. Should these automatically-scored essay items be grouped in the same category as other TEIs such as drag-and-drop items? There are not yet generally accepted guidelines in the field to answer these questions.

The VTAG Approach: The VTAG project is focused on the extent to which different item types measure student knowledge. Because this is a research question related specifically to measurement (as opposed to, for example, a research question related to the method of interaction), the project adopted the level of constraint framework to specify what constitutes a TEI. Based on this framework, it is the degree of constraint put on a response, not the method of interaction, that determines whether an item is considered SR or TE.

For example, Figure 5 shows the use of the QTI hotspot interaction where the student clicks on graphical content to select the content. While hotspot items can often be thought of TE, this item is classified as SR because it places a high level of constraint on the student's response (i.e., the item presents the student with a set of five discrete response options, each of which can be selected). How the student interacts with the item is different than a traditional multiple choice item (they click to circle a picture rather than selecting from a list of A, B, C, etc.), but the degree of constraint is the same. Figure 6 also uses the QTI hotspot interaction but this item is considered TE. Technically, this item could also be re-written using a list of all combinations of sides and angles, and presented as a multiple-choice item, but this would be unwieldy to students. Because the responses to this item are less constrained, it is considered TE.

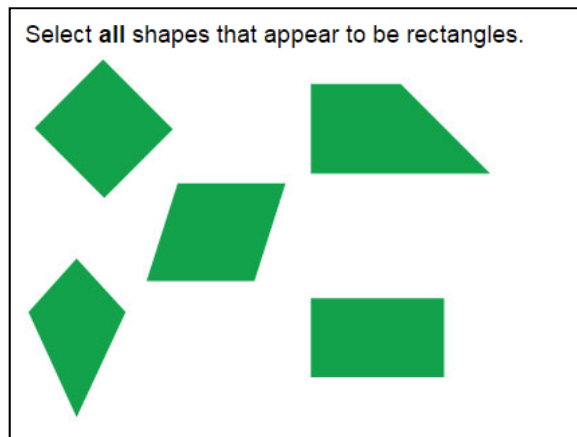


Figure 5: Sample VTAG SR Item (using QTI Hotspot Interaction)

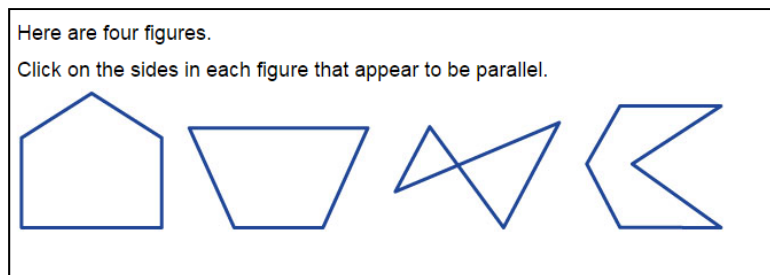


Figure 6: Sample VTAG TEI (using QTI Hotspot Interaction)

The VTAG project considered text-entry items as distinct from TEIs and hence excluded them from the study. The VTAG project set the threshold of SR vs. TEI at approximately 3A/3B in the taxonomy shown in Figure 4. Items from 1A to 3A are considered SR and items from 3B to 6B are considered TEI. Some items were borderline. For example, consider the item displayed in Figure 7. Should this item be considered SR or TE? The researchers classified this item as SR because it is essentially a composite of five multiple choice items, each with three options. The item shown in Figure 8 was classified as TE because while technically this item could be rewritten as a multiple choice item by enumerating all of the possible combinations, that would be prohibitively impractical. These kinds of drag-and-drop labeling or classification items seemed to frequently be the kinds of items that were difficult to classify as SR or TE. The researchers generalized that if one of the draggable pieces of content belonged in more than one bucket (i.e., the draggable content could be “tiled”) or at least one piece of content did not belong in any bucket, then that item would be classified as TE. If neither of these cases held, then the item would be classified as SR. This generalization meant that some items classified as TE could technically be rewritten as a series of multiple choice items with all of the possibilities enumerated, like the item in Figure 8, but the researchers considered the vast number of possibilities sufficient to justify classifying these items as TE.

These examples provide information about decisions made for one particular project whose research questions use an assessment perspective and the level of constraint framework to specify what constitutes a TEI. More importantly, they serve to highlight the subjective and complex nature of defining what constitutes a TEI.

Here are some letters. Decide whether each letter has

- a horizontal line of symmetry, ←-----→
- a vertical line of symmetry, ↑-----↓ or
- no line of symmetry.

Match each letter to the words that tell about its symmetry by connecting the correct word label to the box by each letter. Each label may be used more than once.

A

B

C

D

F

Vertical line of symmetry Horizontal line of symmetry No line of symmetry

Figure 7: Sample VTAG SR Item on the Borderline of the TEI Threshold

Here are some letters. Decide whether each letter has

- a vertical line of symmetry, ↑-----↓,
- a horizontal line of symmetry ←-----→, or
- no line of symmetry.

Drag each letter to the correct answer space. Some letters may be dragged to more than one answer space.

Vertical Line of Symmetry	Horizontal Line of Symmetry	No Line of Symmetry

L H I K M N X

Figure 8: Sample VTAG TEI on the Borderline of the TEI Threshold

Recommendations for Researchers:

- ✓ Be explicit in how you define a TEI for a given project.
- ✓ Select a single framework to classify items. The framework should be chosen based on the nature of the research questions. For example, the research might be focused on measurement characteristics of items, the use of technology to deliver items, the user experience of interacting with items, etc.
- ✓ Within the framework selected, set a threshold for what constitutes a TEI. Be explicit about items that fall on the borders of the threshold.
- ✓ Consider the differing perspectives of your research team and provide multiple exemplar items of what constitutes a TEI based on the nature of the research questions and the framework adopted to ensure all team members have a common understanding.

Challenge: There are multiple open-source platforms for authoring, delivery, scoring.

In recent years, a variety of open-source platforms have become available for use in authoring, delivering, and scoring assessments, including assessments with more interactive items. This is generally a very positive development. However, because these platforms are new, their features are often limited, they are subject to bugs, and they are continually undergoing development. Thus, a given platform can change significantly over the course of a research project. Because the platforms are open-source, support can be limited, particularly for researchers with limited technical expertise. There are advantages and disadvantages of each available platform and limits to how well the platforms interact with each other.

Many of the open-source platforms currently available use QTI. While there are other options for how to implement TEIs (e.g., Flash), QTI provides a standardized language for both the delivery and scoring of items and has been gaining acceptance in the field as the preferred solution. QTI offers the promise of portable items that can be authored and delivered through any QTI-compliant engine. Unfortunately, as is so often the case, the reality is messier. QTI is an exacting specification, and as such is prone to errors in porting work across platforms. Moreover, items authored in one platform may look visually different when delivered by another platform. Identifying and resolving problems or errors in QTI or between platforms takes time and technical savvy, a challenge that is compounded if a research team that consists mainly of people without extensive technical expertise is tasked with resolving these issues.

The VTAG Approach: VTAG has adopted two different open-source platforms, one for authoring and another for delivery and scoring. There have been challenges in porting items across the two platforms. Often, items authored in the first platform look different when delivered by the second. Further, the authoring system offers a limited set of interaction types, which required the development of custom interaction types. It was discovered late in the process that the authoring platform could not import items that use this custom interaction type. Thus, assembling test packages to export to the delivery platform requires manual manipulation of the files exported from the authoring system.

Recommendations for Researchers:

Choosing a Platform:

- ✓ Invest time at the beginning of the project to evaluate the available assessment platforms.
 - Evaluate all components of available platforms prior to adoption.
 - The evaluation should be a collaborative effort among the researchers and the technical team supporting the research effort.
 - Develop a comprehensive list of criteria to guide the evaluation. The criteria should describe the requirements for authoring, delivery, scoring, and reporting.
- ✓ Ensure that the technical team has a strong understanding of the research questions, the framework being used to classify TEIs, and the desired functionality of items and the assessment in order to advise on which platform will best suit the needs of the project. The researchers must provide the following types of information to the technical team:
 - What item types are required?
 - Consider creating a mapping between the classification of items for research purposes and the implementations supported by the platforms.
 - Some item types can be implemented in different ways. If one specific interface or method of student interaction is required for an item type, ensure that the platform supports that or budget for customization.
 - How will items be scored (e.g. partial credit, polytomous, etc.)? What is the test design (e.g., randomization, skipping of items, linear vs. non-linear, adaptive items, provision of feedback, etc.)?
 - What are the characteristics of students in your sample?
 - What is the expected technical fluency of the students or other users?
 - What is the age and other characteristics of the students or other users?
 - Will students who typically use accessibility supports during assessment (e.g. students with low vision, print disabilities, motor impairments, etc.) participate in your research? What accessibility tools are required (e.g. magnification, color contrast, tab/enter navigation of the interface)? How might these tools impact the delivery of TEIs and students' ability access and respond to items?
 - What are the data and reporting requirements (e.g. time on task, response history or final response only, raw response data, scored response data, etc.)?
- ✓ The technical team must ensure that the researchers understand the strengths and weaknesses of platforms from a technical perspective and the implications of weaknesses.
 - For example, some open-source tools have an active community of developers who contribute code and engage with one another to problem solve. The extent to which there is an active community of developers can affect the time it takes to solve unanticipated problems or customize features. This is an example of a technical consideration that researchers might not otherwise consider when evaluating platforms.
- ✓ Consider what support software will be required by the delivery platform (e.g., Flash, Java, etc.) and be prepared to guide teachers in installing the necessary software.

- ✓ Recognize that different phases of the research might present conflicting priorities. It is possible that one platform will better satisfy the requirements for one phase while another platform will better satisfy the requirements for a different phase. But be cautious in choosing to use different platforms for different phases of the research.
 - For example, an authoring platform might provide the most robust support for authoring various item types, but the types of data required for analysis might not be collected by items authored in that platform.
 - In these cases, carefully weigh the potential challenges of dealing with platform interoperability against the cost required to use a single platform and build out its functionality for a particular phase.

Using the Chosen Platform:

- ✓ Staff and budget appropriately. If using a new platform or modifying an existing platform to author, deliver, and score TEIs, plan for a technical team to support the research. Ensure that either a member of the research team has the requisite technical expertise or budget for the time needed for the researchers to learn how to use the technical tools.
- ✓ Different platforms treat graphical content differently. The research team must include expertise in both the technical details of the platform and in design in order to properly use graphics (including making decisions about size, color, fill, formatting, etc.).
- ✓ Anticipate that any open-source platform will require some customization. It is unlikely that an existing system will be wholly sufficient for research needs. Budget accordingly.
- ✓ Given the limitations of working with relatively new and often unsupported platforms, producing an assessment exactly as planned may be unrealistic within the project's budget and timeline. Be prepared to make concessions.
- ✓ Develop an orientation or student practice test so that students can gain practice and experience with the testing platform prior to taking an assessment whose data will be used for research purposes. This will alleviate the risk of the platform and interface introducing construct-irrelevant factors.
 - In some research contexts, you may want teachers to aid students who have difficulty with the technology enhanced features of the test. To this end, consider making the orientation or practice test available to teachers in advance so that they can familiarize themselves with the interface and tools.

Challenge: There are a variety of browsers and devices used in education.

Browser and device compatibility is a challenge with any research that involves technology. The relative newness of TEIs heightens this challenge. For example, browser compatibility for traditional websites is a challenge that has been largely resolved over time. Hopefully compatibility issues will be similarly resolved for assessments, but currently the challenge persists. A single delivery engine can render the same item very differently in different browsers or on different devices. Some renderings make it impossible to respond to the item. For example, the item in Figure 9 shows a drag-and-drop item that rendered correctly in one browser but showed overlapping content in another. The displayed version makes it impossible to respond to the item. Further complicating browser compatibility issues are presented when support software for the selected platform fails to function properly or requires the user to complete a series of (sometimes complex) steps before functioning in some or all browser and device combinations.

Match the figure to the total number of lines of symmetry it has.

Triangle with all sides the same length	<input type="text"/>
Square with all sides the same length	<input type="text"/>
Right triangle with no sides the same length	<input type="text"/>
Rectangle with all sides the same length	<input type="text"/>

0 lines of symmetry 3 lines of symmetry 2 lines of symmetry 4 lines of symmetry

Figure 9: Sample VTAG TEI with Improper Rendering in One Browser

An easy solution to this challenge would be to require and support only a single browser/device combination. This is an insufficient solution, however, because it does not reflect how technology is used in reality. Further, as educational researchers are keenly aware, recruiting sufficient numbers of teachers to participate in research is its own challenge. Restricting the pool of available teachers by restricting the method through which they access an assessment would prove deleterious to the research and would introduce an irrelevant and confounding factor into the sample.

The VTAG Approach: The VTAG project will support Chrome, Safari, and Internet Explorer versions 10 and higher and will not include mobile support. The platforms used in VTAG require the installation of a Java applet. The project will include detailed instructions for teachers on how to install the applet on the three supported browsers.

Recommendations for Researchers:

- ✓ Prepare detailed instructions for using the software on each supported browser.
- ✓ Budget for testing all items on all supported browsers. As much as possible, test *each item on each browser*, as items with the same item type and implementation can render differently in the same browser.
- ✓ Find a balance between supporting enough browsers and devices to allow for broad participation but not so many that item testing becomes too onerous.
- ✓ Consider carefully whether mobile support is required for the research. If mobile support is required, budget accordingly.

Challenge: Technology is flashy and fun!

There is a temptation to write and use TEIs because they are new and exciting. This is often described as “using technology for technology’s sake” and it can be a common pitfall in assessment and research design. If an item is being used to capture student knowledge, then the item type used should be driven by the content: i.e., what is the best way to elicit knowledge of the construct? In other research contexts, the item type might be driven by other research questions (e.g., are tech-savvy students more comfortable with drag-and-drop items versus traditional multiple choice selection?). In any project, the use of TEIs should be driven by the research questions. It’s easy to be swept up in new technology, but this can lead to using items that are not appropriate for the intended purpose.

The VTAG Approach: Because the VTAG project is exploring the measurement properties of items, the researchers were careful to ask the content experts writing the items to think first about how best to elicit student knowledge and then only choose an item type or interaction accordingly. As part of external item review, the researchers asked expert reviewers to identify and provide feedback on items that appeared to use technology for technology's sake, rather than as a way to measure the construct.

Recommendations for Researchers:

- ✓ Think carefully about why you are using a TEI. There are many different reasons to use TEIs, including to provide a better measure of knowledge and to engage students in an assessment. Researchers should have *some* reason for using TEIs and avoid using them only because they are new and exciting.

Challenge: Automated scoring relies on human-defined algorithms.

One of the benefits of TEIs is that these items can be automatically scored. This removes the subjective nature of human scoring and enables instant feedback to be provided at a reduced cost. However, the scoring algorithms used by computers are written by humans. Items that require polytomous or partial credit scoring can be complex to define. Most of the current open-source platforms provide standard scoring for traditional items, but less robust support for more complex scoring. Thus, scoring items in this way currently requires a concerted effort and extensive technical expertise. Many researchers are familiar with the complexities of writing a rubric for human scoring. Translating that rubric into a computer-interpretable definition adds an additional layer of complexity.

Automated scoring also requires specifying acceptable tolerances around correct and partial-credit responses. Consider an item asking a student to plot a point at (1,2). In a paper-based test, a human scorer would likely subjectively decide that a point at (0.99, 1.99) should be scored as correct. In fact, the human scorer might not even notice that this point was different from the correct response. This same item administered as a TEI requires pre-specified tolerances around response options. A computer needs to be told that (0.99, 1.99) is correct. What the acceptable tolerances should be depends on many factors, including the age and technical acuity of the students and the devices likely to be used to enter responses. Further, tolerances around responses must be defined using consideration of how students with motor disabilities might be able to respond to the item.

The VTAG Approach: Because this project is focused on evaluating and comparing the measurement properties of items, it was critical to get the best estimate of student understanding. Because of this, the researchers felt it was necessary to allow for partial credit scoring. A large effort was required to develop partial scoring algorithms for the TEIs. For example, consider the item in Figure 10. The rubric for human-scored partial credit is: 4/5 parallelograms correctly classified (demonstrates understanding of parallelogram); or 2/2 rhombi correctly classified (demonstrates understanding of rhombus); or 5/7 parallelograms and rhombi correctly classified with at most one of these shapes incorrectly classified (demonstrates partial understanding). The QTI to specify the scoring for this item required over 2,800 lines of properly formatted XML because, for example, the various combinations of ways that four out of five parallelograms could be correctly classified must be explicitly listed.

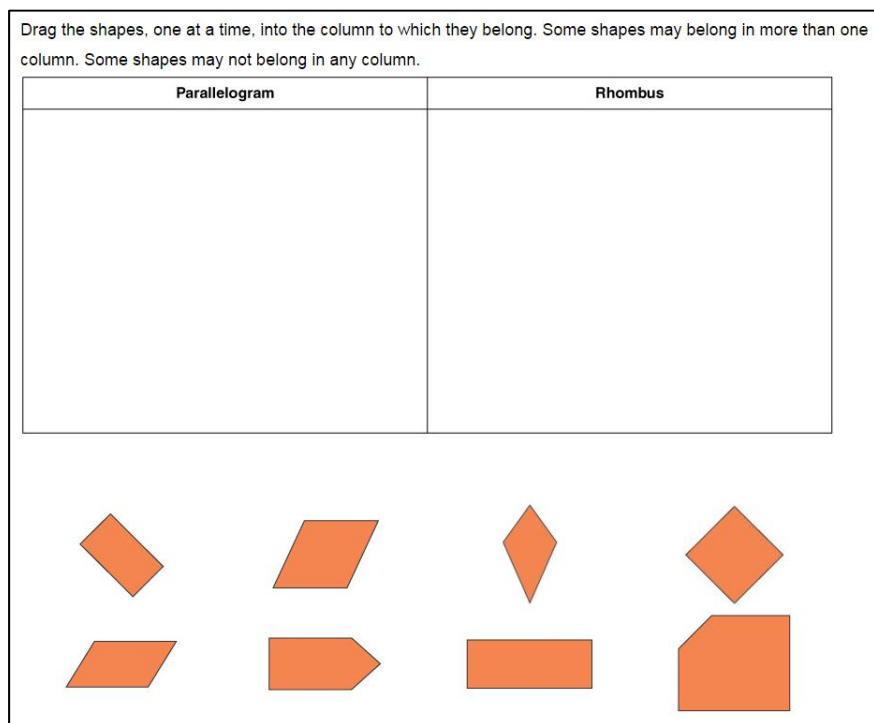


Figure 10: Sample VTAG Item with Partial Credit

Because of the young age of students participating in the research and the infeasibility of field testing due to the project scope, the researchers often used snap-to interfaces (a point placed between two lines on a grid will align or "snap to" the nearest intersection of lines) rather than setting tolerances. The researchers acknowledge that there are trade-offs to this decision.

Recommendations for Researchers:

- ✓ Think carefully about whether partial credit or polytomous scoring is required. These scoring algorithms are more complex than dichotomous scoring. If using more complex scoring algorithms, budget for the increased time required to develop and test the scoring.
- ✓ Consider the research participants when setting thresholds for tolerance. In addition to the current research participants, consider future students who might use or benefit from the TEIs designed for the research.
 - What ages of students will be interacting with the items?
 - What technology will students use (e.g., a mouse vs. touch screen)?
 - Will your sample include students with disabilities?
- ✓ When possible, conduct field testing to determine appropriate tolerances.

Designing Research Studies on TEIs: The next frontier in assessment research.

A small number of research efforts have begun to build a base of research about TEIs, but much more is needed. Considering the current ubiquity of TEIs on both formative and summative assessments, this need is urgent. Given the current dearth of research, there are innumerable research questions that would make a meaningful contribution to the field of education. From an assessment perspective, the most critical need is to gather evidence of the

validity of inference made by TEIs and the ability of TEIs to provide improved measurement over other item types. Without this research, there is no way to ensure that TEIs can effectively inform, guide, and improve the educational process. Substantial and valid critiques have already emerged about the TEIs currently being deployed in high stakes environments (e.g., Rasmussen, 2015). It is the authors' contention that these criticisms could have been avoided if validity research preempted the vast deployment of TEIs. It is not too late, however, to conduct high-quality research and revise current assessments based on the findings.

The current paper has presented some of the practical challenges that researchers are likely to face when conducting validity and other research related to TEIs. As noted throughout the paper, the response to these practical challenges should be motivated by the project's research questions. As is the case with any research project, it is critical to have carefully considered and clearly specified research questions and hypotheses, as these are the key drivers for determining the methods that will be employed. It is not possible to think through all of the details and implementation decisions in the early study design stages. Thus, when unanticipated issues arise and decisions need to be made, researchers should turn to their research questions to inform those decisions. It is often seemingly small details and decisions that, if not considered in light of the research questions, overall objectives, and context of the research, could lead to flaws, limitations, and criticisms of the study. This is the case for *any* research project. However it bears repeating in this context because of the relative newness of TEI research. In more established lines of research, there is often precedence and histories that researchers can review when making practical implementation decisions. Because that is not the case for TEI research, it is critical for researchers to make decisions based on their research questions and document the rationale for such decisions.

Researchers seeking to make comparisons between TEIs and other item types must take several factors into consideration when designing their research. In addition to deciding what types of traditional and TE items to study, consideration must be given to what item, test, and interface features will yield a fair comparison for a stated purpose. For example, in a study comparing student performance on CR and TE items, the researcher needs to decide whether the CR items should be administered on paper and pencil or via computer. If the CR items are delivered on computer, to what extent, if at all, should the computer-based environment mimic the paper and pencil test? Ultimately, as always, these decisions should be based on the objectives and context for the research and should be appropriately documented.

The possibilities embodied in TEIs are arguably both their greatest value and greatest challenge. Researchers have always faced a host of decisions about assessment design. This was true for paper-and-pencil assessments as well as computer-delivered assessments consisting only of traditional items. TEI research requires those same decisions and also introduces a myriad of new choices. Among the most important decisions facing TEI researchers is the pairing of a construct to an appropriate technology-enhanced interface. What is the best interface for an item targeting a particular construct and how is the fit of a construct-interface pairing to be judged? To what extent is the method of interaction between the student and the item relevant to the construct being measured? This is just one example of a question that assessment researchers always considered, but technology now offers an ever-increasing number of possibilities.

Other questions faced by TEI researchers are similarly magnified versions of traditional research questions. For example, researchers must consider the role of randomization. In paper and pencil testing, researchers often used a small number of test forms with different item orders. It would have been unfeasible to randomize each individual test. Technology allows for each

student's test to be randomized and for the response options within items to be randomized. There are many more options for the researcher to consider. As always, the researcher should return to the research questions to make these decisions.

Some of the decisions unique to the design of TEIs have a direct impact on the ability of the item to measure a construct. For example, consider an item that asks a student to draw a line from a given equation. On a paper and pencil item, the student would be free to draw *anything* on the provided paper. With a TEI, the student might have access to a variety of drawing tools (e.g., free draw, line segment, ray, angle, line, etc.) or only a tool that allows only the drawing of a line. The points drawn by the student can be left unrestricted and able to be placed anywhere or instead can be made to snap-to the nearest grid intersection. There can be a limit set to the number of lines drawn or not. When faced with these types of decisions, researchers must first consider the content the item is designed to measure and whether each feature of the environment is appropriate for that purpose. For some items, the snap-to feature, for example, might violate measurement of the construct. Researchers must consider the extent to which the technology-enhanced features of an item cue or otherwise lead students to a correct response. Researchers must also use the technology with the care needed to avoid inadvertently obfuscating a mathematically correct response. Each of these seemingly small decisions, when combined, can have a significant effect on the outcome of the research. Thus it is critical for researchers to carefully consider and document the decisions made. What at first glance might seem like a small detail can have large implications in light of the research questions.

Conclusion

Over the last decade, groundbreaking work has been done to push the boundaries of innovation in assessment in areas such as test development, delivery, scoring, and reporting. Technological innovation in each of these areas is prompted by a common goal: to improve learning through improved assessments. However, during this period of innovation, most large scale standardized K-12 tests continued to be delivered to students in paper test booklets and bubble sheets. As states transition their assessments from paper to computer, educational practitioners and policy makers need to know whether technological innovations such as TEIs result in valid inferences about student learning and whether they provide improved measurement over other more traditional item types. To provide this information, researchers must consider what kinds of items to compare, how to compare them, and to what end.

The field of K-12 assessment is in need of research focused on the validity of inferences made by TEIs and the ability of TEIs to provide improved measurement over other item types. To encourage and support researchers to contribute to the base of research related to technology-enhanced assessment, the authors offers insight into some of the practical challenges that researchers are likely to face when conducting these studies. While this validity and measurement evidence is most critically needed, researchers should consider other investigations that would make a meaningful contribution to the knowledge base and further advance innovation in assessment. Long gone is the notion of delivering paper-based assessments via computer. To fully realize the potential of technology-enhanced assessment, researchers must create and empirically investigate that which was not conceivable in paper-based testing and continue to explore the myriad possibilities that technology offers to measure student understandings.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1316557.

References Cited

- Archbald, D. A. & Newmann, F. M. (1988). *Beyond Standardized Testing: Assessing Authentic Academic Achievement in Secondary School*. Reston, VA: National Association of Secondary School Principals.
- Bennett, R. E. (1993). On the meaning of constructed response. In R. E. Bennett and W. C. Ward, (Eds.), *Construction versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (pp. 1–27). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, 18(3), 5-12.
- Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats. It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11, 385-395.
- Darling-Hammond, L. & Lieberman, A. (1992). The shortcomings of standardized tests. *The Chronicle of Higher Education*, B1-B2.
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, 5(1).
- Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive Lab Evaluation of Innovative Items in Mathematics and English Language Arts Assessment of Elementary, Middle, and High School Students*. Iowa City, IA: Pearson Education.
- Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Erlbaum.
- Harlen, W. & Crick, R. D. (2003). *A Systematic Review of the Impact on Students and Teachers of the Use of ICT for Assessment of Creative and Critical Thinking Skills*. London: Institute of Education, University of London.
- Hickson, S., & Reed, W.R. (2009) Do constructed-response and multiple-choice questions measure the same thing? Department of Economics. Working Paper Series. Retrieved March, 2015 from <http://hdl.handle.net/10092/2465>.
- Huff, K. L. & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*. 20(3), 16-25.
- IMS Global Learning Consortium. (2001). IMS Question and Test Interoperability Specification. Retrieved March, 2015 from <http://www.imsglobal.org/question/>.
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1-15.
- Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, 23(3), 6-14.
- Livingston, S. (2009, September). Constructed-response test questions: Why we use them; how we score them. *ETS R&D Connections*, no. 11.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, 43–52.

- McFarlane, A., Williams, J. M., & Bonnett, M. (2000). Assessment and multimedia authoring- a tool for externalizing understanding. *Journal of Computer Assisted Learning, 16*, 201-212.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: The National Academy Press.
- Rasmussen, S. (2015). The Smarter Balanced Common Core Mathematics Tests are Fatally Flawed and Should Not Be Used. Retrieved March, 2105 from <http://mathedconsulting.com/>.
- Scalise, K. & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment, 4*(6).
- Sireci, S. G. & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329-347). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Strain-Seymour, E., Way, W., & Dolan, R. (2009). *Strategies and Processes for Developing Innovative Items in Large-Scale Assessments*. Iowa City, IA: Pearson Education.
- U.S. Department of Education. (2010). *Transforming American Education: Learning Powered by Technology. National Educational Technology Plan 2010.* Washington, D. C.
- Wan, L. & Henly, G. A. (2012): Measurement properties of two innovative item formats in a computer-based test, *Applied Measurement in Education, 25*(1), 58-78.
- Winter, P. C., Wood, S. W., Lottridge, S. M., Hughes, T. B., & Walker, T. E. (2012). The utility of online mathematics constructed-response items: Maintaining important mathematics in state assessments and providing appropriate access to students. Final research report. Pacific Metrics Corporation. Retrieved March, 2015, from <http://www.pacificmetrics.com/files/OMAP/omap%20final%20research%20report%20body.pdf>.
- Zenisky, A. L. & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362.