Astonishing Impact:

An Introduction to Five Computer-Based Assessment Issues

Written By:

Phoebe C. Winter, Pacific Metrics
Amy K. Burkhardt, Pacific Metrics
Joseph R. Freidhoff, Michigan Virtual University
Rebecca J. Stimson, Michigan Virtual University
Susan C. Leslie, Michigan Virtual University

Michigan Virtual Learning Research
INSTITUTE™
A Division of MVU

Image inside tablet on cover depicts a sample technology-enabled item designed for ONPAR at the Wisconsin Center for Education Research. Their website is http://onpar.us

About

About Michigan Virtual Learning Research Institute

In 2012, the Governor and Michigan Legislature asked the *Michigan Virtual University*® (*MVU*®) to establish a center for online learning research and innovation, and through this center, directed *MVU* to work on a variety of projects. The center, now formally known as the *Michigan Virtual Learning Research Institute*™, is a natural extension of the work of *MVU*. Established in 1998, *MVU's* mission is to serve as a catalyst for change by providing quality Internet-based programs that strengthen teaching and learning for K–12 education. Toward that end, the core strategies of the Institute include

- Research – Expand the K–12 online and blended learning knowledge base through high-quality, high-impact research;
- Policy – Inform local, state, and national public education policy strategies that reinforce and support online and blended learning opportunities for the K–12 community;
- Development – Develop human and web-based applications and infrastructures for sharing information and implementing K–12 online and blended learning best practices; and
- Innovation – Experiment with new technologies and online learning models to foster expanded learning opportunities for K–12 students.

MVU dedicates staff members to *Institute* projects as well as augments its capacity through a Fellows program drawing from state and national experts in K–12 online learning from K–12 schooling, higher education, and private industry. These experts work alongside *MVU* staff to provide research, evaluation, and development expertise and support.

About Pacific Metrics

Pacific Metrics is a privately held corporation founded in July 2000 by an executive team with many years of experience in large-scale assessments, psychometrics, and software development. Their vision and cohesive background was leveraged to create an organization that offers innovative and practical solutions to modern assessment and learning needs that enhance and support the goals of their customers. Collectively bringing many years of hands-on, large-scale assessment experience, along with a creative approach to the application of technology in this environment, the company's visibility in the industry spans numerous states, large districts, the PARCC and Smarter Balanced consortia, and other organizations that administer large-scale online assessments.

Since its founding, the company has achieved recognition for its technical and psychometric work and for being a leading force in the development and deployment of customized, web-based systems. Pacific Metrics provides online platforms to deliver formative and summative assessments as well as interactive, student-based educational learning systems. The company also provides online item management and reporting services and offers automated scoring of constructed-response items—services that are vital to the needs of today's changing educational environment. Pacific Metrics' offerings also include the delivery of specialized psychometric, technical, software, and content development services to organizations that develop, administer, and score large-scale assessments.

Suggested Citation: Winter, P. C., Burkhardt, A. K., Freidhoff, J. R., Stimson, R. J., & Leslie, S. C. (2013). *Astonishing impact: An introduction to five computer-based assessment issues*. Michigan Virtual Learning Research Institute. Retrieved from http://media.mivu.org/institute/pdf/astonishing_impact.pdf

Abstract

Astonishing Impact: An Introduction to Five Computer-Based Assessment Issues is a primer on computer-based assessment research and the effect of rapidly developing technology on high- and low-stakes assessment development. The authors identify and discuss five issues showing potential for significant impact on computer-based assessments that can be delivered via the Internet. They include: New Item Types, Automated Item Generation, Accessibility of Computer-Delivered Tests for Students, Use of Artificial Intelligence in Scoring, and Increased Efficiency with Accountability Testing. Given existing technologies and the pace of technological change, wide-scale implementation of computer-based assessment opportunities will likely be limited primarily by human willingness to embrace and adopt such innovations.

Acknowledgements

The authors would like to thank Dr. Tia Sukin, Jamey Fitzpatrick, and Kristen DeBruler for their review of and helpful comments on earlier drafts of the paper. Thanks also goes out to the editing staff at Pacific Metrics (Valerie Haff and David Thomas) and *MVU* (Dan Keedy) for assisting with final proofing. Finally, thanks to Victoria Odson for working through multiple iterations of the design and layout of the publication.

Astonishing Impact An Introduction to Five Computer-Based Assessment Issues

"We *will* soon be grading essays by computer, and this development *will* have astonishing impact on the educational world."

—*Ellis Page (1966)*

Introduction

In 1966, the father of automated essay scoring, Ellis Page, identified automated essay scoring as an inevitable development that would have profound educational impact. This development has occurred more slowly than Page anticipated perhaps in part due to validity concerns, technological limitations, and human resistance to change. Close to 60 years later, the promise of automated assessments seems to be at a tipping point, and it is not without company. With the maturation of the Internet, the rapid acceleration of technology accompanied by plummeting cost, and the proliferation of personal and professional computing, we are beginning to see glimpses of the astonishing impact foreseen by Page.

Consider, for instance, recent multi-state efforts to create the Common Core State Standards (CCSS) and the U.S. Department of Education's funding of two assessment consortia—the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC)—to measure performance in meeting the CCSS through large-scale, online assessments. This convergence of states toward two assessment systems in English language arts (ELA) and math has created a concentrated focus on how to appropriately develop online assessments that take advantage of technology to better measure student achievement.

The concentrated focus, however, does not mean uniformity of approach as the assessment models used by the two consortia differ. PARCC will use linear test forms—that is, test forms with a fixed group of test items on them, where every student in a grade receives the same form or one of several equivalent forms. SBAC will use computer adaptive tests (CATs), where students receive items according to how well they did on the previous item or set of items they received. In its simplest form, a CAT assigns an easier item to students who answered the previous item incorrectly and a more difficult item to students who answered the previous item correctly. Because each test administered is tailored to the student's performance, a CAT can provide a more precise estimate of student achievement after administering the same number of items as on a linear form or an equally precise estimate after administering fewer items than on a linear form. A challenge for the CAT model is that it requires a large pool of items to draw from in order to function effectively.

While there is divergence in their assessment models, the two consortia share similarities in the types of items they are constructing. Both consortia include technology-enabled items (TEIs), such as drag and drop, graphing, and text selection or text-based computer-scored constructed-response items. Both consortia are investigating how more complex tasks, such as scenario-based items, might be included in the assessment system. Researchers are reviewing extant research and conducting new research to determine how states can get the most from using technology while maintaining the validity of inferences about students and comparability of scores across students and tests.

Even though many of the technological challenges and validity issues have been resolved over time, challenges still remain when it comes to human acceptance. Whether it be concerns about high-stakes testing in general, funding for such tests, or even acceptance of the CCSS themselves, regardless of how the politics play themselves out, the work the consortia are doing in online assessment provides a valuable foundation for moving assessment design into the future.

Consider also examples of low-stakes, formative forms of online assessment that are emerging. ASSISTments, developed at Worcester Polytechnic Institute in collaboration with Carnegie Mellon,[1] is one such example. The ASSISTments system contains the hallmarks of formative assessment as outlined in a recent Council of Chief State School Officers report: "consistently working from students' emerging understandings within the ZPD [zone of proximal development], supporting learning through instructional scaffolding, including feedback, and the active involvement of students in the assessment/learning process" (Heritage, 2010, p. 15). With the ASSISTments system, the student is presented assessment items, and if the student does not get the answer correct, he or she is provided with a tutoring session, which includes providing supporting questions at the student's current level of learning as well as offering hints to help the student answer the question. Research suggests that data captured in the ASSISTment system can predict math proficiency better than pen and paper benchmark tests (Anozie & Junker, 2006) and as well as standardized tests (Feng, Beck, Heffernan, & Koedinger, 2008).

The ALEKS® system is another example and is similarly designed to provide information to teachers and instant feedback to students.[2] Other key features of this system are its explicit foundation in a cognitive theory (knowledge space theory), its extensive use of item formats other than multiple-choice questions, and its use of artificial intelligence in scoring. Using adaptive techniques, ALEKS is designed to pinpoint a student's status and provide results that indicate the degree to which a student has mastered specific concepts. Within each concept area, students are provided with items specific to that area; when students respond to each item, they receive instant feedback and can access explanations of the concepts assessed. ALEKS courses are offered primarily in mathematics, with a few science and business course offerings. As in the ASSISTments systems, teachers can receive both individual and aggregated information about student progress on the ALEKS system.

Both ASSISTments and ALEKS are designed so that they can be incorporated into teacher-directed courses, delivered either online or face-to-face. Carnegie Mellon's Open Learning Initiative[3] allows the delivery of entire courses online, with or without face-to-face instruction and tutoring. These examples illustrate how technology and cognitive science are affecting instructionally-focused assessments and are the subject of active research, undergoing changes and growth as new information is collected. As more such programs are designed and implemented, both formal research and user-based evaluations of their characteristics and effectiveness will foster rapid change and, if evaluations are used appropriately, rapid improvements. As one measurement researcher recently noted, we are at the beginning of merging the longstanding principles and theories behind formative assessment with the possibilities of technology (E. Matthew Schulz, personal communication, July 3, 2013).

Though, as evidenced by Page, it may not be the beginning. This paper will identify areas of

[1] http://www.assistments.org, retrieved July 2, 2013
[2] http://www.aleks.com/, retrieved July 2, 2013
[3] http://oli.cmu.edu/

ongoing research in the assessment community that are likely to pose significant potential for reshaping both high-stakes and low-stakes assessments. In particular, it will narrow the scope to issues and opportunities around computer-based assessments that would be capable of being delivered via the Internet. Five assessment issues were chosen that are having or will have significant impact. They include

- New Item Types
- Automated Item Generation
- Accessibility of Computer-Delivered Tests for Students
- Use of Artificial Intelligence in Scoring
- Increased Efficiency with Accountability Testing

New Item Types

Most high-stakes tests, and low-stakes tests for that matter, rely heavily on multiple-choice questions, yet there is general agreement that many critical aspects of standards-based knowledge and skills are not well-assessed by multiple-choice items (Ruiz-Primo, February 2009; Herman & Linn, 2013). Although carefully crafted multiple-choice items can measure higher-order skills, they more typically address skills such as recall and application. **The integration of technology in assessment, paired with the application of recent research in cognition, allows for measuring student knowledge and skills in ways that have the potential for more closely addressing the construct of interest. Technology-enhanced items that use drag-and-drop, hot spot, highlighting, graphing, and other techniques; essays and constructed-response items that are machine scored; and simulations that require students to address a complex task from planning through solution can provide more direct information about student learning than constrained selected-response items.**

A simple example, supplied by Pacific Metric Corporation, of a drag-and-drop item can be seen below. In this example, the respondent drags equations from the right-most column into the middle column, placing them in the proper order for solving the initial equation.

Show how to solve this equation for X.

3(2x-5) + 9 = 12

Slide selected equations to the Solution Steps column and place them in the correct order under the given equation.

You must show at least 4 steps in the correct order in order to receive full credit.

Leave unneeded equations in the Equations column.

Step	Solutions Steps	Equations
Given	3(2x-5)+9=12	
1		2x-2=4
2		6x=18
3		3(2x-5)=3
4		2x-5=1
5		2x=6
		6x-15+9=12
		x=3
		2x-5+3=4
		6x-6=12
		6x-15=3

[4] http://www.nwea.org/
[5] http://www.renlearn.com/se/
[6] http://www.ctb.com/ctb.com/control/ctbProductViewAction?productFamilyId=444&productId=30675&p=products
[7] http://www.ctb.com/ctb.com/control/ctbProductViewAction?productFamilyId=444&productId=30675&p=products
[8] http://www.pearsonschoolsystems.com/products/schoolnetforpowerschool/

Although research in this area is just beginning, one study of a large, multi-state sample of grade 7 math and Algebra I students found that a test consisting of technology-enabled and constructed-response items was more reliable and provided more information about students than a test consisting of multiple choice items measuring the same skills (Winter, Wood, Lottridge, Hughes, & Walker, 2012).

At present, most items embedded in online learning materials are multiple-choice, short-answer, or of a form that can be deterministically scored, that is, by entering in parameters that specify how to score specific responses (e.g., graphing items) (Scalise & Gifford, 2006). Similarly, many computer-delivered assessments designed for formative or interim use more easily-scored items and are transitioning to the inclusion of deterministically scored technology enhanced (TE) items (e.g., Northwest Evaluation Association's Measures of Academic Progress®;[4] Renaissance Learning's STAR™ assessments;[5] CTB/McGraw-Hill's Acuity® assessments;[6] Acuity® InFormative Assessment™;[7] Pearson's Schoolnet for Powerschool[8]). An exception to the prevalent use of easy-to-score items in online systems is the formative assessment of writing skills, in which students write and submit essays. Several companies offer assessments designed for students to practice, revise, and finally submit their essays for a final score, providing feedback to the student along the way.

Automated Item Generation

Along with technology enabling new test item types, two trends in online assessment are fueling the demand for a high volume of test items and tasks that meet explicit criteria for quality. First, as technology and online access become more widely available in schools, more large-scale assessment programs, most notably SBAC, are developing computer adaptive tests as their summative assessments. The move to computer adaptive testing platforms for the administration of large-scale assessments brings a demand for item banks consisting of several thousand items (American Institutes for Research, 2013). Second, demands for in-classroom computer-based assessments, often as a non-summative part of large-scale systems, are increasing. This is in part due to the potential for quick scoring of items and tasks that can be embedded as part of instruction, making results available for real-time adjustments in individual instruction and instructional plans. At the same time, **the increased use of assessment in the classroom has led to the need for more items targeting specific areas of the learning progressions or other curricular units. Thus, there is a need for scalability in item development where items can be created in an efficient and economic manner. Automatic Item Generation (AIG) serves as an algorithmic approach to creating items. Along with the scalability of this method, AIG offers other benefits, such as reducing manual errors, decreasing item authoring costs, improving item security, increasing reliability, and providing items of consistent quality, founded in a deep understanding of the cognitive context.**

While several variations for the implementation of AIG and the construction of item models exist (Luecht, 2013), they largely fall within two categories: weak or strong theory (Drasgow, Luecht, & Bennett, 2006). Methods relying upon weak theory make use of item construction guidelines and subject matter experts to determine which item features of a parent (i.e., model) item can be manipulated without altering such psychometric properties as difficulty. Conversely, strong theory employs a cognitive model that defines the interaction between item features and examinee response processes. Strong theory makes it possible to predict the psychometric properties (e.g., difficulty) of the generated items. Examples where this approach has been successful include the constructs of spatial reasoning (Bejar, 1990), abstract reasoning (Embretson, 2002), and a host of quantitative skills (Graf & Fife, 2013).

There have been a number of more recent successes employing AIG methods to assess academic content. In one study, 1,248 multiple-choice items for medical examinations were created from a single item model (Gierl, Lai, & Turner, 2012). Another approach, known as "min-max," experienced a low item rejection rate even after field-testing (Arendasy & Sommer, 2012). College Board's Advanced Placement assessment for biology (and other areas of science) approached its assessment development design by articulating claims and the associated evidence required to support each claim and then constructed task models and item templates to implement AIG (Huff, Alves, Pellegrino, & Kaliski, 2013). Biology items developed from four templates resulted in a total of 1,787 items.

Beyond the mere number of items that can be generated, researchers are also focusing efforts on varying the linguistic context of items. For example, mathematical word problems require setting up a context within which to apply mathematical skills. However, automating the construction of contexts can be challenging. One example includes employing a Natural Language Generation approach for developing rich contexts (Deane & Sheehan, 2003). Within the construct of ELA, very little progress has been made, with the exception of reading comprehension, specifically, determining the meaning of words and sentences (Gierl & Leighton, 2010). Even then, the number of items generated from templates is limited (e.g., four to five items per template).

Investigating the technical quality of items developed via AIG is a natural next step to developing the field of AIG. Gierl (2012) identifies two ways to assess the technical quality for items generated algorithmically. The first way is to investigate validity and reliability through item analysis. Across AIG methods, researchers have supported their conclusion regarding the technical quality of these items by means of psychometric analyses (Arendasy & Sommer, 2012; Embretson, 1998, 1999; Holling, Bertling, & Zeuch, 2009). A complementary form of validity inspection involves analyzing the quality of the cognitive item model using the expert judgment of educators, cognitive scientists, and other content specialists before, during, and after item generation.

AIG has the potential to increase drastically the availability of high-quality items with known psychometric and/or construct-related properties for both summative and formative use in instructional settings as well as for expanded practice opportunities. Items and tasks available online can be tagged at multiple levels, from content standards to suggested instructional sequence within a curriculum, allowing teachers and other educators to easily select items and tasks pinpointed for the specific purpose and desired content of assessments used for ongoing formative evaluation. Some Learning Management Systems (LMS) have a mastery/proficiency level testing that can result in the student branching to build capacity (strong theory) or being locked out of proceeding with content until an instructor or mentor has been involved in review. Such a pool of items and tasks could also be used by instructors and other educators to create parallel test forms that could be used across multiple administrations, for example, for pre- and post-tests.

Concerns about compromised test items may diminish significantly if AIG is used to create parallel large-scale assessments for a single administration, as the likelihood of students actually receiving one or more compromised items is lower than with traditional assessments using small banks of items. Multiple forms of tests used in a summative manner, such as unit tests, increase assessment options and can also alleviate local security concerns about specific test content being revealed from year to year or between the original test and a make-up opportunity or remediation.

As techniques and models for AIG are being developed, it has become clear that the appropriate person to act as the original item generator must have knowledge of the content area, instruction, and programming. The individual must create question patterns that coordinate with the goals of the course while meeting standards and outcomes, determine the difficulty and complexity of the items, and set parameters such as constraints and values. It is likely cost prohibitive for local schools or even districts to develop original questions. Fortunately, as automated item generation procedures become more sophisticated and the field matures, there are a number of possibilities for more flexible use of procedures and templates by local educators. For example, interfaces may be developed that allow teachers to specify which item and task features they want to vary in a set of items and their related scoring templates.

With AIG still considered a revolutionary science, future research will likely focus on expanding applicability to alternative item formats (i.e., beyond multiple-choice) and text-rich constructs (e.g., history and ELA) and the use of hybrid models that combine the work of item developers and an item/task model to develop more complex items and tasks. Additionally, we will see continued movement toward the integration of test development and learning sciences as researchers attempt to assess student achievement using cognitively based models such as learning progressions.

Accessibility of Computer-Delivered Tests for Students

Tests are accessible to the degree that students with various physical, cognitive, sensory, linguistic, or other barriers are provided the opportunity to demonstrate the knowledge, skills, and abilities (KSAs) intended to be measured—the targeted KSAs (Winter, Kopriva, Chen, & Emick, 2006). Aspects of the testing situation that interfere with measuring the targeted KSAs are minimized through providing appropriate settings, tools, and accommodations for students, preferably by building these options into the testing situation rather than adding them after the fact. As new types of test items become possible and the numbers of items we can generate expand with AIG, technology must also be leveraged to make these items more accessible to students. Using computers for testing provides an opportunity to make tests more appropriate for, or accessible to, a greater number of students—particularly students with disabilities and students with limited English proficiency—through the provision of embedded tools (e.g., text to speech, roll-over translations and definitions, and magnification) and accommodations (e.g., sign language avatars and refreshable Braille) (Almond et al., 2010).

The National Center on Educational Outcomes for Students with Disabilities (NCEO) has published guidelines for accessibility (Thurlow, Lazarus, Albus, & Hodgson, 2010), but more research and development is needed to ensure that the tests are both accessible and are producing scores that allow for valid inferences to be made about student performance. Researchers and users of scores from tests with online accommodations are not yet in accord about the comparability of scores from some particular accommodations (see for example, Randall, Sireci, Li, & Kaira, 2012; Winter, April 2009; Winter, 2010). Of particular interest is the provision of alternative representations of items to students as an accommodation, so that the same construct is assessed in a way that a student's accessibility needs do not interfere with obtaining valid information (Russell, 2010; Russell, Mattson, Higgins, Hoffmann, Bebell, & Alcaya, 2011). For example, an item that cannot be Brailled may have a corollary item or set of items that can be rendered in Braille.

[9] http://www.imsglobal.org/apip/, retrieved July 7, 2103

Current work is being conducted to define the Accessible Portable Item Profile (APIP) standard that would allow for standardizing the file formats used to transfer test items from one system to another, for example, from an item development system to an online test delivery system. The APIP standard is intended to serve two purposes. The first purpose is to enhance the transportability of items from one system to the next so that items from one test banking system can be used in another APIP-compliant test banking system. The second purpose is to provide a test delivery interface with all the information and resources required to make a test and an item accessible for students with a variety of disabilities and special needs.[9]

Ideally, APIP-compliant items can be administered according to information provided by a student access profile (Russell, Mattson, Higgins, Hoffmann, Bebell, & Alcaya, 2011), an electronic record that contains information about the tools, test conditions, and accommodations that a student is eligible to receive given his or her accessibility needs. The test delivery system would retrieve this information and deliver the tests and items within a test with appropriate features for the particular student.

More research is needed on implementing the student access profile for assignment of accommodations, including the degree to which it is feasible for schools to upload such student information into a single database. A model for doing so can be found in research conducted by Kopriva and her colleagues on creating an access profile of English learners (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007). The authors created a system combining multiple sources of information—from parents, teachers, the student, and school records—to develop a profile that led to specific recommendations about appropriate accommodations. This idea of personalizing instruction to meet the specific needs of the student has been argued essential not only for students with identified access needs but for all students (Public Sector Consultants & Citizens Research Council, 2013).

Much of the effort and energy around accessibility research and development has focused on high-stakes assessments, fueled in large part by state-mandated tests used for student, school, and district accountability. There is, however, an emerging realization that little attention has been paid to accessibility issues involving low-stakes testing or assessments designed to guide classroom instruction in real time or to summarize the effectiveness of instruction, despite the fact that fewer barriers may exist for implementing accessible assessments for these purposes. For instance, assessments used during the learning process have greater flexibility in the types of tools that can be offered to students than those used for high-stakes purposes, since rigorous research exploring the effects of such tools on measurement of the construct is not needed for low-stakes assessments.

As advances in assistive technologies, such as the touchscreen capability of mobile phones and tablet computers and student-controlled text-to-speech technology, rapidly expand the possibility of what can be done in schools at an affordable price, it is likely that more tools and accommodations applicable to classroom instruction will be developed and available to students and teachers. These accommodations will minimize the barriers to students trying to demonstrate their knowledge and skills in the targeted area.

The accessibility of assessment for all students opens doors enabling valid assessments at every stage of the learning process. This, according to West (2011), will improve learning and drive educational change. Personalization or scaffolding learning and measuring it for each and every student, is attainable as a result of this accessibility in online testing.

Use of Artificial Intelligence in Scoring

Just as computer-based testing makes possible the development of new types of technology-enhanced test items and tasks and makes them more accessible to a range of students, the availability of increasingly sophisticated automated scoring algorithms is showing great potential. Acceptance of automated scoring for both low- and high-stakes testing is also influenced by quick turnaround time, cost effectiveness, and reliable scores. The efficiency gained through the automated scoring process can allow for an increase in the use of both TEIs and traditional text-based constructed-response items, from simple drag-and-drop items to complex scenario-based tasks and essays. Automated scoring is being used to score a variety of items, including essay length writing prompts, short-answer constructed-response items, and technology-enhanced math items (Lottridge, Shultz, & Mitzel, 2013).

Four states currently use automated scoring software to assist in scoring student responses. The automated scoring engine often serves the role of a "second reader" that checks the reliability of the human readers. Online courses such as EdX (funded by Harvard and MIT), Coursera, and Udacity (both started by Stanford) are also committed to integrating machine-scored assessments into their free course offerings (Markoff, 2013). Criterion® from Educational Testing Service offers a low cost option for elementary and secondary schools. Using ETS's e-rater as its scoring engine, Criterion has been favorably reviewed (Lim & Kahug, 2012).

There is a growing pool of evidence supporting the validity of automated scoring (Shermis & Burstein, 2013), most focusing on essay scoring. While even the first automated scoring system showed that the patterns of agreement between machine and human scores were indistinguishable from patterns between scores from two human raters (Page, 1966), researchers acknowledge that the current state of automated essay scoring is limited, for example, not able to appropriately score a subset of writing constructs (Attali, 2013). Researchers have also pointed out reliability and validity concerns in human ratings of essays and recommend combining automated and human scoring in ways that improve the validity of scores. Automated scoring can also serve as a monitoring tool to mitigate the risks of rater bias in essay scoring because computer scoring has been found to be more consistent than human raters (Lottridge, Shultz, & Mitzel, 2013).

In 2012, the Hewlett Foundation sponsored a competition: the Automated Student Assessment Prize (ASAP). In the first phase of this competition, nine vendors with state-of-the art essay scoring software scored eight essays that were 150 words or longer. The competition results demonstrated that most automated essay scoring systems performed similarly to each other and to the human raters (Shermis & Hamner, 2012). The second phase of this ASAP competition examined the performance of machine scoring for constructed responses that were around 50 words in length. For this second phase, automated scoring underperformed relative to the human raters (Shermis, 2013). Initial research in scoring more constrained item responses such as graphing items or items requiring an expression or equation as a response have shown promise. In general, the machine scoring engine performs as accurately as human scorers (Winter, Wood, Lottridge, Hughes, & Walker, 2012).

The current direction of automated scoring research is to continue the investigation of the viability of using automated scoring in the new generation of assessments of the PARCC and Smarter Balanced consortia (Shermis & Hamner, 2012). For now, in high-stakes assessments, most essays and complex constructed-response items are human scored or scored using a combination of human and machine scoring. Active research is addressing ways of developing better automated scoring algorithms. Eventually, machine scoring may provide the primary item score with humans checking a sample of scores for accuracy.

Currently, the use of machine scoring as the sole source of student scores and feedback is limited in high-stakes assessment. However, the low- and medium nature of formative assessments (and other tests that are instructional based) make machine scoring as the sole source of a score a much more attractive option (see Lottridge, Winter, and Mugan, 2013, for a discussion of scoring models for different item types and test stakes). Along with machine scoring providing immediate feedback, a prerequisite for useful formative assessment, it also allows for the use of more complex items, including constructed-response and essay items, while a student is learning.

Concerns about "gaming the system" or learning the rules of the scoring algorithm are also alleviated when automated scoring is used for formative feedback. These low-stakes assessment opportunities can give students and teachers a better idea of what the student has learned, an opportunity to build on that learning, and details regarding what the student needs to revisit. With the development of more sophisticated machine scoring techniques tied to learning research, such scoring could provide detailed information about student understanding and point to different learning paths based on student responses.

Increased Efficiency with Accountability Testing

Even though the advantages of computer-based testing such as automated scoring may provide more opportunities for frequent testing of students, it may also be leveraged to help reduce the amount of time students spend on high-stakes assessments. Current education legislation focuses heavily on test-based accountability and requires performance information at various levels of aggregation, including the student level for the measurement of individual achievement, the school level for comparisons of consecutive grade assessments, and school, district, and state levels to provide aggregative information about school effectiveness (Weiss, 2010). These competing demands on test results pose a challenge in designing a comprehensive assessment program that can validly and reliably report all information in an efficient manner. A carefully constructed computer adaptive test might provide one solution to reducing the amount of testing needed for accountability (Gage, 2013). Another solution that would also be appropriate for linear (non-adaptive) tests is to leverage computer-based technologies with a previously-proposed method of reporting aggregate and individual information: the duplex design.

The duplex design couples two different assessment methods to capture both individual and aggregate data (Bock & Mislevy, 1988). The first form of assessment is to test for individual student achievement. Because reliable student scores cannot be assumed to provide generalizable scores at the class, district, or state level, another method must also be employed to obtain aggregative information. Matrix sampling addresses benchmarks of learning in a manner where a group of items covering the full range of desired content is distributed across examinees so that no individual responds to every item. This practical approach has been used for accountability in the past (Bock & Mislevy, 1988), assuring generalizability of the group mean without spending an inordinate amount of time testing, and is resistant to the effects of teaching to the test.

Now that computer-based assessments are becoming more common, researchers propose a revamped version of the duplex design in order to provide a single assessment used to satisfy all competing needs from legislation (Bejar & Graf, 2010). The assessment is designed to be a three-stage test, where the first and second stages are used to precisely classify students into achievement levels for a given grade. The third stage of the test can employ matrix sampling techniques to gather aggregate performance information. In order to shorten this assessment to a feasible length, automated scoring and adaptive testing must be used. Automated item generation would be used to economically generate comparable results at the school and district levels and support the validity of the inferences made from the test. Therefore, implementing such a design that couples matrix sampling and individual achievement relies on the validity, acceptance, and progress of the technologies it employs.

An obvious implication of more efficient accountability assessments is more time for instruction, incorporating assessments to evaluate learning and provide feedback as students learn. In addition, the principles underlying large-scale assessments designed to provide both useful individual scores and valid information about group progress can be used for classroom assessments that are more summative in nature; for example, those used to determine whether students have met targeted competency standards after a period of instruction and assessments that contribute to a student's grade, such as end-of-unit tests.

These principles are less applicable to the use of formative assessments; indeed, most measurement professionals warn against using the same assessments for both formative and summative purposes. This is due to technical concerns and because of the effects such uses can have on the utility of information stemming from results. An important distinction between formative and summative assessments is that the former is used to direct and adjust instruction as it is proceeding (assessment *for* instruction), while summative assessments such as unit tests are used primarily to evaluate the effectiveness of instruction delivered over a period of time (assessment *of* instruction) (Wiggins, 2011; Heritage, 2010).

An example from the learning progressions literature illustrates this difference. Student understanding of the concept of buoyancy tends to develop along a pathway ranging from understanding that the mass of an object affects buoyancy to understanding that buoyancy is dependent on the interaction between the density of an object and the density of the medium in which the object is placed (Draney, Galpern, & Wilson, 2005). A formative assessment designed to guide immediate instruction might include items and tasks that both locate where along the continuum a student's understanding lies and more specific items and tasks that elicit information about the characteristics of that level of understanding. A summative assessment designed to determine whether the students as a whole learned the intended outcomes of a unit on buoyancy might include more items focusing on the broad range of concepts needed to reach those outcomes as well as questions targeted specifically on the intended outcomes of instruction. It can be seen that the design of these two assessments would be different.

Depending on the purpose of the summative assessment, the duplex design or a purposefully-designed CAT could be used. If the summative assessment was both to show how well the class as a whole learned the topic after the unit was completed and to evaluate student learning to determine what topics needed to be reinforced for a student or the class as whole, such a model might be useful. The utility of these models is enhanced when assessments used for summative purposes are administered or designed as part of a systematic program, such as one instituted or supported by a school district.

Conclusion

This paper offered glimpses into five assessment issues that have the potential to significantly impact high-stakes and low-stakes assessment, changing how assessment items are used in classrooms with sufficient online access and boosting the speed at which online instructional models can be developed and implemented. Commercial interim assessments and formative item banks, once consisting solely of multiple-choice and fill-in-the-blank items, can expand to include technology-enabled items and machined-scored constructed-response items, providing better information to teachers and students about progress and next steps in learning. The application of better item development and scoring technology has the potential to rapidly expand the use of online learning systems with embedded self-assessments, the utility of emerging instructional trends such as flipped instruction, and the possibilities for more individualized learning. As Page foretold, it can appear that "the times they are a-changin." But there are other axiomatic conclusions that could occur—"the more things change, the more they stay the same."

What seems likely given Moore's law and the rapid rate of technological innovation is that the limiting factor in changing education will no longer be technology, but rather human desire for and willingness to embrace these changes. The desire for this kind of buy-in is evidenced by the U.S. Department of Education's invitation to the public and to assessment experts in general to comment on how to best evaluate the assessments required by the Elementary and Secondary Education Act.[10] Ironically, perhaps one of the best ways to improve the quality of education in the United States is to improve the quality of education of the American public and policymakers about new assessment issues.

[10] http://www.ed.gov/blog/2013/08/help-ed-improve-how-we-evaluate-state-assessment-systems/ posted on August 7, 2013.

References

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *Journal of Technology, Learning, and Assessment, 10*(5). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1605/1453

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*(4), 341–359. doi: 10.1111/j.1745-3984.2007.00043.x

American Institutes for Research. (2013). *Smarter Balanced Assessment Consortium: Online pilot test administration manual*. Retrieved from http://sbac.portal.airast.org/Pilot_Test/resources/SmarterBalancedPilotTestAdministrationManual_updated031113.pdf

Anozie, N. O., & Junker, B. W. (2006, July). *Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system*. Paper presented at the Association for the Advancement of Artificial Intelligence's Twenty-first National Conference on Artificial Intelligence, Boston, MA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.6701&rep=rep1&type=pdf

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences, 22*(1), 112–117. doi:10.1016/j.lindif.2011.11.005

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*(3), 237–245. doi: 10.1177/014662169001400302

Bejar, I. I., & Graf, E. A. (2010). Updating the duplex design for test-based accountability in the twenty-first century. *Measurement: Interdisciplinary Research & Perspective, 8*(2–3), 110–129. doi:10.1080/15366367.2010.511976

Bock, R. D., & Mislevy, R. (1988). Comprehensive educational assessment for the states: The duplex design. *Educational Evaluation and Policy Analysis, 10*(2), 89–105. doi: 10.3102/01623737010002089

Community for Advancing Discovery Research in Education. (2012, April). *New measurement paradigms*. Retrieved from http://www.academia.edu/2773351/New_measurement_paradigms

Deane, P., & Sheehan, K. (2003). *Automatic item generation via frame semantics: Natural language generation of math word problems*. Educational Testing Service. Retrieved from http://ccl.pku.edu.cn/doubtfire/semantics/automaticitemgenerationviaframesemantics-by-deane.pdf

Draney, K., Galpern, A., & Wilson, M. (2005, November). *Designing and using an embedded assessment system to track student progress*. Paper presented at the National Science Teachers Association conference, Chicago, IL.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: Praeger Publishers.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380–396. doi:10.1037/1082-989X.3.3.380

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407–433. doi:10.1007/BF02294564

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219–250). Mahwah, NJ: Erlbaum.

Feng, M., Beck, J., Heffernan, N., & Koedinger, K. (2008). Can an intelligent tutoring system predict math proficiency as well as a standardized test? In R. S. J. d. Baker, T. Barnes & J. E. Beck (Eds.), *Educational data mining 2008: 1st international conference on educational data mining, Proceedings* (pp. 107–116). Montreal, Canada. Retrieved from http://www.educationaldatamining.org/EDM2008/uploads/proc/full%20proceedings.pdf

Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). Addressing the testing challenge with a web-based e-assessment system that tutors as it assesses. *WWW '06 Proceedings of the 15th international conference on World Wide Web* (pp. 307–316). New York: Association for Computing Machinery. doi:10.1145/1135777.1135825

Gierl, M. J., & Leighton, J. P. (2010, April). Developing cognitive models and construct maps to promote assessment engineering. In R. M. Luecht (Chair), *Application of assessment engineering to multidimensional diagnostic testing in an educational setting*. Paper presented in symposium conducted at the annual meeting of the National Council on Measurement in Education, Denver, CO.

13

References

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757–765. doi:10.1111/j.1365-2923.2012.04289.x

Graf, E. A., & Fife, J. H. (2013). Difficulty modeling and automatic generation of quantitative items: Recent advances and possible next steps. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 157–179). New York: Routledge.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Washington, DC: Council of Chief State School Officers (CCSSO). Retrieved from http://www.edweek.org/media/formative_assessment_next_generation_heritage.pdf

Herman, J. L., & Linn, R. L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia* (CRESST Report 823). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://www.cse.ucla.edu/products/reports/R823.pdf

Holling, H., Bertling, J. P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation, 35*(2–3), 71–76. doi:10.1016/j.stueduc.2009.10.004

Huff, K., Alves, C. B., Pellegrino, J., & Kaliski, P. (2013). Using evidence-centered design task models in automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp.102–117). New York: Routledge.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11–20. doi: 10.1111/j.1745-3992.2007.00097.x

Lottridge, S., Winter, P., & Mugan, L. (2013). *The AS decision matrix: Using program stakes and item type to make informed decisions about automated scoring implementations*. Paper presented at the National Conference on Student Assessment. Retrieved from https://ccsso.confex.com/ccsso/2013/webprogram/Handout/Session3711/ASDecisionMatrix_WhitePaper_Final.pdf

Lottridge, S. M., Schulz, E. M., & Mitzel, H. C. (2013). Using automated scoring to monitor reader performance and detect reader drift in essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 233–250). New York: Routledge.

Luecht, R. M. (2013). Automatic item generation for computerized adaptive testing. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp.196–216). New York: Routledge.

Markoff, J. (2013, April 4). Essay-grading software offers professors a break. *The New York Times*. Retrieved from http://www.nytimes.com/2013/04/05/science/new-test-for-computers-grading-essays-at-college-level.html?pagewanted=all&_r=0

Measured Progress; ETS Collaborative. (2012). *Smarter Balanced Assessment Consortium: Technology-enhanced items guidelines*. Washington, DC: SBAC. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/TechnologyEnhancedItems/TechnologyEnhancedItemGuidelines.pdf

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards.* Washington, DC: Authors. Retrieved from http://www.corestandards.org/the-standards

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

Page, E. B. (1966). The imminence of grading by computers. *The Phi Delta Kappan*, 47(5), 238–243.

Public Sector Consultants & Citizens Research Council. (2013). *Moving Michigan farther, faster: Personalized learning and the transformation of learning in Michigan*. Lansing, MI: Michigan Virtual University. Retrieved from http://www.mivu.org/LinkClick.aspx?fileticket=90cm7nhVPIE%3d

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice, 31*(4), 2–12. doi: 10.1111/j.1745-3992.2012.00252.x

Ruiz-Primo, M. A. (2009, February). *Towards a framework for assessing 21st century science skills*. Paper commissioned for the Workshop on Exploring the Intersection of Science Education and the Development of 21st Century Skills. Washington, DC: The National Academies. Retrieved from http://sites.nationalacademies.org/DBASSE/BOSE/DBASSE_080127

Russell, M. (2011). Computerized tests sensitive to individual needs. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students* (pp. 255–273). New York: Springer.

References

Russell, M., Mattson, D., Higgins, J., Hoffmann, T., Bebell, D., & Alcaya, C. (2011). *A primer to the accessible portable item profile (APIP) standards*. Minnesota Department of Education. Retrieved from http://apipstandard.org/archive/papers/APIP%20Primer%20-%20Final.pdf

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). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653/1495

Shermis, M. (2013, April). *Contrasting state-of-the-art in the machine scoring of short-form constructed responses*. Paper presented at the 2013 National Council for Measurement in Education Conference, San Francisco, CA.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York: Routledge.

Shermis, M. D., & Hamner, B. (2012, April). *Contrasting state-of-the-art automated scoring of essay: Analysis*. Paper presented at the 2012 National Council of Measurement in Education Conference, Vancouver, BC. Retrieved from http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007, June). *An assessment for learning system called ACED: Designing for learning effectiveness and accessibility*. Educational Testing Services. Retrieved from http://www.ets.org/Media/Research/pdf/RR-07-26.pdf

The Gordon Commission. (2013). *To assess, to teach, to learn: A vision for the future of assessment*. Princeton, NJ: Author. Retrieved from http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

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. Retrieved from http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis78/Synthesis78.pdf

Weiss, J. (2010). *Race to the Top assessment competition: Overview for state applicants*. U.S. Department of Education. Retrieved from http://www2.ed.gov/programs/racetothetop-assessment/competition-overview.pdf

West, D. M. (October 6, 2011). *Using technology to personalize learning and assess students in real-time*. Center for Technology Innovation at Brookings. Retrieved from http://www.brookings.edu/~/media/research/files/papers/2011/10/06%20personalize%20learning%20west/1006_personalize_learning_west.pdf

Wiggins, G. (2011). Kappan classic: A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 92*(7), 81–93.

Winter, P. C. (Ed.). (2010). *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/2010/Evaluating_the_Comparability_of_Scores_2010.pdf

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 from http://www.pacificmetrics.com/files/OMAP/omap%20final%20research%20report%20body.pdf

Michigan Virtual Learning Research
INSTITUTE™
A Division of MVU

3101 Technology Blvd., Suite G | Lansing, MI 48910 | www.mivu.org | 888.532.5806