Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 22, Number 1, February 2017 ISSN 1531-7714

Developing a Strategy for Using Technology-Enhanced Items in Large-Scale Standardized Tests

William Bryant, *BetterRhetor Resources LLC*

As large-scale standardized tests move from paper-based to computer-based delivery, opportunities arise for test developers to make use of items beyond traditional selected and constructed response types. Technology-enhanced items (TEIs) have the potential to provide advantages over conventional items, including broadening construct measurement, increasing measurement opportunities, and improving test-taker engagement. However, TEIs also come with some potential disadvantages, including difficulty in determining with precision what it is they measure beyond conventional items, if anything. This paper examines TEIs in light of the need by test makers to develop an evidence-based argument for their use. It offers some guiding questions and considerations toward the creation of a coherent strategy for incorporating TEIs into large-scale assessments.

Large-scale, high-stakes standardized tests in education, such as the Smarter Balanced and PARCC assessments, increasingly make use of computer-based delivery platforms. These platforms create opportunities for adding new types of items -- technology-enhanced items (TEIs) -- to the conventional mix of selected and constructed response item types. Test makers are motivated to incorporate TEIs into their assessments in answer to criticisms of multiple-choice-only tests (Martinez, 1999; Muckle, 2012; Sireci & Zenisky, 2006), and in response to market desires for machine-scorable items that are more authentic, engaging, and demanding (Florida Department of Education, 2010; Washington State, 2010). There is a broad faith among test developers and test consumers alike that TEIs add such value, but the exact nature of that added value, if it exists, is difficult to verify and describe (Dolan, et al, 2011; Parshall, et al, 2002; Scalise & Gifford, 2006).

Generally speaking, test makers have not developed or deployed TEIs according to any coherent strategy or with a confident understanding of the value they add, if any, compared with conventional multiple choice (MC) and constructed response (CR) items (Huff & Sireci, 2001; Parshall & Harmes, 2008; Russell, 2016). They have incorporated TEIs into assessments without a clear grasp of their differences, measurement implications, cost-benefit tradeoffs, or effects on test takers.

This paper is aimed at identifying key issues that should be considered as decisions are made about the use of TEIs in large-scale standardized tests. It proceeds under the assumption that decisions about their use should be justified by evidence-based arguments that identify the impact of TEIs on such factors as cost, measurement validity, and test-taker experience.

This examination confines itself to TEIs that appear on large-scale, summative, standardized assessments. There are, of course, other contexts in which TEIs are developed and deployed, such as interim and formative assessments. These contexts,

however, are not necessarily constrained by the same considerations of timing, accessibility, security, standards alignment, cost, and other factors, compared with large-scale, summative standardized tests.

Overview of Technology-Enhanced Items

Potential Advantages and Limitations

TEIs, if they are well-designed, can provide advantages over conventional MCs and CRs (Boyle & Hutchinson, 2009; Jodoin, 2003; Kane, 2006; Parshall, Harmes, Davey & Pashley, 2010; Tarrant, Knierim, Hayes & Ware, 2006). They can:

- Broaden construct measurement;
- Present more authentic contexts for the demonstration of skills and knowledge;
- Reduce the effects of random guessing;
- Reduce construct irrelevance;
- Increase measurement opportunities;
- Facilitate time- and cost-efficient scoring of constructed responses;
- Improve test-taker motivation through greater engagement.

At the same time, TEI's have some potential limitations (Bachman, 2002; Haigh, 2011; Huff & Sireci, 2001; Parshall & Harmes, 2014; Sireci & Zenisky, 2006):

- They can be more expensive to develop and administer, since they depend upon advanced authoring, delivery, and scoring technologies;
- Their psychometric and performance characteristics are not always well understood, compared with conventional items types;
- They may introduce construct-irrelevant variance;
- Impacts on test takers are not well understood;
- The variety of technology-enhanced item types makes it impossible to draw universal conclusions about the properties and performance of TEIs as a class;
- Different TEI formats function differently and each needs its own independent research;
- In the absence of adequate research, it is difficult to make informed cost-benefit evaluations that can guide the use of TEIs;
- The use of TEIs is generally driven by the functionalities offered by item authoring and test delivery platforms, not by the constructs identified by test developers; that is, technology drives measurement, not vice versa.

Given that TEIs can have a significant impact on costs, test-taker performance, measurement, inference claims, and test marketability, it seems imperative that test makers gain a better understanding of all of these variables in order to arrive at a coherent, integrated strategy for developing and deploying TEIs.

Definitions:

There is no single definition for "technology-enhanced items." The term, along with a bevy of similar names -- technology-enabled items, innovative items, technology-enhanced innovative items, computer-based items, innovative computerized test items, and more -- is generally used to identify computer-delivered items that are not conventional MCs or CRs. For some, "technology-enhanced" can include AI scorability, computer adaptive testing, or the inclusion of audio, video, or animation -- even while the test questions themselves ultimately take the form of conventional MCs or CRs.

For the purposes of this article, "technology-enhanced items" refers to computer-based items that make use of formats and/or response actions not associated with conventional MCs and CRs.[1]

It is worth noting that specialized formats and response actions do not automatically *enhance* assessment. In fact, it is likely that many TEIs do not appreciably add to the quality, breadth, or validity of measurement, compared with conventional item

[1] Broad as it is, this definition is a model of specificity compared with other attempts. One set of researchers, for example, defines the TEI as a "test item that uses technologies that use features and functions of a computer to deliver assessments that do things not easily done in traditional paper-and-pencil format" (Parshall, Harmes, Davey, & Pashley, 2010). Smarter Balanced, as another example, defines TEIs as "computer-delivered items that include specialized interactions for collecting response data. These include interactions and responses beyond traditional selected-response and constructed-response" (Smarter Balanced Assessment Consortium, 2012).

formats, despite the name. To the degree that TEIs *can* enhance assessment, it is by virtue of the alternative response actions, formatting, types of stimulus, and measurement data they afford.

To make informed decisions about the use of TEIs, it is thus important first to define the goals for incorporating them into assessments, and then to determine, through research, the degree of their success in meeting those goals.

It is not particularly easy to classify TEIs, since references to "item types," "response actions," "interaction types," and "item formats," tend to blend and overlap with one another. Further complicating classification is the fact that new item types and response actions are being developed all the time, so there is no fixed or exhaustive catalogue. One broadly-cited classification scheme (Parshall, Harmes, Davey & Pashley, 2010) identifies seven TEI dimensions that can be organized on a continuum ranging from least to most "innovative," where *more innovative* generally means more dependent upon complex computer functionality.

1. Assessment Structure: this encompasses the range of possible formats; for example, discrete items, item sets, constructed responses, situated tasks, simulated environments.
2. Complexity: the number and variety of elements the test taker must consider.
3. Response action: the physical action required of the test takers (i.e., click, drag-and-drop, type, speak into a microphone, etc.)
4. Media inclusion: the use of interactive graphics, animation, audio, video, etc., in stimulus and/or answer choices.
5. Level of interactivity: the extent to which an item reacts or responds to input from the examinee.
6. Fidelity: the degree to which the item resembles or represents real-world contexts. Authenticity.
7. Scoring model: type and mode of response data collection; for example, recording selected responses; the application of artificial intelligence to CRs; multi-part items in which one response is dependent upon another; etc.

This taxonomy could be useful for evaluating and comparing TEIs so as to better understand their required technologies and the effort and cost involved in authoring, delivering, and scoring them. It is not particularly helpful for understanding TEIs as item types designed to elicit targeted knowledge and skills, however.

Another way of classifying TEIs is by the degree to which they constrain responses. One useful categorization schema moves from fully selected responses to fully constructed responses, and, within each category, from less to more complex (Scalise, 2009)[2] . This schema sifts TEIs into seven categories:

1. Multiple Choice
2. Selection/Identification
3. Reordering/Rearranging
4. Substitution/Correction
5. Completion
6. Construction
7. Presentation

Categorizing TEIs this way recognizes that the same response action can be used to achieve different assessment objectives. For example, drag-and-drop functionality can be used to reorder information into a sequence, and also used to complete a sentence or mathematical expression. Reordering and completion are two different item types, presumably measuring different cognitive skills, but both can make use of the same response action.

Content development systems and test delivery platforms classify TEIs according to interactions, but this does not help test developers understand and specify the cognitive skills elicited by the TEIs. For example, these systems make use of organizing terms such as "matching interactions," "order interactions," and "hotspot." Matching and ordering would seem clearly to be cognitive skills, whereas "hotspot" only signifies a type of computer functionality.[3]

[2] This taxonomy adapts for TEIs a model originally designed to categorize various constructed-response and performance formats (Bennett, Ward, Rock, and LaHart, 1990; see also Bennett, 1993).

[3] Some researchers make the point that the combination of response actions and input devices is endemic to TEIs, so a thorough understanding must Some researchers make the point that the combination of response actions and input devices is endemic to TEIs, so a thorough understanding must take into

In addition to clarifying the distinction between item types and response actions, classifying TEIs by degree of constraint begins to suggest how they may elicit different cognitive skills -- i.e., selecting, sequencing, correcting, completing, and so forth. In truth, however, the relationship between degrees of constraint and associated cognitive skills has not been mapped and is not well understood. Test makers do not really know, for example, what, if anything, typing a word into a sentence gap measures beyond what *dragging and dropping* the word measures. They generally suppose, as they do for CRs compared with MCs, that less constraint on test taker responses corresponds to greater cognitive demand; that generating a response is more challenging than selecting one.

Aims and Limitations of TEIs

Generally, the desired direction of technology enhancement for summative tests is toward functionality that moves interaction closer to real-world fidelity: for example, highlighting and annotating texts, as when reading in authentic contexts; math equation editors that allow construction rather than selection; interactive elements simulating real-world science experiments; multimedia sources in History, and so forth. Underlying these enhancements is the assumption that heightened authenticity increases test validity (Huff & Sireci, 2001).

The other major goal for technology enhancement is increased automated scorability. How might the tests broaden assessment constructs (as conventional CR items do), yet keep scoring costs affordable (as conventional MC items do)? Advances toward this goal depend in large part on understanding what a given TEI measures that is different from a conventional selected response item.

Some of the most innovative enhancements are taking place in professional credentialing assessments (Ziv, Sidi & Berkenstadt, 2007). Knowledge and skills assessments in diverse areas, such as medical licensing, accounting, architectural design, patient management, and computer networking incorporate simulations, multimedia stimuli and responses, situated problem solving, and many other computer-based innovations. These new modes of assessment are leading in turn to innovations in automated scoring and new measurement models. Even so, the benefits they add over cost, and their validity compared with conventional tests, are not clearly established.

Perhaps the most fertile area of innovation within education testing is in formative assessments (Bertling, Jackson, Oranje, & Owen, 2015). Researchers are working with simulations, virtual worlds, games, social networks, augmented reality, and other innovations to create new assessments that, unlike conventional summative tests, can provide actionable feedback, be integrated into instruction, engage students over extended time periods, and exhibit greater fidelity to real-world scenarios. Many are incorporating advances in learning and cognitive science. Compared with the summative environment, robust but expensive innovations involving simulations or complex interactions are doubtless more suitable for formative assessments, since they can be re-used, enough perhaps to justify their costs, and do not have the same time, security, and accessibility constraints around them. Nevertheless, cost-benefit ratios and the incremental validity of these innovations are not yet well understood.

Large-scale summative tests in education for the most part have not been able to take advantage of the technological innovations found in credentialing and formative assessments. For one thing, the summative tests are dependent upon the functionality built into available authoring and delivery platforms, and that functionality is comparatively limited (Russell, 2016). In addition, development costs restrict investment in tech-enhancement on summative tests. Further, the time it takes students to learn new formats and interactions or perform extended tasks, the ability to adapt TEIs for accessibility to all populations, the administrative and technological limitations of test delivery, and the fact that it is still very much a paper-and-pencil world in most U.S. schools, all work against advanced item formats on summative tests.

Because not all schools can accommodate computer-based assessments, test makers such as PARCC and Smarter Balanced must produce parallel paper-based versions of their test forms. The need for equivalency between paper and computer versions inevitably calls into question any measurement-based justification for using TEIs. That is, enhancing items with technology by definition runs counter to the objective of equivalency between paper-based test forms and their computer-based counterparts. Such

consideration the use of mouse, keyboard, touch screen, joystick, trackball, speech recognition, or other mechanisms (see Sireci & Zenisky, 2006).

equivalency may be elusive in any case: PARCC, for instance, reported test scores that were lower for students who took the computer version of their test, compared with students who took the paper version, in 2015 (Herold, 2016).

Factors and Implications

Arriving at a coherent TEI strategy for standardized summative tests entails developing criteria upon which to gauge the relative value of TEIs so as to make informed decisions about their use. In general, TEIs appear to reach their greatest potential utility when they do both of two things: 1) enrich measurement compared with traditional MC items, and 2) improve upon test efficiency and scoring costs compared with traditional CR items. This potential for being both richer than MCs and more efficient than CRs would seem to be a prime consideration in devising a strategy for the development and use of TEIs. However, such advantages, where present, need to be clearly identified, articulated, and mapped to construct domains if they are to support assessment inferences in a verifiable way.

Construct Representation

Interest in TEIs is to some degree a response to a longstanding complaint that multiple-choice items are limited in what they can measure (Martinez, 1999; Sireci & Zenisky, 2006; Muckle, 2012). The extensive research comparing MCs to CRs complicates this complaint, but in general it is fair to say that constructed response and performance response items can reach a broader range of cognitive skills than multiple-choice items can (Darling-Hammond & Adamson, 2010).

The hope for TEIs is that they too might elicit higher order cognitive skills, providing more information about reasoning processes, say, or problem-solving strategies. According to Sireci and Zenisky (2006), "The ability to increase representation is perhaps the greatest potential of innovative item formats in computer-based testing" (p. 330). However, to-date there is not much research evidence clarifying what TEIs measure in excess of multiple-choice questions, if anything (Dolan, et al, 2011; Huff & Sireci, 2001). In any case, TEIs are too diverse as a class to warrant blanket statements about their measurement capabilities. Individual technology-enhanced item types might be productively compared to conventional MCs (see, for example, Jodoin, 2003; Bennett & Sebrechts, 1997; Bennett, et al., 1999), but in the absence of focused studies, it is difficult to support general claims about what most TEIs are measuring compared with other types of items.

Authenticity

Another longstanding criticism of MC items is that they do not present examinees with authentic contexts in which to demonstrate their knowledge and skills (Sireci & Zenisky, 2006). More authentic item formats, it is argued, better elicit the skills associated with the construct measured by the test, and thus lead to more valid test score interpretations (Kane, 2006). TEIs can in some cases present problems and questions, or facilitate response actions, that more closely match real-world conditions.

Engagement

Authenticity is an important dimension of student engagement, which studies show is strongly related to student performance on assessment tasks (SCOPE SCALE, 2015). TEIs can enhance assessment validity to the extent that they better engage students in the tasks that elicit their knowledge and skills -- by, for example, presenting real-world problems with greater fidelity, eliciting higher order thinking skills, and increasing agency by providing test takers with opportunities to devise and exercise their own problem-solving strategies.

Non-Relevant Constructs

Increasing authenticity also is a means for reducing non-relevant constructs (Huff & Sireci, 2001). A computer-based math item making use of equation editor technology, for example, might allow a test taker to solve a multiplication problem by writing it out on-screen, as he or she would on a piece of paper. There would be no further demand to locate and fill-in a bubble on the appropriate line of a separate answer sheet. Presumably, then, the TEI would increase test validity by eliminating some of the artificiality of the exam experience.

Of course, for every opportunity to reduce construct irrelevance, TEIs present opportunities to introduce it as well. Examinees must know how to use computer-based tools in order to successfully negotiate a TEI. This can be a challenge for test takers of all types, but especially where the "digital divide" is widest,

or in the case of young children, for many of whom even typing is an unfamiliar skill. In addition to such known sources of construct-irrelevant variance, increasing authenticity can potentially introduce unknown sources as well.

Guessing

TEIs can also improve validity by reducing the possibility that a test taker answers correctly by guessing at random. Multiple-choice questions on paper-based tests are typically restricted to only four or five choices, and machine scoring of printed MC bubble sheets does not allow for flexibility in the number of choices or the number of correct answers. Thus, test takers responding to conventional MC items generally have a 20-25% chance of guessing the correct answer. TEIs can reduce the effects of guessing, compared with conventional MC items, by allowing for more choices and multiple correct answers (Parshall & Harmes, 2014; Tarrant, Knierim, Hayes & Ware, 2006).

Differences in Content Areas

It becomes clear from actual assessment practice that particular technology-enhanced item types are better suited for some content areas than for others. Reading tests, for example, may make better use of Selection/Identification item types compared with Construction TEIs, whereas the reverse may be true for Math and Science tests. A thorough understanding of technology-enhanced item types and their relationships to the skills and abilities particular to different content areas would seem to be a valuable component of any coherent TEI development strategy. Test makers might build such understanding into TEI authoring guidelines to achieve a greater degree of efficiency and standardization in development.[4]

Psychometric Considerations

There are few psychometric research studies on the properties of TEIs (Bennett & Sebrecht, 1997; Bennett et al, 1999; Jodoin, 2003; Gutierrez, 2009; Wan and Henly, 2012). The studies that do exist tend necessarily to focus on particular technology-enhanced item types, not TEIs as a class. Among the questions that might be aimed at TEIs broadly, however, are:

- How do TEIs compare with other item types in terms of
    - Difficulty
    - Discrimination
    - Guessing
    - Reliability
    - Information
- Do test takers take more time to answer TEIs?
- Do TEIs assess something different from conventional items?
- Do TEIs differ in some other psychometric properties, such as item drift or model-data fit?

One study addressing such questions was conducted by the National Council of State Boards of Nursing (Woo, Kim & Qian, 2014). It compared conventional multiple-choice to fill-in-the-blank calculation questions, multiple-response items, and ordered response questions, finding, among other things, that:

- Fill-in-the-blank calculation questions and multiple response items were significantly more discriminating than MC items.
- TEIs tended to provide more information, but also took more examinee time.
- TEIs with high authenticity, such as those incorporating audio and exhibits, tended to measure higher order thinking skills, but other TEI types did not necessarily do so.

Such findings, of course, are confined to the particular test under investigation, but they suggest the kind of analysis that might be conducted by other test makers so as to gain an understanding of the differences among technology-enhanced item types, and the differences between TEIs and conventional items. In addition to the questions above, such analyses might also investigate the impact of TEIs on domain sampling, scoring rules, variance and variability, and other dimensions of measurement.

Scoring Considerations

One of the most attractive attributes of TEIs is their potential for expanding construct representation without the need for expensive hand scoring. Any

[4] See Muckle (2012) for a description of how a major credentialing organization approached the development of technology-enhanced item writing guidelines.

coherent TEI strategy should include a clear understanding of how to make use of this value (Bennett & Bejar, 1998).

Some of the new measurement opportunities afforded by TEIs can create additional scoring considerations as well. Multiple-response questions, for example, can force decisions about partial credit. The utility of multiple-response items can depend on test delivery and scoring platform capabilities. Some platforms do not readily accommodate automated scoring of "composite" items, which pair MC questions with CR questions, or contain two CR questions in a single item, especially when credit for one part depends upon a correct answer in the other.

Greater measurement precision, along with more complex scoring rules, may become possible with TEIs, but greater complexity can increase costs and conflict with standard measurement assumptions (Parshall, Harmes, Davey & Pashley, 2010).

New Constructs

TEIs potentially provide opportunities not only to broaden existing constructs, but to measure altogether new constructs as well. Speaking, listening, research, collaborative problem solving, and other "hard to assess" skills may come within easier reach thanks to computer-based tests that make use of innovative item formats, multimedia, and new input devices or response actions.

Of course, an evaluation of TEI capabilities depends first on clear construct definitions and measurements that are free of construct-irrelevant variance.

Subgroups

An effective TEI strategy must take into account the impact of new item types on test-taking subgroups. Do items that call heavily upon computer skills advantage or disadvantage any test takers based on differences in socio-economic status or cultural markers, such as gender or race? This is an area that needs extensive study. The few studies available focus mostly on the effects of moving from paper-based to computer-based tests, not on TEIs specifically (Parshall & Kromrey, 1993). In general, researchers have found that the online mode makes little or no difference among subgroups. A study of the GRE from 2002, in fact, found that minority populations did somewhat better on computers: "Although all differences were quite small, some consistent patterns were found for some racial-ethnic and gender groups. African-American examinees and, to a lesser degree, Hispanic examinees appear to benefit from the CBT [computer-based test] format. On some tests, female examinees' performance was relatively lower in the CBT version" (Gallagher, Bridgeman, and Cahalan, 2002).

Accessibility

A coherent TEI strategy must take into account the goal of making tests accessible to the entire test-taking population, including people with disabilities and English learners. Computer-based items present both opportunities and challenges for accessibility. On one hand, special tools, such as magnifiers and glosses, can be built into standard items. On the other, TEI formats using color, interactivity, response actions requiring fine motor skills, and other features can be difficult or impossible for some test takers.

Research suggests that, overall, the digital testing environment can improve access to testing for students with disabilities (Thompson, Thurlow, Quenemoen & Lehr, 2002; Thurlow, Lazarus, Albus & Hodgson, 2010). A TEI development strategy must consider how best to incorporate universal design principles and leverage computer functionality to achieve accessibility objectives.

Test Security

In general, digital testing environments tend to reduce test security issues. TEIs themselves, however, are typically fewer in number and more memorable than the conventional items on a standard test. They thus present some higher degree of potential for being memorized and shared (Harmes, Kaliski & Barry, 2007).

Issues and Considerations

A coherent TEI strategy must take shape in the context of larger organizational and market considerations. Some of the key issues and questions for test makers are raised below:

- What role can and should TEIs play in helping the test maker achieve its mission and goals?
    - Do they assist the organization in meeting its stated objectives? If so, how important and effective are they in this?

- To what degree do TEIs present opportunities to improve the real and/or perceived quality and value of the organization's assessments?
- What market considerations must be understood?
    - What is the level of market desire for TEIs? What drives that desire, and to what extent should the test maker respond to it?
    - What role do TEIs play in the competitiveness and marketability of the organization's products? What role will they play in the future?
- Can or should the test maker adopt a policy that defines under what conditions, and on what basis of evidence, it will develop and incorporate TEIs?
- What should be the test maker's commitment to TEI research?
    - Clearly, to make best use of TEIs there is much to discover about their impact on constructs, validity, and timing, to name only a few critical areas. Moreover, there is much to understand about the functions of different technology-enhanced item types in different content areas, different grades, and different assessment programs.
    - How important is such research to the test maker's test development and/or assessment design decisions?
    - If it is important, to what extent should the test maker conduct the research itself, versus contracting it out, or relying on the findings of others?
- What should be the test maker's level of commitment to TEI development?
    - Does the organization want to "catch up," by adopting solutions that prove successful for others, including competitors? Are there opportunities for the organization to "jump ahead," by experimenting and innovating on its own, or with partners?
    - Does the organization want to push for new authoring and delivery platform capabilities or work within the capabilities of available platforms as they advance independently?
- What are the cost impacts of developing and deploying TEIs? How do the cost-benefit tradeoffs compare with those of conventional items?
    - On what basis should decisions be made to incorporate TEIs, given that their costs are likely greater than those of conventional items, and that, at least in the short-term, the test maker may not have empirical evidence of incremental value in measurement and validity?
- What is the role of TEIs in test design?
    - What strategy and rationale should the test maker employ for determining when to use TEIs, how many TEIs to use, of what sort, at what grade level, etc.?
    - What is the organization's approach to the differences (e.g., assessed constructs, timing, difficulty) created by TEIs when tests appear in both online and paper modes?
    - What role do TEIs play in determining whether to assess new constructs?

Conclusion

TEIs can have a significant impact on costs, test-taker performance, measurement, inference claims, and test marketability, Where possible, test makers employing TEIs should commission or conduct empirical analyses of assessment data, aimed specifically at understanding the performance of TEIs compared with one another and with conventional items. Test makers should strive to gain a thorough understanding of TEIs in order to arrive at a coherent, integrated strategy for developing and deploying them.

References

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in

educational performance assessments. *Educational Measurement: Issues and Practice*, 21(3), 5-18.

Bennett, R.E. (1993). On the meaning of constructed response. In R.E. Bennett & 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., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17, 9-17.

Bennett. R.E., & Sebrechts, M.M. (1997). A computer-based task for measuring the representational component of quantitative proficiency. *Journal of Educational Measurement*, 34, 64-77.

Bennett, R.E., Morley, M., Quardt, D., Rock, D.A., Simgley, M.K., Katz, I.R., & Nhouyvanisvong, A. (1999). Psychometric and cognitive functioning of an under-determined computer-based response type for quantitative reasoning. *Journal of Educational Measurement*, 36, 233-252.

Bertling, M., Jackson, G. T., Oranje, A., & Owen, V. E. (2015, June). Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning. In International Conference on Artificial Intelligence in Education (pp. 545-549). Springer International Publishing.

Boyle, A., & Hutchinson, D. (2009). Sophisticated tasks in e-assessment: what are they and what are their benefits? *Assessment & Evaluation in Higher Education*, 34 (3), 305-319.

Darling-Hammond, L. & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

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.

Florida Department of Education (2010). Race to the Top Assessment Program Application for New Grants. Retrieved September 5, 2013 from http://www2.ed.gov/programs/racetothetop-assessment/rtta2010parcc.pdf

Gutierrez, S.L. (2009). Examining the psychometric properties of a multimedia innovative item format: comparison of innovative and non-innovative versions of a situational judgment test. Unpublished dissertation.

Haigh, M. (2011). An investigation into the impact of item format on computer-based assessments. Cambridge, England: Cambridge Assessment.

Harmes, J. C., Kaliski, P. K., & Barry, C. L. (2007, November). Are they really more memorable? Implications of innovative items for test security. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

Herold, Benjamin, (2016). PARCC scores lower for students who took exams on computer. Education Week. February 3.

Huff, K. L., & Sireci, S. G. (2001). Validity Issues in Computer‐Based Testing. *Educational Measurement: Issues and Practice*, 20(3), 16-25.

Jodoin, M.G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement,* 40, 1-15.

Kane, M. (2006). Content-related validity evidence in test development. In S.M. Downing & T.M. Haladyna (Eds.) Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.

Martinez, M.E. (1999), Cognition and the question of test item format. *Educational Psychologist*, 34 (4), 207-218.

Muckle, Timothy (2012). Beyond multiple choices: strategies for planning and implementing an innovative item initiative. Washington D.C.: Institute for Credentialing Excellence.

Parshall, C. G., & Harmes, J. C. (2014). Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology*, 10(1), 1-20.

Parshall, C. G., & Harmes, J. C. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. CLEAR Exam Review, 19(2).

Parshall, C.G., Harmes, J.C., Davey, T., & Pashley, P. (2010). Innovative items for computerized testing. In W.J. van der Linden & C.A.W. Glas (Eds.) Computerized adaptive testing: theory and practice (2nd. ed.). Norwell, MA: Kluwer Academic Publishers.

Parshall, C. G., & Kromrey, J. D. (1993). Computer Testing versus Paper-and-Pencil Testing: An Analysis of Examinee Characteristics Associated with Mode Effect.

Parshall, C. G., Spray, J., Kalohn, J., & Davey, T. (2002). Issues in Innovative Item Types. In Practical

Considerations in Computer-Based Testing (pp. 70–91). New York: Springer.

Russell, M. (2016). A Framework for Examining the Utility of Technology-Enhanced Items. *Journal of Applied Testing Technology,* 17(1), 20-32.

Scalise, K. (2009). Computer-based assessment: intermediate constraint questions and tasks for technology platforms. University of Oregon.

Scalise, K. & Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *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 & T.M. Haladyna (Eds.) Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.

Smarter Balanced Assessment Consortium (2012). Technology-enhanced items guidelines. Developed by Measured Progress/ETS Collaborative.

SCOPE SCALE (2015) Engagement Toolkit Definitions Checklist and Review Tool. Retrieved from https://scale.stanford.edu/sites/default/files/SCOPE%20SCALE%20Engagement%20Toolkit%20Definitions%20Checklist%20and%20Review%20Tool.pdf

Tarrant, M., Knierim, A., Hayes, S.K., & Ware (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, 26 (8), 354-363.

Thompson, S.J., Thurlow, M.I., Quenemoen, R.F., & Lehr, C.A. (2002). Access to computer-based testing for students with disabilities (Synthesis report 45). Minneapolis, MN: University of Minnesota. National Center on Educational Outcomes.

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

Washington State. (2010). Race to the Top Assessment Program Application for New Grants. Retrieved September 5, 2013 from http://www2.ed.gov/programs/racetothetop-assessment/rtta2010smarterbalanced.pdf

Woo, A., Kim, D., & Qian, H. (2014). Exploring the psychometric properties of innovative items in CAT. National Council of State Boards of Nursing. Paper presented at 2014 MARCES Conference: Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring from an Interdisciplinary Perspective. University of Maryland.

Ziv, A., Rubin, O., Sidi, A., & Berkenstadt, H. (2007). Credentialing and certifying with simulation. *Anesthesiology clinics*, 25(2), 261-269.

Citation:

Bryant, William (2017). Developing a Strategy for Using Technology-Enhanced Items in Large-Scale Standardized Tests. *Practical Assessment, Research & Evaluation*, 22(1). Available online: http://pareonline.net/getvn.asp?v=22&n=1

Corresponding Author

William Bryant
Founder and CEO
BetterRhetor Resources LLC

email: wbryant [at] better-rhetor.com