Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2018, Vol. 37, No. 4, pp. 7–15

Educational Assessment: What to Watch in a Rapidly Changing World

Randy E. Bennett, *Educational Testing Service*

This article is a written adaptation of the Presidential address I gave at the NCME annual conference in April 2018. The article describes my thoughts on the future of assessment. I discuss eleven likely characteristics of future tests and, for each characteristic, why I think it is important and what to watch with respect to it. Next, I outline what is unlikely to change. The article concludes with a comment about the probable state of education 10 years on and what that state might mean for members of the assessment community.

Keywords: assessment, educational futures, formative assessment, testing

In this article, I describe what I think will be some likely characteristics of future educational assessments in the various forms those future assessments might take and diverse uses to which they might be put. For each characteristic, I state why that characteristic is important but, perhaps more interestingly, what to watch with respect to it. This first part of the article is, then, about change. In the second, shorter portion, I discuss what I think will *not* change. I close with a brief comment about the state of education as I envision it 10 years from now and what that state might mean for the future of educational assessment.

I begin with a list of eleven characteristics, some of which may be obvious because they are already happening but, in all cases, their presence is still quite limited. The claim is that these characteristics will become more widespread, going well beyond the research projects and high-end operational settings in which some of them might already be occurring.

I expect assessments of the future to:

- Be technology based
- Measure "new" constructs
- Be built from richer underlying models of cognition and learning
- Make greater use of more complex tasks
- Be "personalized"
- Attempt to improve learning
- Be better at accounting for context
- Be "embedded" and distributed across time
- Use automated scoring
- Incorporate new approaches to modeling and analysis
- Provide more effective reporting

Be Technology Based

Electronic delivery is important for at least three reasons. First, it allows for traditional competencies to be measured more effectively and efficiently by, for example, speeding assessment presentation and response return. Second, it permits the measurement of new competencies that couldn't be assessed with traditional methods. Examples include writing on computer, reading in hypertext environments, collaborating with remote partners in virtual spaces, and executing problem-solving processes themselves. Finally, electronic delivery makes possible collecting and analyzing the "big data" coming from online learning activities.

What should we watch? We should watch the leading international, national, and state assessments. Because of their prominence, size, and resources, these programs can and do innovate as one mechanism for demonstrating their relevance and leadership.

Among the international assessments, the Program for International Student Assessment (PISA) was given in 2015 on school equipment to about 400,000 fifteen-year olds in 57 countries and other political jurisdictions (M. von Davier, personal communication, November 21, 2017; Organization for Economic Cooperation and Development [OECD], 2017). Assessments were administered in reading, math, science, problem solving, and financial literacy. What is unique about PISA is that its technology-based assessments were administered in 90 language versions, an enormous challenge in management, content development, task presentation, response collection, scoring, and analysis.

ePIRLS, the online segment of the Progress in International Reading Literacy Study, was given in 2016 to approximately 85,000 fourth-grade students in 16 education systems on school computers (Institute for Education Sciences [IES], n.d.; Mullis & Prendergast, 2017). Unique to ePIRLS was its focus on measuring online informational reading, a construct arguably quite different from the one measured in the reading comprehension portions of more traditional assessments. In ePIRLS, the text selections included hyperlinks and material housed on tabs, making navigation dissimilar to the linear process commonly used to read ordinary text.

With respect to national assessments, the National Assessment of Educational Progress (NAEP) gave its reading and math assessments in 2017 on tablets to 150,000 students in

Randy E. Bennett, Norman O. Frederiksen Chair in Assessment Innovation, Educational Testing Service, Princeton, NJ 08541; rbennett@ets.org.

each of Grades 4 and 8 (National Assessment Governing Board [NAGB], n.d.a). It also administered its writing assessment in the same way to roughly 20,000 students in each of those grades (IES, 2018). Unique to NAEP was that it brought machines to schools in an attempt to better control the variation in performance that would otherwise occur from differences among school computers. NAEP was able to take this approach only because it tests comparatively few students in each of a relatively small number of schools. NAEP's methodical approach to technology implementation has also been unusual, beginning with substantial research (Bennett, Persky, Weiss, & Jenkins, 2007; Horkay, Bennett, Allen, & Kaplan, 2005; Sandene, Bennett, Braswell, & Oranje, 2005), moving next to small operational measures (2009 Interactive Computer Tasks in science; IES, 2010), then to assessments administered only to national samples (writing, Technology and Engineering Literacy; IES, 2012; NAGB, n.d.b), and most recently to its main assessments in large state and national samples.

Beginning in 2018, the Australian National Assessment Program (NAPLAN) administered its Online Literacy and Numeracy assessments to students in years 3, 5, 7, and 9 (NAPLAN, 2018). NAPLAN administered these assessments on school machines as well as on "approved personal devices," apparently the only large assessment program to have done so.

Finally, among U.S. states, the California Assessment of Student Performance and Progress (CAASPP) is highly significant. A press release dated September 27, 2017, describes the administration of the CAASPP online tests. It states, "California testing went smoothly for 3.2 million total students. On a single day (May 9, 2017), 500,000 students took the online tests, the largest single day of such assessments ever" (California Department of Education [CDE], 2017). CAASPP's accomplishments certainly demonstrate the feasibility for technology-based delivery at scale. That said, it is important to note that most K–12 testing around the world—and, by examinee volume, much of K–12 assessment in the United States—continues to be done on paper.

In addition to the innovation these programs will bring, we should watch for the effect that technology can have on the meaning of assessment results. In particular, we should watch for variation in meaning over (1) mode in the transition from paper to digital delivery, (2) students with different levels of computer familiarity, (3) demographic groups, (4) time as technology changes, and (5) languages. Watch also for the substantive, technical, policy and political challenges that preserving meaning poses (e.g., when NAEP state comparisons are challenged because of questions about how the change in delivery mode differentially affected performance; Barnum, 2018). Finally, given that technology delivery is still novel in some testing programs, watch for its implementation to sometimes go awry (Herold, 2018a), with the resulting calls for retrenchment an inevitable part of moving forward.

Measure New Constructs

I expect that a second characteristic of future assessments will be the measurement of so-called new constructs. Why? Because there is far more needed for success in education, in the workforce, and for meaningful citizenship than we currently assess or teach. At the individual level, examples include problem-solving processes—for which technology will facilitate measurement (Zhang & Deane, 2015)—and a host of socioemotional competencies like "grit," social awareness, and self-awareness (Belfield et al., 2015). At the group level, team functioning or collaboration is often cited as a novel competency (Fiore et al., 2017). At the institutional level, such contextual factors as classroom and school climate are becoming increasingly critical to monitor and understand (Cohen, Pickeral, & McCloskey, 2008/2009; Thapa, Cohen, Guffey, & Higgins-D'Alessandro, 2013).

What should we watch? Watch the extent to which measures of these constructs are adopted for consequential decision-making in education (e.g., school accountability, postsecondary admissions). To date, very limited use has been made of such assessments because of the perception (and reality) that they are not yet ready (Duckworth & Yeager, 2015), the greatest concern being their susceptibility to manipulation. Also watch the extent of use for formative purposes, to which they may well be more suited in selected instances.

Be Built From Richer Underlying Models of Cognition and Learning

I believe that a third characteristic of future assessments will be their being built from richer underlying models of cognition and learning. These theory-based models offer more coherent organizations of subject matter than do curriculum standards. As a consequence, such models can be a more useful guide for assessment design and item writing (Bennett, Deane, & van Rijn, 2016). Of course, it's important to substantiate, supplement, and sometimes even discover these models through data. These models may also potentially allow for more meaningful measurement of change through learning progressions, which describe frequently used paths to proficiency for particular competencies (Corcoran, Mosher, & Rogat, 2009; Daro, Mosher, & Corcoran, 2011; Deane & Song, 2015).

What should we watch? Watch the extent to which tests built from such models function as the models predict (van Rijn, Graf, & Deane, 2014; Graf & van Rijn, 2016). For example, when a test is intended to measure one or more learning progressions, do the item difficulties align with the learning-progression levels for which the items were written? When a test is designed from an underlying theory, that theory's propositions can be evaluated each time the assessment is given (Bejar, 1993). We should expect, for instance, to find that the response patterns for most students line up with the pattern that would be predicted based on that underlying theory. Also watch the extent to which teachers find the models helpful as an aid to organizing instruction and guiding classroom assessment, something that learning progressions are claimed to do (Bennett et al., 2016; Heritage, 2008).

Make Greater Use of More Complex Tasks

Fourth, I would expect tests of the future to make greater use of more complex tasks. Why? Because the activities that characterize proficiency in a discipline often take the form of extended problem-solving episodes. These episodes have been very difficult to replicate in assessment settings, largely due to practical constraints. The disconnect between the tasks we typically use and the ones that define a discipline is, I think, one reason why only 58% of U.S. public school parents say that tests do a good job measuring how well their child is learning (Phi Delta Kappa [PDK], 2017). Our past attempts to approximate those extended activities have taken the form of essays, hands-on experiments and, much more rarely,

portfolios of work (Camp, 1993; Wolf, 1993). But more recently, extended activities in the form of technology-based simulations and games have been used or proposed (Gee & Shaffer, 2010; Mayrath, Clarke-Midura, Robinson, & Schraw, 2012).

What should we watch? For ideas and examples of task types to adopt or adapt, watch educational games and simulations; watch occupational and professional assessments, which include approximations of some of these extended activities (Dillon & Clauser, 2009); and watch NAEP, one of the few K–12 programs using simulation tasks, as well as other types of performances. Examples from NAEP can be found in its Interactive Computer Tasks administered during the 2009 science assessment and in the 2014 Technology and Engineering Literacy assessment (IES, 2010; NAGB, n.d.b).

But we should also be cognizant of old problems of breadth of coverage, fairness for groups and individuals, development and scoring cost, and the examinee time required to take such tasks (Shavelson, Baxter, & Gao, 1993; Stecher & Klein, 1997). We can expect new (or newly applied) approaches to mitigate some of those problems. For example, some programs combine performance tasks with sections of shorter questions, an approach employed for many years by the Advanced Placement Program and by NAEP. A newer idea is to use structured performance tasks, which break up an extended task into a shorter series of less locally dependent ones than might be found in a full-blown performance activity (e.g., Deane et al., 2018). Third, we should expect more sophisticated development tools to emerge (e.g., based on the design patterns of Evidence Centered Design; Mislevy & Haertel, 2006), as well as automated methods for scoring to be adopted, which in combination should lower front- and back-end costs.

Be "Personalized"

I expect tests of the future to be in some sense "personalized." The reasoning here is simple. Students come to education, and to assessment, with different levels and types of competency, diverse backgrounds, and varied interests. Our characterizations of what students know and can do might be improved if we could, somehow, better accommodate diversity. That idea, broadly speaking, is behind the movement towards personalized learning so often discussed in the education press (Burnette, 2017).

There are several dimensions along which assessment might be personalized, all of which bear watching. One such dimension is accessibility. Attempts to make assessments accessible go back a long way, at least to the Scholastic Aptitude Test's Braille and large-type editions, which first appeared in 1938 (Saretsky, 1983). Much more recently, Smarter Balanced Assessment Consortium introduced a wide array of built-in assistive technologies (Smarter Balanced Assessment Consortium, 2017), that have significantly raised the bar with respect to what assessment programs might be expected to provide. Still more recently, the Graduate Record Examinations (GRE) General Test deployed technologies examinees commonly use in their everyday interactions with computers (e.g., the most commonly used, commercially available screen reader instead of an unfamiliar one custom-created for the test) (Educational Testing Service [ETS], 2018).

Another way to personalize is through adaptive testing. In their simplest form, adaptive tests match item difficulty to estimated skill level, an idea dating to at least the Stanford-Binet Intelligence Scales of 1916 (Becker, 2003). A version of that notion based on item response theory (IRT) was implemented in the College Board's Computerized Placement Tests in 1986 (Ward, 1988), one of the first such adaptive measures. A more modern and challenging twist would be to try to match not only item difficulty to student ability, but item content to student background or interest. That type of matching might enhance engagement and also reduce the unfairness inherent in the person-by-task interaction commonly observed for performance tasks (Linn & Burton, 1994; Shavelson et al., 1993). This idea, of course, brings fairness issues of its own: What if the examinee doesn't agree with our view of what it is he or she should be interested in?

Yet another way to personalize is through examinee-determined problem choice. This type of personalization is exemplified in Section II, Part B of the Advanced Placement U.S. History examination (College Board, 2017, p. 34), which allows the student to choose from three essay questions differing on the time period in focus (e.g., Civil War, World War II, Vietnam War), but measuring the same reasoning skills. This type of personalization raises the validity concern of what happens if examinees choose poorly? The research, which is largely quite dated, gives mixed results (Powers & Bennett, 1999). Bridgeman, Morgan, and Wang (1997), for example, asked students to choose in advance the essay prompt they believed they would do better on but then had the students respond to both prompts. Although most students chose well, 3 in 10 individuals chose an essay on which they subsequently scored lower than on the essay they excluded.

A last way to personalize is on choice of goal or curriculum standard. A simple example is when students are allowed to select which subject test scores to submit for university admission. A more complex case is the Advanced Placement Studio Art 3-D Design Portfolio, Section II: Concentration, for which there is "no preferred (or unacceptable) style or content" (College Board, 2014, p. 17). Submissions may include figurative or nonfigurative sculpture, architectural models, metal work, glass work, installation, performance, assemblage, and 3-D fabric/fiber arts. In the extreme, then, the assessment is built to measure whatever specific goals the school and student negotiate. For AP Studio Art, this negotiation takes place within the constraints of common, high-level goals and evaluation criteria that become particularized to the student. The scoring challenge is approached through a detailed but general rubric, and (one presumes) very considerable rater training.

Attempt to Improve Learning

State accountability tests have traditionally been intended to help improve learning indirectly by giving information for policy action (e.g., identifying which schools need special attention or on which content standards teachers might benefit most from professional development). The value of such testing, however, is being increasingly questioned to the point that state assessments are perceived negatively by a significant segment of the population, at best as a waste of instructional time and at worst as harmful to students (PDK, 2017, pp. K23–K25). Such attitudes caused President Obama to call for capping the time devoted to federally mandated tests in 2015 (Associated Press [AP], 2015, para. 3). The Smarter Balanced Assessment Consortium, the Partnership for Assessment of Readiness for College and Careers

(PARCC), and many states have since shortened their assessments. In some states, students have opted out of testing altogether (Bennett, 2016). In New York State, for example, 19% of students in Grades 3–8 refused to take the state assessment in 2017 (Tyrrell, 2018).

What should we watch? Watch attempts to create tests that not only measure well, but include tasks designed to be instructional models for teachers and guides to learning for students (Bennett et al., 2016). Look for tests that try to cause incidental learning for students by helping them become more informed about an important topic (Bennett, 2010, p. 76). Finally, watch for tests that provide qualitative feedback to encourage student self-reflection about their problem-solving processes (e.g., how a student composed his or her essay or conducted a simulated science experiment).

Be Better at Accounting for Context

Seventh, I believe tests of the future will be better at accounting for context. Large-scale summative tests are designed to assess "out of context," ignoring the social, learning, and teaching environment for an individual or group in an attempt to produce inferences generalizable across many contexts. How a student or group performs on such a test is a fact. But *why* the student performed that way is an *interpretation* requiring knowledge of context. We attempt to account for context in only the most basic ways, however, by using devices such as school, teacher, and student background data questionnaires.

Well worth watching are electronic learning environments that bring embedded assessment with them because embedding assessment *into* the learning context ought to make results more actionable since the content, knowledge representations, and tools called for by the assessment are the same as used in the learning environment (Bennett, 2015). Such embedding essentially makes the context of assessment and instruction identical.

Also worth watching are testing programs that try to become more "embedded" themselves.

Why would a testing program try to become (or appear to become) more embedded? A reason just cited is to account better for the learning context. But a second reason a testing program might be attracted to this idea is to prevent electronic learning (and other technology) companies from displacing it. Such displacement, or disruption, can occur when companies from related industries become competitors offering simpler, lower-cost alternatives (Christensen, 1997).

What would it mean for a testing program to become embedded? One meaning is to become a mechanism for gathering a casual (but extensive) sampling, distributed over time, of whatever students are doing (or learning) in school (or elsewhere). I say, a "sampling," because we can never record everything that occurs. I describe this sampling as "casual" since it wouldn't be *designed* to provide evidence of particular competencies in a given range of proficiency to a known standard of quality. Rather, it would be an incidental sampling. This incidental sampling would, however, be "extensive" in that it would provide lots of data. It would be essentially what would occur if all students used a common electronic learning environment for educational activity, or if data from all the environments students used could be assembled together. In the extreme case, it would be fully instrumented learning—all keystrokes and mouse clicks recorded, with sensors, cameras, and microphones distributed throughout the learning space. It would be fully instrumented learning leading to very *Big Data* leading to . . . what?

That is, how might we use a casual (but extensive) Big Data sampling of behavior? One way in which we might use such a sampling is descriptively. That is, we could use it to exemplify exactly what students were doing or learning. I think this use would be relatively easy to accomplish. It would also be extremely valuable. It would allow us, for the first time, to describe in exquisite detail how instruction differed from one classroom to the next, from one teacher to the next, from school to school, district to district, and demographic group to demographic group. It would allow us to link outcome data, like value-added, to hypothesized causal mechanisms in ways we now cannot.

An additional use we could propose for this incidental sampling of behavior is inferential, that is, to draw comparable conclusions about what students know and can do. That use, I believe, would be very difficult to engineer meaningfully, given the incidental nature of the data.

However, a second meaning of "embedded" might be a recording of what occurs in response to a series of designed events inserted into the curriculum at specified points. Imagine that each such event took the form of what we will call a "learning challenge." That challenge could occur over one or more classroom periods. It could involve items, games, simulations, or other performance tasks—whatever evidence-gathering opportunities fit the intended assessment claims and use-case best. These periodic assessments would, of course, provide a much less extensive sampling of behavior. They would be less attuned to the learning context because they would be coming from outside of that context. But they would be *designed*, arguably supporting stronger and more comparable inferences about what students knew and were able to do with respect to broad competencies of interest than would a more casual sampling.

Watch those entities that try to collect data through both meanings of embedded. Look, in particular, for attempts to combine information from casual, fully embedded approaches with data from designed, contextually more distant methods—attempts that use the results of one approach to complement the results of the other, and that investigate causes of disagreement, perhaps leading to improvements in both approaches. Also watch parental, educator, and policy maker concerns about privacy with respect to the continuous recording of teacher and student behavior (Herold, 2018b).

Use Automated Scoring

Ninth, future assessments are likely to depend more heavily on automated scoring. Such scoring will allow greater use of complex tasks, increase scoring efficiency in time and cost, and provide more detailed feedback about performance.

Warranting special attention will be black-box algorithms. These algorithms can often predict with reasonable accuracy how a human judge would have scored or classified some types of response like an essay or speech sample (Bennett & Zhang, 2016). Black-box algorithms may be fine for formative purposes, where the consequences of wrong decisions are relatively low and the decisions themselves are more easily reversed. Such algorithms are more problematic for high-stakes decisions because they use inscrutable or

otherwise proprietary methods—hence the name, "black box." In other words, we may not know how they arrive at an individual's score. For prediction, they may be using correlates without regard to the justification of those correlates in terms of the target construct. As a consequence, improving standing on the correlates will raise test score but not necessarily imply an increase on the target competency. Essay length is an example that appears to be used, directly or indirectly, in most automated scoring systems. In such a case, adding well-formed, grammatically correct, but vacuous text, will increase one's score but not make one's writing better.

With respect to black-box technology, we should watch for legal action like the European Union's General Data Protection Regulation (GDPR; European Commission, n.d.). That regulation has received much attention because it has caused many U.S. companies to revise their privacy policies, resulting in notifications to millions of customers. Less well known is that as of 2018, GDPR gives individuals the right to an explanation of an algorithmic decision that significantly affects them (Meyer, 2018). One can imagine how the regulation of algorithmic decisions might be important in the awarding of consumer loans, for example. Regulations like GDPR are encouraging the emergence of the field of XAI (explainable artificial intelligence; Kuang, 2017; Maglieri & Comande, 2017).

Following from the GDPR, the goal for automated scoring ought to be to make it explainable in ways that align with construct definition. As an example, for a test intended to measure argumentative writing skill, the automated scoring should focus on such essay response characteristics as the extent to which reasons support the stated position, the strength of evidence backing each reason, and the quality of rebuttal of likely counter-arguments. Those who have experience with artificial intelligence will know that this type of analysis is very difficult for a machine to do currently. But, as Tom Hanks said to Madonna in the film *A League of Their Own*, "If it wasn't hard, everbody'd be doing it. It's the hard that makes it great!" Watch for the "great."

Incorporate New Approaches to Modeling and Analysis

Tenth is that assessment of the future will almost certainly need to incorporate new approaches to modeling and analysis. New types of data from online learning and online assessment are emerging. A salient example is process data (Cope & Kalantzis, 2016). With technology-based assessment, we can capture the type of action an examinee takes, when that action occurred, how long it lasted, and what preceded and followed it, all of which may hold useful information. Traditional psychometric models were developed, of course, for much simpler data, dichotomously or polytomously scored item responses generated under much more restrictive assumptions.

Worth attending to are advances coming from several fields. These fields include educational data mining (EDM) and learning analytics (LA), both concerned with using big data in instructional applications and for advancing the science of learning (Siemens & Baker, 2010). Finally, we should continue to watch statistics, from which significant segments of modern measurement have evolved (Braun, 2017), and which remains a primary generator of new methods taken up in the above arenas.

Provide More Effective Reporting

Finally, I expect tomorrow's assessments to provide more effective reporting for at least three reasons. First, reporting is an important part of the user experience, with the desire for a useful result often being a motivation for taking a test in the first place. Second, reporting is a vastly underutilized mechanism for positive impact. It is a potential path toward guiding the thinking and actions of teachers, students, parents, and policy makers, as well as a route to improving perceptions of testing. Finally, the state of the art in reporting has progressed so slowly relative to other aspects of testing. Consider what we have in operational testing programs today: electronic delivery to millions of students in CAASPP; adaptive testing in Smarter Balanced and the GRE General Test; simulation tasks in NAEP and in the United States Medical Licensing Examination; and automated scoring in the GRE Analytical Writing section, the Test of English as a Foreign Language, and Smarter Balanced. For reporting, we have seen far less innovation in operational programs.[1]

What should we watch? Look for reporting to be more tailored to examinees and to the intended assessment use. As an example, look for performance replay, which is used universally by sports coaches and athletes to analyze what an athlete or team did and to help them identify how to improve. Performance replay might be very useful for problem-solving instances where the process is part of what's judged. The scientific practice of planning and carrying out investigations (National Research Council [NRC], 2012) offers an example: the end result of an experiment is made credible or, alternatively, immediately dismissible to the degree that the process used to obtain it was defensible. Performance replay might also be valuable in situations where the process is *not* what is being judged. In these instances, if the end product is deficient, a look at the process might suggest means for improvement. Writing would be an example. Watching a replay of how an essay was composed could make clear that the student began entering text very soon after encountering the prompt, with no evidence of planning; that the student did no editing; or that he or she submitted the essay following the last keystroke, leaving no time for rereading. Observing and reflecting on that behavior could be helpful to both student and teacher. Similarly, reviewing samples of the modal behavior of a group might assist policy makers and the public in understanding more concretely what aggregated test results mean.

In addition to performance replay, we should look for reports that try to educate users in simple, intuitive, and engaging ways (e.g., by using game elements for students). PISA goes in a very good direction in terms of simplicity and ease of use. Figure 1 gives a screen showing results from the 2015 Science assessment. Countries are color-coded in nine categories according to performance level, allowing the viewer to quickly see which countries participated and where they fell in the achievement distribution. Clicking on any participating country leads to a detailed report on that and other assessments administered in the cycle.

What Is Unlikely to Change

Those are my thoughts about how assessments of the future are likely to be different. But what is unlikely to change? Unlikely to change are the fundamental characteristics that define assessment. At a high level, assessment is about only four things. First, it is about engineering opportunities to

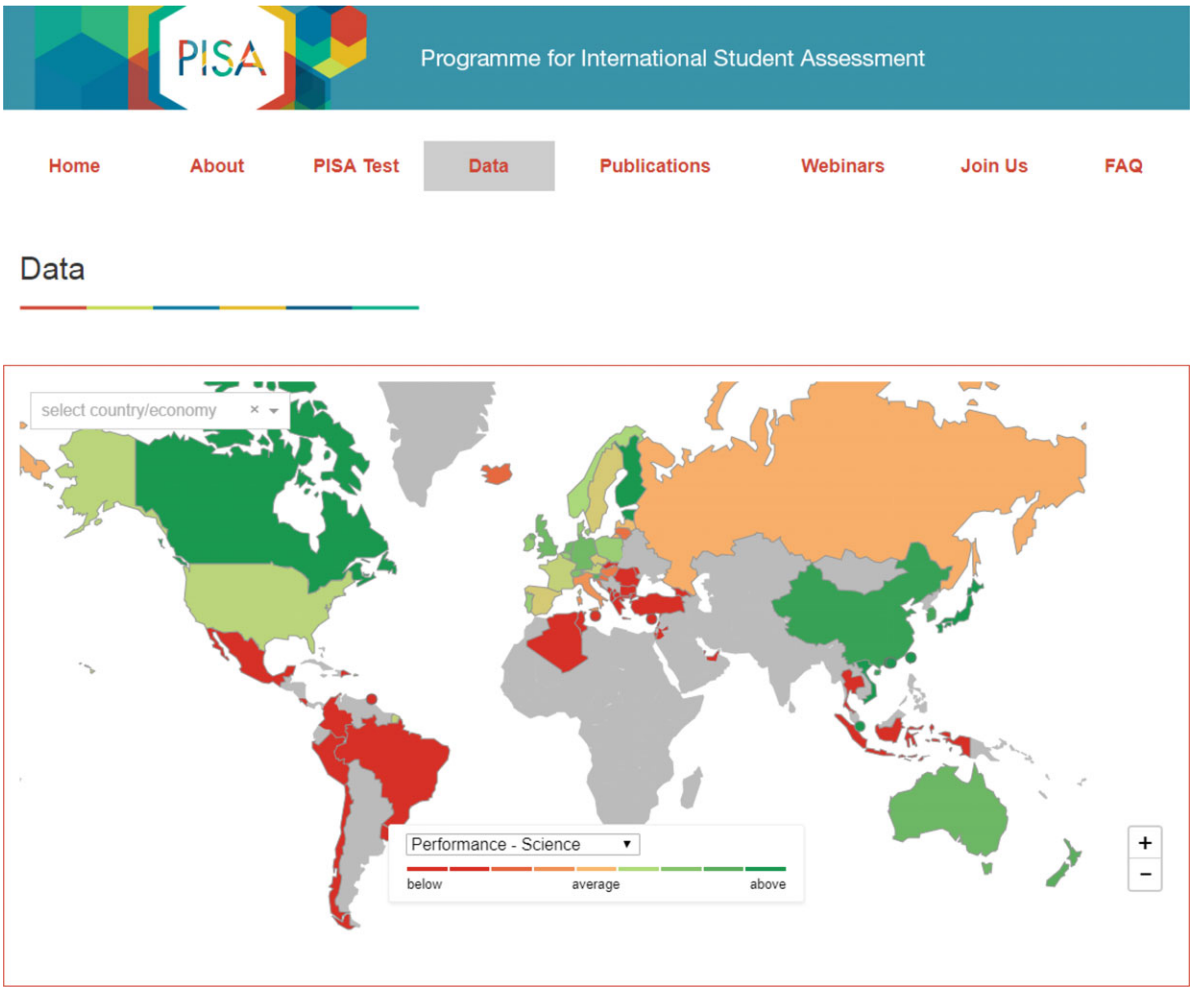

FIGURE 1. Interactive PISA reporting screen showing results from the 2015 Science assessment. ©OECD 2017. Used by permission. [Color figure can be viewed at wileyonlinelibrary.com]

observe evidence of the competencies we wish to make claims about and then making those observations. Narrowly construed, these activities are test design, item creation, and test administration. A second fundamental characteristic is connecting the evidence we observe to meaningful characterizations of individuals, groups, or institutions (along with estimates of the uncertainty associated with those characterizations). This activity is measurement modeling. Third is communicating results for use in decision making, or reporting. The last characteristic is evaluating the quality and impact of the evidence-gathering opportunities, characterizations, and decisions—or validation. At this level of description, assessment is likely to remain much the same.

What is also unlikely to change are the big social problems toward which assessment is directed. These problems include documenting effectiveness of education systems, monitoring achievement gaps among important social groups, providing information for resource allocation to individuals (e.g., admissions to top-tier universities), and helping improve learning and teaching. These social problems have been with us for decades. There can be little doubt that they will persist.

Unlikely to change also are the social values that underlie assessment (Messick, 1994). Those social values include validity, fairness, comparability, and reproducibility, a more generally accessible term than "reliability" for "show me again." If we can't generate a similar result by using another sample of behavior, a different rater, or another temporally close occasion, the results won't be credible.

Finally, what is unlikely to change is the need for distinct summative and formative approaches to assessment. Of course, summative assessments may be able to (secondarily) provide some formative information (e.g., performance replay when items are disclosed; Bennett, 2011). Similarly, formative assessments may be able, in some instances, to (secondarily) contribute to judgments of what students know and can do by providing qualitative evidence of learning in a specific context. A periodically proposed idea is to replace high-stakes summative assessment entirely with quantitative data collected from assessment embedded in classroom learning activity (e.g., Bennett, 1998, pp. 11–14; Gee & Shaffer, 2010; Pellegrino, Chudowsky, & Glaser, 2001, pp. 283–287; Tucker, 2012). I believe that this use is unlikely to happen for several reasons.

One reason is the dramatic variation in the type and quality of data that exists across districts, schools within districts, and even classes within schools. This variation occurs because of local differences in learning goals, in the curricula employed to achieve those goals, in the e-learning resources used with those curricula, and in the quality and types of evidence the different resources collect. That dramatic variation reduces

the possibilities for making inferences about student proficiency that are comparable across students, classes, schools, and districts.

A second reason is that it is intuitively appealing to policy makers and the public for measures of school quality to be at least somewhat independent of the school being measured. This appeal is motivated by the belief that the more the measure is part of the fabric of everyday learning at a given school, the less valuable that measure would be as an indicator of how well a student would be able to apply what was learned if moved to another school, or to other contexts more generally. The more local the measure, the more local the interpretation may need to be.

Third are the privacy and political concerns that attend to the continuous monitoring of student and teacher behavior. A long series of breaches of personal information held by private companies and governments (Fung, 2018; Larson, 2017; McCoy, 2017), coupled with high-profile instances of inappropriate data use (e.g., Granville, 2018), have contributed to a climate of institutional distrust extending to schools, state and federal education agencies, and education companies (Herold, 2014).

Last are potential negative effects on teaching and learning. Learning often involves experimentation and that experimentation inevitably leads to failure, some of which may be productive (Kapur, 2010). A negative side effect of using the continuous collection of student and teacher performance data for consequential decision making could well be to discourage attitudes and habits of mind, like risk taking, that facilitate learning. A somewhat similar concern is that the absence of a culminating test will have the unintended side effect of removing an opportunity for practice and for the consolidation of learning. To the extent the test is a good representation of content standards and teachers prepare students for it by broadly instructing to those standards, preparing for that culminating assessment should be a desirable activity. That ideal, unfortunately, was compromised considerably by associating with test performance such rewards and punishments for educators as promotion, tenure, bonuses, and dismissal. That association narrowed the focus of instruction, raised anxiety for students and parents, and resulted in a public backlash against testing (Bennett, 2016). As state requirements for basing teacher evaluation on student test scores are scaled back (Loewus, 2017), and if state assessments can become more effective in their representation of standards, perhaps a more balanced approach to teaching can emerge in which the summative assessment helps encourage good instructional and learning practices.

Summary

To reiterate the main points of this article, we should expect change along a number of dimensions:

- the competencies we consider important to measure (e.g., the addition of socioemotional learning);
- the nature of the opportunities we engineer to observe evidence of those competencies (e.g., using more complex tasks built from richer learning models; employing more in-context observations);
- how we connect evidence to characterizations (e.g., through new quantitative models);
- how we communicate results for use in decision making via better, more interactive reporting; and
- how we evaluate quality and impact by, for example, giving greater attention to the extent to which tests have positive effects on teaching and learning, and on public perceptions of testing.

At the same time, we should expect assessment fundamentals, the big social problems toward which assessment is directed, the social values underlying assessment, and the distinctions between major assessment types to endure.

Looking forward to education itself 10 years from now, we should also anticipate considerable change. That change is bound to affect the purposes for which education is conducted (e.g., to prepare students better for citizenship), the goals associated with existing purposes as the nature of work evolves in our society, and the methods of teaching and learning due to the increasing use of technology. Obviously, assessment must follow suit because, if it does not, it will become an anachronism. To remain relevant, assessment community members will need to think out-of-the-box, experiment, adopt what works, and always remain focused on core values and fundamental principles. Without that focus on core values and fundamental principles, we risk chasing fads that, while perhaps profitable and exciting in the short-term, end up compromising teaching and learning in the long run.

Acknowledgments

This article is adapted from the Presidential address given at the annual meeting of the National Council on Measurement in Education, New York, April 2018. I appreciate the comments of Bob Mislevy, Andreas Oranje, and Rebecca Zwick on an earlier version.

Notes

[1]See Zapata-Rivera (2019) for a recent review of research and applications.

References

Associated Press (AP). (2015). Obama calls for capping time devoted to standardized tests. Retrieved from http://www.pbs.org/newshour/rundown/obama-calls-cap-class-time-devoted-standardized-tests/

Barnum, M. (2018). Did computer testing muddle this year's NAEP results? Testing group says no; others are unconvinced. *Chalkbeat*. Retrieved from https://www.chalkbeat.org/posts/us/2018/04/10/did-computer-testing-muddle-this-years-naep-results-testing-group-says-no-others-are-unconvinced/

Becker, K. (2003). *History of the Stanford-Binet Intelligence Scales: Content and psychometrics* (Stanford-Binet Intelligence Scales, Fifth Edition Assessment Service Bulletin No. 1). Itasca, IL: Riverside Publishing. Retrieved from https://www.hmhco.com/~/media/sites/home/hmh-assessments/clinical/stanford-binet/pdf/sb5_asb_1.pdf?la=en

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Lawrence Erlbaum.

Belfield, C., Bowden, B., Klapp, A., Levin, H., Shand, R., & Zander, S. (2015). *The economic value of social and emotional learning*. New York: Teachers College, Columbia University. Retrieved from http://blogs.edweek.org/edweek/rulesforengagement/SEL-Revised.pdf

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Policy

Information Center, Educational Testing Service. Retrieved from https://www.ets.org/research/policy_research_reports/pic-reinvent

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*, 70–91. https://doi.org/10.1080/15366367.2010.508686

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice 18*, 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*, 370-407. https://doi.org/10.3102/0091732X14554179

Bennett, R. E. (2016). *Opt out: An examination of issues (RR-16-13)*. Princeton, NJ: Educational Testing Service. http://doi.org/10.1002/ets2.12101

Bennett, R. E., Deane, P., & van Rijn, P.W. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist*, *51*, 82–107. http://doi.org/10.1080/00461520.2016.1141683

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007–466)*. Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid = 2007466

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). New York, NY: Routledge.

Braun, H. I. (2017). Research on statistics. In R. E. Bennett & M. Von Davier (Eds.), *Advancing human assessment: The methodological, psychological, and policy contributions of ETS*. Cham, Switzerland: Springer Open.

Bridgeman, B., Morgan, R., & Wang, M. (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement*, *34*, 273–286. https://doi.org/10.1111/j.1745-3984.1997.tb00519.x

Burnette, D. II (2017). States take steps to fuel personalized learning. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2017/11/08/states-take-steps-to-fuel-personalized-learning.html

California Department of Education (CDE). (2017). State schools chief Tom Torlakson announces results of California Assessment of Student Performance and Progress online tests (Release #17-67a). Retrieved from https://www.cde.ca.gov/nr/ne/yr17/yr17rel67a.asp

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement* (pp. 183–212). Hillsdale, NJ: Lawrence Erlbaum.

Christensen, C. M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail*. Boston, MA: Harvard Business School Press.

Cohen, J., Pickeral, T., & McCloskey, M. (2008/2009). The challenge of assessing school climate. *Educational Leadership*, *66*(4). Retrieved from http://www.ascd.org/publications/educational-leadership/dec08/vol66/num04/The-Challenge-of-Assessing-School-Climate.aspx

College Board. (2014). *Studio art course description*. New York, NY: Author. Retrieved from https://apcentral.collegeboard.org/pdf/ap-studio-arts-course-description.pdf?course=ap-studio-art-3-d-design

College Board. (2017). *AP United States history practice exam: From the course and exam description*. New York, NY: Author. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-united-states-history-ced-practice-exam.pdf

Cope, B., & Kalantzis, M. (2016). Big data comes to school: Implications for learning, assessment, and research. *AERA Open*. https://doi.org/10.1177/2332858416641907

Corcoran, T., Mosher, F. A., & Rogat A. (2009). *Learning progressions in science: An evidence-based approach to reform*. New York, NY: Consortium for Policy Research in Education (CPRE).

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction (RR-68)*. Philadelphia, PA: CPRE.

Deane, P. D., & Song, Y. (2015). *The key practice, discuss and debate ideas: Conceptual framework, literature review, and provisional learning progressions for argumentation (RR-15-33)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12079

Deane, P. D., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M., Bennett, R. E., . . . Zhang, M. (2018). The case for scenario-based assessment of written argumentation. *Reading and Writing*. Advance online publication. https://doi.org/10.1007/s11145-018-9852-7

Dillon, G. F., & Clauser, B. E. (2009). Computer-delivered patient simulations in the United States Medical Licensing Examination (USMLE). *Simulation in Healthcare*, *4*(1), 30–34. https://doi.org/10.1097/SIH.0b013e3181880484

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Educational Testing Service (ETS). (2018). Accommodations for test takers with disabilities or health-related needs. Retrieved from https://www.ets.org/gre/revised_general/register/disabilities?WT.ac=rx28

European Commission. (n.d.). 2018 reform of EU data protection rules. Retrieved from https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., . . . von Davier, A. (2017). Collaborative problem solving: Considerations for the National Assessment of Educational Progress. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/researchcenter/collaborative_problem_solving.pdf

Fung, B. (2018). Equifax's massive 2017 data breach keeps getting worse. *Washington Post*. Retrived from https://www.washingtonpost.com/news/the-switch/wp/2018/03/01/equifax-keeps-finding-millions-more-people-who-were-affected-by-its-massive-data-breach/?utm_term=.6a07f261799d

Gee, J. P., & Shaffer, D. W. (2010). Looking where the light is bad: Video games and the future of assessment. *Edge*, *6*(1), 3–19. Retrieved from http://edgaps.org/gaps/wp-content/uploads/EDge-Light.pdf

Graf, E. A., & van Rijn, P. W. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd Ed.) New York, NY: Routledge.

Granville, K. (2018). Facebook and Cambridge Analytica: What you need to know as fallout widens. *New York Times*. Retrieved from https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: CCSSO. Retrieved from http://www.k12.wa.us/assessment/ClassroomAssessmentIntegration/pubdocs/FASTLearningProgressions.pdf

Herold, B. (2014). inBloom to shut down amid growing data-privacy concerns. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/DigitalEducation/2014/04/inbloom_to_shut_down_amid_growing_data_privacy_concerns.html

Herold, B. (2018a). Online state testing in 2018: Mostly smooth, with one glaring exception. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2018/05/09/online-state-testing-in-2018-mostly-smooth.html

Herold, B. (2018b). How (and why) ed-tech companies are tracking students' feelings. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2018/06/12/how-and-why-ed-tech-companies-are-tracking.html?cmp=eml-enl-eu-news1&M=58515710&U=1605540

Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based*

Assessment Project (NCES 2005–457). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Institute for Education Sciences (IES). (2010). Hands-on tasks (HOT) and interactive computer tests (ICT) for the 2009 science assessment. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/sample_design/2009/2009_samp_nation_science_hot_ict.aspx

Institute for Education Sciences (IES). (2012). *The nation's report card: Writing 2011*. Washington, DC: Author. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid = 2012470

Institute for Education Sciences (IES). (2018). Writing assessment. Retrieved from https://nces.ed.gov/nationsreportcard/writing/

Institute for Education Sciences (IES). (n.d.). *Progress in International Reading Literacy Study (PIRLS)*: Countries. Retrieved from https://nces.ed.gov/surveys/pirls/countries.asp

Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science*, *38*, 523–550.

Kuang, C. (2017). Can A.I. be taught to explain itself? *New York Times Magazine*. Retrieved from https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Larson, S. (2017). Every single Yahoo account was hacked—3 billion in all. *CNN*. Retrieved from http://money.cnn.com/2017/10/03/technology/business/yahoo-breach-3-billion-accounts/index.html

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5–8. https://doi.org/10.1111/j.1745-3992.1994.tb00778.x

Loewus, L. (2017). Are states changing course on teacher evaluation? *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2017/11/15/are-states-changing-course-on-teacher-evaluation.html

Maglieri, G., & Comande, G. (2017). Why a right to legibility of automated decision-making exists in the General Data Protection Regulation. *International Data Privacy Law*, *7*, 243–265. https://doi.org/10.1093/idpl/ipx019

Mayrath, M. C., Clarke-Midura, J., Robinson, D. H., & Schraw, G. (Eds.). (2012). *Technology-based assessments for 21st century skills*. Charlotte, NC: Information Age.

McCoy, K. (2017). Target to pay $18.5M for 2013 data breach that affected 41 million consumers. *USA Today*. Retrieved from https://www.usatoday.com/story/money/2017/05/23/target-pay-185m-2013-data-breach-affected-consumers/102063932/

Meyer, D. (2018). AI has a big privacy problem and Europe's new data protection law is about to expose it. *Fortune*. Retrieved from http://fortune.com/2018/05/25/ai-machine-learning-privacy-gdpr/

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment*, *10*(1), 1–9.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Mullis, I. V. S., & Prendergast, C.O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.). *Methods and procedures in PIRLS 2016* (pp. 1.1–1.29). Chestnut Hill, MA: IEA.

National Assessment Governing Board (NAGB). (n.d.a). 2017 NAEP mathematics and reading assessments. Retrieved from https://www.nationsreportcard.gov/reading_math_2017_highlights/

National Assessment Governing Board (NAGB). (n.d.b). 2014 NAEP technology and engineering literacy (TEL). Retrieved from https://www.nationsreportcard.gov/tel_2014/

National Assessment Program (NAPLAN). (2018). FAQs. Retrieved from https://www.nap.edu.au/online-assessment/FAQs

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

Organization for Economic Cooperation and Development (OECD). (2017). *PISA technical report*. Paris, France: Author. Retrieved from http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Phi Delta Kappa (PDK). (2017). *The 49th annual PDF poll of the public's attitudes toward the public schools*. Arlington, VA: PDK International. Retrieved from http://pdkpoll.org/assets/downloads/PDKnational_poll_2017.pdf

Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education*, *12*, 257–279. https://doi.org/10.1207/S15324818AME1203_3

Sandene, B., Bennett, R. E., Braswell, J., & Oranje, A. (2005). Online assessment in mathematics. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project (NCES 2005–457)*. Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Saretsky, G. (1983). SATs for the blind offered 45 years ago. *Examiner*, *13*(7), 3.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Siemens, G., & Baker, R. S. J. d. (2010). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. Retrieved from http://www.upenn.edu/learninganalytics/ryanbaker/LAKs%20reformatting%20v2.pdf

Smarter Balanced Assessment Consortium. (2014). Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Stecher, B., & Klein, S. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis*, *19*(1), 1–14. https://doi.org/10.3102/01623737019001001

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, *83*, 357–385. https://doi.org/10.3102/0034654313483907

Tucker, B. (2012). Grand test auto: The end of testing. *Washington Monthly*. Retrieved from http://www.washingtonmonthly.com/magazine/mayjune_2012/special_report/grand_test_auto037192.php

Tyrrell, J. (2018). It's test time for NY students in Grades 3–8. *Newsday*. Retrieved from https://www.newsday.com/long-island/education/ela-math-opt-out-test-refusals-1.17901520

van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Psicologia Educativa*, *20*, 1–7. https://doi.org/10.1016/j.pse.2014.11.004

Ward, W. C. (1988). The College Board Computerized Placement Tests: An application of computerized adaptive testing. *Machine-Mediated Learning*, *2*, 271–282.

Wolf, D. P. (1993). Assessment as an episode of learning. In R. E. Bennett & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement* (pp. 213–240). Hillsdale, NJ: Lawrence Erlbaum.

Zapata-Rivera, D. (Ed.). (2019). *Score reporting research and applications*. New York, NY: Routledge.

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features (RR-15-27)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12075