The Future of Assessment in Technology-Rich Environments: Psychometric Considerations

Kristen E DiCerbo, Valerie Shute, and Yoon Jeon Kim

Abstract

A number of assessment experts have advanced a vision of assessment in schools that relies on ongoing, performance-based, formative assessment. While there are many potential benefits of such a system, it also raises concerns about assessment quality. This chapter is a review of the current state of the evidence for psychometric properties undergirding the vision of ongoing assessment in technology-rich environments. We discuss how reliability, validity, and fairness can be examined in individual instances of assessment in technology-rich environments (e.g., game-based assessment, simulation environments) and in a potential system of ongoing assessments covering large domains. The review suggests two areas of need in the area of ongoing assessment research: (1) modification of conceptualizations and procedures for establishing evidence of reliability, validity, and fairness to incorporate new, ongoing types of assessment and (2) collection of broader evidence to support arguments for their psychometric soundness.

Keywords

Ongoing assessment • Validity • Reliability • Fairness

K.E. DiCerbo (✉)
Pearson, Phoenix, AZ, USA
e-mail: Kristen.DiCerbo@Pearson.com

V. Shute
Florida State University, Tallahassee, FL, USA
e-mail: vshute@fsu.edu

Y.J. Kim
MIT, Cambridge, MA, USA
e-mail: yjkim82@gmail.com

J.M. Spector et al. (eds.), *Learning, Design, and Technology*,
DOI 10.1007/978-3-319-17727-4_66-1

Contents

Introduction

Imagine an educational system, maybe 10–20 years hence, where students would be immersed in many different learning contexts, all of which capture and measure their dynamic growth in knowledge and skills (both cognitive and noncognitive), and then the system uses the information to further enhance their learning. In this complex, interconnected, digital world, people are learning constantly and producing copious digital footprints or data.

This vision, described in Shute, Leighton, Jang, and Chu (2016), does not involve administering assessments more frequently (e.g., each week, each day) but, rather, continually collecting data as students interact with digital environments. It relies on what Shute (2011) calls stealth assessment, the process by which data are unobtrusively gathered while students are playing/learning in carefully designed environments that then allow inferences to be made about relevant competencies. As the various data streams coalesce, we obtain more evidence about what students know and can do across multiple contexts. The vision of assessment in technology-rich environments involves high-quality, ongoing, unobtrusive assessments that can be aggregated to inform a student's evolving competency levels (at various grain sizes) and also aggregated across students to inform higher-level decisions (e.g., from student to class to school to district to state to country). The key aspects of assessment in this vision include that it is:

- Ongoing – Assessment is not a single event. Rather, evidence is gathered over time (i.e., continuous) and across contexts (i.e., ubiquitous). Estimates of students' knowledge, skills, and other attributes are continually updated based on

multiple observations in diverse contexts rather than a single observation at one point in time.
- Performance based – Students are asked to complete tasks or produce complex responses and show the processes involved with solving meaningful problems.
- Formative – Information derived from the assessment is meant to inform instructional decisions and provide scaffolding and other support to learners.

When all three of these features come together in an assessment, with the goal to improve learning (e.g., Shute (2009)), we have a vision of a new kind of educational assessment. Throughout this chapter, we refer to this vision simply as *ongoing assessment* in technology-rich environments.

Benefits of Ongoing Assessment in Technology-Rich Environments

In the vision of assessment described above, the time spent administering tests, handling makeup exams, and going over test responses is not particularly conducive to learning. If we were to eliminate testing, according to Nelson (2013), we could add from 20 to 40 min of instruction per day to school. Given the importance of time on task as a predictor of learning, reallocating the current testing time into activities that are more educationally productive is a potentially large benefit that would apply to almost all students in all classes.

Second, by having assessments that are continuous and ubiquitous, students are no longer able to "cram" for an exam. Although cramming provides good short-term recall, it is a poor route to long-term retention and transfer of learning. Decades of research confirm that spaced practice is superior to massed practice. Thus, many current assessment practices in school lead to assessing students in a manner that is in conflict with their long-term success. With an ongoing assessment model in place, the best way for students to do well is to do well every day. By moving students toward a model where they will retain more of what they learn, we are enabling them to better succeed in cumulative domains, such as mathematics and science.

The third direct benefit is that this shift addresses growing concerns of an increasing number of educators and parents regarding pencil-and-paper high-stakes tests (Kamenetz, 2015). Our vision is essentially ongoing assessment without tests. Assessment is a general term that includes testing. Although progress toward educational goals is typically assessed through testing, we believe it can be evaluated without the instrument we typically recognize as a test. That is, rather than engaging in the traditional teach-stop-test model of gathering information, we can make inferences about student ability based on evidence from ongoing assessments.

Concerns About Ongoing Assessment in Technology-Rich Environments

Shute et al. (2016) identify four hurdles to this vision of assessment: assessment quality, identification of learning progressions, data privacy and security, and impediments to flipped classrooms. In this chapter we focus on the first hurdle, i.e., issues of assessment quality. To quote Samuel Messick (1994), "such basic assessment issues as validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made" (p. 13).

Concerns about the use of formative- and performance-based assessment in the classroom have centered on the lack of standardization in task requirements, task presentation, and scoring, resulting in weak claims of validity and reliability (Shavelson, Baxter, & Gao, 1993). Some may argue that ongoing formative assessments should not be required to show evidence of sound psychometric properties in similar ways to high-stakes summative assessments. However, the validity and reliability of the assessment data affect the accuracy of the student diagnosis, and the diagnosis informs instructional support. If the first part of the chain is weak, the rest (i.e., diagnostic accuracy and effective instructional support) would consequently be compromised (Shute & Zapata-Rivera, 2010). In addition, the fairness of assessment impacts the likelihood that different groups will differentially be given different instructional interventions, so this is necessary as well.

The use of digital technology presents new opportunities to standardize presentation and score activities, which might improve reliability of scores. It also offers the promise of allowing for assessment activities to be closely aligned to real-world activities and performances, which in return can improve the validity of scores. However, the impact of these affordances on reliability, validity, and fairness must be investigated rather than assumed. Many of the techniques used for measuring the psychometric properties of assessment were developed in the context of standardized assessment, consisting of discrete items specifically designed to assess a single construct, scored as correct or incorrect. Much of the evidence gathered from assessment in technology-rich environments (e.g., time spent, sequence of events) is not scored as correct/incorrect and often relates to multiple constructs. In addition, there is often variability even in what activity is presented to learners. In game-based assessment, for example, players' previous choices and actions, and the immediate feedback received, will result in differing activity sequences from player to player and from play occasion to play occasion for the same player (Kim & Shute, 2015). As a result of these differences, in some cases the models and techniques for estimating particular types of reliability and validity evidence will need revision to be applicable for new types of activity and data.

Given the changes described above, a review of the current state of the evidence for psychometric properties undergirding our vision of ongoing assessment in technology-rich environments is warranted. The next sections of this chapter discuss how reliability, validity, and fairness can be examined in individual instances of

ongoing assessment (e.g., game-based assessment, simulation environments) and in a potential system of such assessments covering large domains.

Reliability

In general, reliability refers to whether a test yields stable, consistent estimates of the construct under consideration. As defined by classical test theory, a score observed on a test is made up of the learner's true score plus error. Therefore, reliability can be viewed as precision, or the extent to which scores are free of measurement errors (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014). In this section, we examine traditional (e.g., interrater reliability, test-retest reliability) and contemporary views on reliability (e.g., reliability coefficients, generalizability) and discuss implications of those methods for ongoing assessment.

Interrater Reliability

Although agreement between human raters can be reached for relatively complex performance assessments, it requires continual monitoring and calibration. In most performance-based formative assessments happening day to day in classrooms, this calibration does not occur. As a result, different teachers may score the same performance in different ways (or even different performances in the same way), introducing error into the estimates of students' proficiency. Alternately, ongoing assessment in rich technology environments relies almost entirely on automated scoring of student work products, with consistent rules applied, reducing the error introduced by human variability but raising the question of whether the automated scoring procedures are as good as the best human judgments.

Substantial work has gone into efforts to create automated methods of scoring, particularly in scoring written essays (e.g., Williamson et al., 2010). For example, Foltz, Laham, and Landauer (1999) use latent semantic analysis (LSA), which is both a computational model of human knowledge representation and a method for extracting the semantic similarity of words and passages from text. To assess essay quality, LSA is first trained on domain-representative text. Then student essays are characterized by LSA representations of the meaning of their contained words and compared with essays of known quality on degree of conceptual relevance and amount of relevant content. Over many diverse topics, the scores from automated scoring programs were as similar to those from human experts as scores from experts were to each other (Williamson et al., 2010) without suffering from human rater issues such as halo effects, drift, and fatigue.

Recently, machine learning models (referred to as detectors) have been used with a number of other types of performances to broaden the potential of automated scoring. In general, the detectors are trained to model the best human rater judgments and can then be deployed to reliably score subsequent student performances. For

example, in Physics Playground (formerly known as Newton's Playground), Shute et al. (2013) were able to create an identification system to determine which simple machine (e.g., pendulum, ramp, springboard, lever) a player had drawn using crayon-like sketchings with an accuracy of >95 % compared to human raters (Shute, Ventura, & Kim, 2013, 2013). Similarly, Gobert and colleagues (Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012) used process data from students' log files from science inquiry simulations and built detectors that reliably identify when a student is demonstrating inquiry skills. In summary, the automated scoring in these new digital assessments can be trained to produce similar ratings to good human scorers while eliminating much of the error they introduce.

Test-Retest Reliability

The traditional notion of test-retest reliability is that if a test is stable, it should yield the same result when given multiple times under similar conditions. Therefore, a common method of assessing reliability is to give the same test on multiple occasions, often weeks apart, and compute the correlation between the scores. Most descriptions of test-retest reliability caution that testing must be done on individuals who are not expected to change on the trait under measurement, since the goal is to see if the assessment produces the same result. However, a common feature of ongoing assessment is that data are gathered as students are learning. The very act of engaging in the experience from which we are gathering data is increasing students' knowledge and skills. Using a digital game as an example, as students play through a game, they are provided with feedback and tasks specifically designed to increase their knowledge and/or skills. The second time they play through the game, we would expect higher scores. To complicate matters further, in many digital environments, the student can have a different experience on a second attempt if the system adapts based on his or her responses or different paths chosen.

Rousson, Gasser, and Seifert (2002) point out that learning effects are not a defect of the assessment, but should be encouraged and then accounted for in procedures. They argue that the product-moment correlation should be used to assess test-retest reliability (as opposed to the intraclass correlation (ICC)). However, this method assumes that the amount of learning across individuals will be constant. That is, those who score high the first time will score high the second time, those who score low will score low again, and there will simply be a shift in means. It is not at all clear that this assumption is warranted. It is likely that there will be variations in the amount of learning that are not directly related to levels of previous knowledge. Weir (2005) suggests that if learning effects are present, trials should be added until a plateau is reached and then the ICC calculated on trials in the plateau region. However, this would likely to create a restriction of range in the analyses.

One solution might be to create a learning parameter that could be added to the equation that would take into account the average learning, along with potential variability by the individual. However, such a procedure does not appear to have yet

been fully developed, requiring further work from researchers if test-retest reliability is to be evaluated for ongoing assessments.

Reliability Coefficients

A common measure of reliability is Cronbach's alpha, a measure based on a combination of internal consistency and test length (Davenport, Davison, Liou, & Love, 2015). Evidence from individual games and other assessment-for-learning systems suggests that alpha ranges are quite reasonable. For example, a game-based measure of persistence showed an alpha coefficient of 0.87 (DiCerbo, 2014), and a pre-Algebra assessment-for-learning system yielded an alpha coefficient of 0.88 (Shute, Hansen, & Almond, 2008).

However, recent work has pointed to flaws in the use of Cronbach's alpha, particularly because of its assumption that all items on an assessment have equal loadings on a single factor and that the errors among the items are uncorrelated (Yang & Green, 2010). It is likely that these assumptions are often violated in educational assessment, and it is well documented that coefficient alpha is negatively biased when the equivalency is violated and inflated when there are correlated errors (Green & Hershberger, 2000).

Factor analytic methods can be employed to address the potential violations of the assumptions of alpha (Green & Yang, 2015; Yang & Green, 2010). In this case, the factors represent the true score and errors represent measurement error. Covariation between the factors and between the errors can be modeled. Reliability can be computed as a function of factor loadings, factor variances, factor covariances, item error variances, and item error covariances (Yang & Green). This approach was used by Kim and Shute (2015) in relation to the measurement of physics understanding in two versions of the game Physics Playground. While the Cronbach's alpha estimates were 0.63 and 0.50, the factor analytic-based estimates were 0.96 and 0.92. Green and Yang suggest that when factor analysis suggests multifactor solutions, omega coefficients indicating the proportion of subscale variance due to the general factor and the proportion of variance in total scores due to all factors should be reported. Given the complex nature of many digital learning environments, it is likely that evidence from these systems will consist of multiple factors with differential loadings and correlated errors, so approaches that allow for these are likely to provide more accurate estimates of reliability.

Generalizability Theory

Generalizability (G) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a unifying framework for various measures of reliability, allowing for the specification of different characteristics of the assessment situation (e.g., raters, items, and so on) called facets, such that the researcher can quantify the amount of error associated with each. G studies use analysis of variance to quantify the

contribution of each error source to the overall error, or lack of precision in scores. Although it emphasizes the different sources of error, G-theory also provides a generalizability coefficient, analogous to a reliability coefficient, which summarizes the precision of the estimates from an assessment. Commonly identified facets include rater and item facets, which can be seen to correspond to interrater reliability and internal consistency.

Brennan (2011) notes that occasion is another facet and notes that it is often ignored. However, he also states that in order to estimate the error associated with an occasion facet, the researcher must be reasonably sure that the examinees' scores have not changed. This of course brings us back to the learning-related issues raised in the previous test-retest reliability section. In addition, the requirements of G studies require fully crossing levels of facets in an analysis of variance (ANOVA) design, for example, crossing items by forms by individuals. Even "unbalanced" designs in which not all items are taken by all individuals, for example, can only take into account one form of unbalance (such as all examinees not taking all forms or all forms not having all combinations of items). Therefore, a situation in which game players are playing different scenarios and those scenarios produce different evidence (i.e., the case in many games) is not currently addressable by a G study.

Test Length

Test length is strongly related to reliability. This makes intuitive sense; the more observations we see of someone doing something, the better idea we will get of how proficient that person is at that thing. Anomalies resulting from a student's experiences on a given day (e.g., mood, hunger, and fatigue) and the environment (e.g., broken air conditioners, fire alarms, and poor lighting) will not have as big an impact on our estimates if we gather pieces of evidence over time. Traditional point-in-time assessments generally contain relatively few items assessing a particular skill. A system of linked ongoing assessments allows us to greatly expand the number of pieces of evidence (or observables) that can inform us about a given knowledge, skill, or attribute. Over the course of a unit, with interactions in games, simulations, and computer tutoring systems, it would be possible to collect hundreds of observations that could provide evidence about a student's understanding and proficiency.

The Spearman-Brown prophecy formula shows us how test length affects reliability:

$$\alpha^{new} = \frac{m\alpha^{old}}{1 + (m-1)\alpha^{old}}$$

The reliability of a new test is a function of the old reliability and the new test length divided by the old test length (m), assuming the added items have the same reliability as the previous items. As an example, if we have a 10-item test with a reliability of 0.70 and we add 10 items for a total of 20 items, the reliability would increase to 0.82. The formula shows us that adding ten items will improve reliability

differentially depending on how many items are already on a test. Assuming the evidence is well aligned to the constructs (as discussed below), this explosion in the amount of information available should increase the precision of our estimates.

Validity

An assessment may be highly reliable, giving very stable and consistent results, but may not actually be measuring the construct of interest accurately. The question of validity is whether the results of an assessment are an accurate measure of the target construct. In other words, given that the bathroom scale tells you the same number every day (it is reliable), is it revealing your actual weight? Without evidence supporting the inferences to be made from a test in a given context, it is unclear how the results can be useful. The combined American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education *Standards for Educational and Psychological Testing* define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (2014, p. 9). In order to demonstrate validity, evidence must be gathered from a variety of sources. The *Standards* state, "The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (p. 9). We can think of work we do to establish validity in two stages: (1) work done during the design phase of assessment and (2) information gathered after the assessment is designed.

Design Phase: Developing the Argument

During design, we build the evidence for what we are (and are not) measuring, the types of tasks that measure those things, the type of information gathered from those tasks, and the combination of that information to tell us about the thing we are measuring. By making each of these things explicit, we can point to this chain to demonstrate how our score relates back to the area we were trying to measure. It lays out the process by which we can argue that we can make inferences about proficiency from scores on an assessment.

One of the advantages of the evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003) framework is that it builds the construction of the validity argument into the design of the assessment. ECD makes explicit the chain of reasoning that goes from defining the domain to representing it, to creating tasks to assess it, to identifying the important features of those tasks, and to developing statistical models to summarize the performance.

The Conceptual Assessment Framework is the component of ECD most focused on formally establishing the assessment argument. It defines the following models:

Student model – what we want to know about the student

Task model – what activities the student will undertake

Evidence model – how we link the work produced in the task to the constructs in the student model consisting of two parts:

Scoring model – how we identify evidence in the students' work product

Measurement model – the statistical techniques we use to link the evidence to the elements in the student model

It is the careful definition of the links between activity, evidence, and constructs that creates the basis for the argument that the inferences made from observing student performance on activities are valid.

Post-Design: Confirming the Argument

Once an assessment is developed, then there is the need to gather evidence to confirm and refine our assumptions from the ECD work in order to provide support of the inferences made from the assessment. The literature on validity today is largely influenced by the writings of Messick (1995) who argued that validity is not a property of a test, but of the interpretation of the test scores. Scores are a function of not just the tasks, but also the person and context. Thus, evidence of validity must be gathered in relation to a context, person, and intended use of test results.

Traditionally, researchers talked about three kinds of validity: construct, content, and criterion. Messick (1995) argued that construct validity was the overarching validity concept, subsuming the other types. In a response to these arguments, the *Standards* (American Psychological Association et al. 2014) enumerated five sources of evidence for validity: (1) evidence based on test content, (2) evidence based on response processes, (3) evidence based on internal structure, (4) evidence based on relations to other variables, and (5) evidence based on the consequences of testing. An impression of construct validity emerges from examination of a variety of empirical results that, together, make a compelling case for the assertion of construct validity for a given measure. To examine the validity of ongoing assessments, we next discuss different types of evidence.

Evidence Type 1: Content Relevance and Representativeness

Evidence of content relevance and representativeness addresses questions about whether the tasks of the assessment are relevant to and adequately cover the breadth of the targeted domain. An assessment measuring a particular domain may under- or overrepresent various elements of the domain, thereby influencing the extent to which the overall result gives an accurate picture of the construct of interest. Assessment designers working in technology-rich environments suggest that the affordances of technology allow for better coverage of the construct of interest (Sireci & Zenisky, 2006). That is, the additional types of interactions available via technology can allow for the assessment of previously unassessed aspects of a construct.

The challenges to representation in a single traditional test versus a system of ongoing assessments are quite different. In a single test, it is relatively easy to map

out a test blueprint and identify coverage. However, there is a significant challenge in being able to sample all of the relevant aspects of a construct in a single setting. As a result, we see standardized testing times creeping up but with limited ability to report on subskills due to their small sampling. A typical end of the year test only contains one to two questions per subskill, not enough to create reliable estimates of subskills. Alternatively, in an ongoing formative assessment environment, evidence is gathered across multiple settings and time points as students interact with learning materials. Therefore, it is far easier to get broad and deep coverage of a construct. The difficulty becomes mapping the evidence across different activities and ensuring that subskills are not over- or underrepresented in estimates of the larger constructs. This mapping however is easily accomplished by expanding Q-matrix techniques (Tatsuoka, 1983) to multiple activities.

The advantage of the ongoing assessment approach becomes clear when looking at the Next Generation Science Standards (NGSS). The standards have three strands: crosscutting concepts, science and engineering practices, and disciplinary core ideas. The National Research Council who developed the standards was clear that these are not intended to be measured by separate tasks but integrated into complex activity (National Research Council, 2012). The difficulty then becomes separating evidence for each of the strands when the same piece of evidence might indicate either a lack of understanding of a core idea or lack of application of a practice, for example. One way to address this is to have students complete many activities with different pairings of concepts, practices, and core ideas. This allows for analysis of commonality across skills. However, completing enough complex activities to implement this across even a portion of the standards would take significant time. Imagining how the spirit of the NGSS could be captured in a one-time summative assessment is difficult.

Another major threat to validity is construct-irrelevant variance, and it seems that digital activities have the potential to introduce significant variance due to issues unrelated to the constructs of interest. Experience playing games, for example, has been shown to be a confounding variable when attempting to make inferences from game performance (Gaydos & Squire, 2011). Even players who are expert in a content area may exhibit poor performance if they are unfamiliar with game play. However, these issues of variance due to familiarity with the context can be addressed in a number of ways. First, tutorials to address game play can be implemented (Sireci & Zenisky, 2006). Second, evidence can be gathered from later stages and levels, while early levels are meant to introduce game play. This strategy was used by the team creating SimCityEDU (DiCerbo et al., 2015). Third, we could potentially try to model individuals' game play familiarity as part of the model. Each of these is theoretically promising, but more research is needed to examine their ultimate ability to reduce construct-irrelevant variance.

Finally, ongoing assessment often does not appear as a "test" to students, and this can influence students' intentionality and goal adoption (Slota, 2014). That is, validation of those assessment activities will require information about how the student perceives the goal of activity (e.g., DiCerbo & Kidwai, 2013) and how the adopted goal influences the way the student interacts with the assessment task, which

can completely change the evidentiary strength and focus of assessment. For example, Kim and Shute (2015) conducted an A/B test where students played two versions of Physics Playground. The two versions were identical except for the way that players proceeded to the next level (i.e., linear vs. nonlinear). Players who played the linear version of the game essentially perceived the goal of the game as unlocking the next level (like many players tend to do with games with linear sequences such as Angry Birds) rather than achieving badges (therefore, higher scores).

Evidence Type 2: Response Process

Questions about response process ask whether examinees used the cognitive processes intended by the activity authors to complete a task. That is, do the cognitive processes used by examinees to complete the task align with what was intended to be measured? On a multiple choice exam, for example, if the correct answer can be deduced from the distracters or previous questions, the participants are using processes other than those being assessed to answer the question.

Although evidence related to response processes is not often addressed in traditional measurement (Cizek, Rosenberg, & Koons, 2008), technology-rich environments like digital games add significant complexity to the assessment environment with unknown implications for cognitive processing and may require special attention to substantive validity issues. They also have the potential for allowing successful completion through brute force or trial-and-error strategies, rather than the use of skills of interest.

The most common way to gather evidence regarding response processes is through think-aloud protocols that allow researchers to observe the techniques students use to solve problems. Hickey, Wolfe, and Kindfield (2000) used this method with an assessment for an online system for genetics learning. They found evidence of students using cues from within the question and from previous questions to answer more difficult items, demonstrating that students got correct answers without requisite knowledge. In a nice turn of phrase, they dubbed this construct-irrelevant easiness. DiCerbo, Frezzo, and Deng (2011) recorded game actions and player thoughts in a digital game targeting computer networking skills. Cognitive processes related to troubleshooting cycles were detectable, and differences in the processes of beginning and advanced students were revealed. However, the puzzles in the game were not observed to be solvable by brute force or trial and error. Rather, players had to use the intended skills to complete the tasks. Similarly, researchers on SimScientists, a system of science simulations, analyzed the results of the implementation of think-aloud protocol with 28 students working through the activities. They found that 84 % of their items elicited the targeted knowledge and practices (Quellmalz, Timms, Silberglitt, & Buckley, 2012).

Taking this a step further, Baker and colleagues (Baker, Corbett, Koedinger, & Wagner, 2004) focused on detecting user behaviors that they categorize as attempts to succeed in an educational task by taking advantage of properties in the system rather than thinking through the material. For example, intelligent tutor systems often have hint systems which provide a series of hints in which the last one is the

answer. A student "gaming the system" might hit the hint button quickly a number of times in order to get the answer without doing the work of the problem. Using machine learning techniques, the researchers developed a method of identifying when students used these strategies. The method transfers across students and specific curricular material. They then developed strategies to intervene when this behavior is detected, including the use of an animated agent who both signals the behavior and provides supplementary exercises covering the material the student skipped (Baker et al., 2006). This work to ensure response processes is possible in the context of digital environments used over an extended period of time in the classroom because of the large amounts of data generated by the ongoing interactions.

Evidence Type 3: Internal Structure

Evidence about the internal structure of an assessment asks if the pieces of an exam all relate to the main construct of interest and to each other. The measurement model component of ECD describes ways to combine evidence from multiple tasks to inform probabilities of latent variables. In most of these models, the unobservable variable is what influences the observed scores. This entails using the measured scores to make inferences about a thing we can't actually see. However, if we believe the values of the observed variables are the result of the underlying construct, the observed variables should be correlated with each other, and we should be able to build models that allow us to estimate the amount of variance in each observed variable that is explained by the latent variable.

DiCerbo (2014) used confirmatory factor analysis to investigate the relationship of observed indicators of player persistence in the game Poptropica. This allowed for the investigation of the fit of the data to a model with one underlying factor and also to examine the factor loadings, which indicate the amount of variance in the indicator that is explained by the underlying factor. Similarly, Shute and Moore (in press) used confirmatory factor analysis to determine the relationship among the observed indicators in Physics Playground relative to physics understanding. Quellmalz et al. (2012) examined the fit metrics of a multidimensional item response theory model to determine whether items fit the intended mapping onto latent proficiency variables, finding acceptable fit for nearly all items. Other than these few examples, evidence of internal consistency is not often reported for ongoing assessment and therefore remains an area in need of more research.

Evidence Type 4: Relations to Other Variables

The degree to which a new assessment is related to other measures of the same construct, and not related to measures of dissimilar constructs, is known as convergent validity and divergent validity, respectively. That is, if our assessment is related to other assessments already known to measure a construct, we can infer that our assessment measures that too. When we look at measures that we expect to be less or unrelated, we should first ask what other constructs our assessment might measure. For example, a test of knowledge might actually be assessing language proficiency.

If we compare scores on the test of knowledge to scores on a test of language skills and find low correlations, this suggests that our test is not in fact a test of language.

The relationship to measures of the same construct is somewhat unclear with assessment based on digital activities. If games and other digital activities are measuring constructs at a deeper level than traditional measures, through the inclusion of process data, for example, very high correlations would not be expected between the digital and traditional measures. However, traditional psychometricians would argue that if correlations are not that high, then the measures are assessing different constructs. In some cases, Cohen's (1992) definition of a large correlation as 0.50 or above is used as a cutoff for whether a test demonstrates convergent validity. In other cases, a significant, nonzero correlation is viewed as acceptable evidence.

In examining reported relationships, Quellmalz et al. (2012) report correlations between 0.57 and 0.64 between scores on their embedded simulation-based assessments and more traditional measures of the science constructs. Shute, Moore, and Wang (2015) found significant correlations involving an estimate of problem-solving skill in Plants versus Zombies 2 with two external measures of problem-solving skill – Raven's Progressive Matrices ($r = 0.40$) and MicroDYN ($r = 0.48$) with just a small sample of middle school students ($n = 0.52$). Delacruz, Chung, and Baker (2010) reported a beta weight of 0.67 predicting traditional posttest scores from game scores in a game targeting pre-algebra skills. No studies using other evidence for convergent validity, including confirmatory factor analysis (including evidence from both measures in the model), were found.

The question under consideration here is not whether a single digital experience correlates to accepted measures of the construct, but whether a series of measures over time does. To date, there have been no known studies that examined correlations between (a) a series of ongoing assessments in multiple contexts over an extended period of time (weeks or months) and (b) a one-time summative assessment at the beginning or end of the experience. There is potential, for example, in the use of dynamic Bayesian networks, as described by Conati, Gertner, and VanLehn (2002) to aggregate evidence across time and activities in a way that produces a final estimate that could be related to other measures of the same construct.

Evidence Type 5: Consequences

Evidence about the consequential validity of an assessment relates to the appropriateness of outcomes that result from the use of an assessment. For example, if a test is used to place students into different classes, examination of consequential validity will look at whether the resulting class placements were appropriate for the students (we might ask the instructors of the classes to rate this). Questions of fairness can also be examined when discussing consequential validity if students of a particular background tend to score in ways such that they have disproportionately different consequences (e.g., overrepresented in classes for the learning disabled).

Interestingly, there was no evidence to be found in the published literature about ongoing assessments being implemented with any consequence, from simple class decisions to high-stakes decisions. Therefore, there is very little to be said about

consequential validity, and this is clearly an area offering potential for further research.

Fairness

Fairness in educational assessment has four meanings: (1) lack of bias, (2) equitable treatment in the testing process, (3) equality in the use of outcomes from testing, and (4) equal opportunities for different subgroups to learn (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Fairness should be addressed with particular care in ongoing assessment because different subgroups of students (e.g., male vs. female, urban vs. suburban) might differently interact with the particular technology.

For example, the literature generally reports that compared with females, males play all genres of games more frequently than females and for a longer duration. Males are also more willing than females to sacrifice other activities to play games (Rideout, Foehr, & Roberts, 2010). Therefore, using games as the vehicle for assessment, without minimizing the influence of the player's gaming ability relative to the accuracy of measuring his or her proficiency in the target skills and knowledge, can be problematic. For example, Kim (2014) investigated the fairness of Physics Playground relative to students' gaming abilities and gender. She found that males who are also gamers may have an unfair advantage over females who are not gamers in terms of obtaining gold badges (i.e., optimized solutions), regardless of physics understanding.

The addition of technology into assessment brings with it concerns about the digital divide. Given that students from homes with lower socioeconomic status have less access to computing devices than those with higher socioeconomic statuses (Purcell, Heaps, Buchanan, & Friedrich 2013), their performance on assessments that rely on those technologies may reflect underlying differences in experience rather than differences in the constructs of interest. The growing use of 1:1 device initiatives in schools may help ameliorate some of these differences, but careful consideration of the technology's potential impact on scores remains important.

Discussion

The previous sections discussed psychometric considerations in the use of ongoing assessments. Table 1 summarizes current evidence regarding the reliability of such assessments. Table 2 summarizes the validity evidence.

There are two different types of work implied by the findings here: (1) modification of conceptualizations and procedures for establishing evidence of reliability, validity, and fairness to incorporate new, ongoing types of assessment and (2) collection of broader evidence to support arguments for their psychometric soundness. Ongoing assessment involves evidence where:

Table 1 Reliability evidence

Reliability measure	Summary of available evidence
Interrater	Automated scoring in digital environments reduces rater error
Test-retest	Methods of estimating test-retest reliability often do not apply because technology-rich environments involve learning in addition to assessment, so second attempts would be expected to change
Internal consistency	Evidence of acceptable values using known criteria suggests using methods allowing for correlation among constructs and among errors
G-theory	Common methods employed in G-theory studies are not applicable to the common situation in new assessments where not all students complete all activities and many pieces of evidence are related to multiple constructs

Table 2 Validity evidence

Validity type	Summary of available evidence
Content relevance and representativeness	Ongoing assessment offers opportunities to increase broad content coverage. There are a number of techniques available to reduce construct-irrelevant variance, but it remains a threat
Response process	Automated detectors of intended as well as unintended response processes have been created and implemented
Internal structure	A few studies report evidence for this type of validity; traditional methods should be applicable
Relation to other variables	Correlations with measures of similar constructs tend to be significant but moderate in size. Consideration should be given to what level of correlation would be acceptable
Consequences	No evidence available

- Each piece of evidence may provide information about multiple constructs.
- Evidence will not be independent of other piece evidences.
- Learning occurs as a result of interaction with the activities.
- Different students will interact with different activities.

Consequently, many of the assumptions of our traditional methods of establishing validity and reliability are violated. In particular, methods of assessing the consistency and stability of scores need to be examined. In addition, acceptable standards for correlation with measures of similar constructs measured via traditional means need to be established. As Linn, Baker, and Dunbar wrote in (1991), "There are, of course, well established psychometric criteria for judging the technical adequacy of measures. Key among these are criteria that stem from the fundamental concepts of reliability and validity, but expanding on their traditional conceptions seems appropriate considering the stated virtues of many new approaches to assessment" (p. 17).

A second line of research involves increased effort to gather a variety of evidences about the validity of these new measures. There is too little evidence regarding the internal structure or consequential validity of ongoing assessments. In addition, we have theoretical techniques for reducing construct-irrelevant

variance, but little quantitative evidence of the success of these methods. Even working within existing definitions, more can be done to establish the psychometric properties of ongoing assessments.

Use Cases and Psychometric Considerations

The various configurations of actors, information, and processes that define an assessment process can be thought of as use cases (Mislevy et al., 2014). Use cases can include providing formative decision support to learners and teachers, providing information to assessment designers, summarizing performance in a unit, or making high-stakes accountability decisions. The assessments described here are specifically for the use case of providing information for and about learning while learning is in process. It may be that the psychometric standards to which assessments are held can vary by use case. It also may be the case that different standards hold differing importance in different use cases. For example, if a strict cutoff on a test is going to be used to make graduation decisions, the error in the test must be very low and the reliability high. If, however, the test is going to be one piece of information a teacher uses, combined with an existing mental model of her students' skills, to make a decision about groups for one lesson on 1 day of class, perhaps the test can be somewhat less reliable. The process by which teachers and others aggregate, weight, and make decisions based on data without the presence of strict rules becomes more important in formative environments. This suggests more attention should be paid to aspects of consequential validity. These differences in emphases and standards for psychometric properties by use case are issues on which consensus will need to be established in the learning and assessment communities. However, this does not mean that we should disregard any of the measures described here.

Implications for Implementation

When making decisions about implementation of any assessment, the intended use of the assessment information is paramount. Mislevy et al. (2014) outlined a number of potential use cases of assessment information, including (a) for students to make decisions about their learning, (b) for teachers to support their instructional decisions, (c) to be used as end-of-course assessment, (d) to be used as accountability assessment for teachers or schools, (e) to be used as high-stakes assessment of students' learning, and (f) to inform assessment designers. It is clear from the evidence above that the current state of these assessments is not appropriate for making high-stakes decisions about students, teachers, or schools. Even such things as placement in programs based on results from many of the assessments described in this chapter would be questionable based on current evidence. However, even at this relatively early stage, it appears that estimates from some of these assessments

could be used to inform daily instructional decisions such as which students to group for a daily lesson or what topics a class needs to revisit during a review session.

Far more instances of well-designed ongoing assessments will be required in order to achieve the vision described herein. Currently there are isolated examples of such assessments/environments assessing a variety of disconnected skills. The vision of an interconnected set of digital activities that produce psychometrically sound evidence will require more careful development using principled design frameworks at a scale that has not yet been achieved.

Conclusions

This review of existing psychometric evidence for ongoing assessment indicates there is work to be done both in reconceptualizing some of our current understandings of evidence for reliability and validity and in the gathering of a broad base of that evidence for these new assessments. As a result, a system of ongoing, performance-based, formative assessment in technology-rich environments remains aspirational. However, it is based on the experiences of educators and students who are moving forward in their embrace of digital technologies and rejection of many of our existing assessment paradigms. Our challenge is to move beyond the understanding of new technology as means to acquire previous ends and to reinvent our conceptualizations to take advantage of a digital-first world.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baker, R. S. J. d, Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students game the system. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 383–390). New York: Association for Computing Machinery.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., Evenson, S. E., Roll, I., Wagner, A. Z., Naim, M., Raspat, J., Baker, D. J., Beck, J. (2006) Adapting to when students game an intelligent tutoring system. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (392–401). New York: Springer.

Brennan, R. L. (2011). *Using generalizability theory to address reliability issues for PARCC assessments: A white paper*. Iowa City, USA: University of Iowa Retrieved from https://www.parcconline.org/sites/parcc/files/gt-PARCC-9-9-11.pdf.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–412.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Conati, C., Gertner, A., & Vanlehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction, 12*(4), 371–417.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined in Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice, 34*(4), 4–9.

Delacruz, G. C., Chung, G. K. W. K., & Baker, E. L. (2010). *Validity evidence for games as assessment environments*. CRESST Report #773. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology and Society, 17*(1), 17–28 Retrieved from: http://www.ifets.info/journals/17_1/3.pdf.

DiCerbo, K. E. & Kidwai, K. (2013). *Detecting player goals from game log files*. Poster presented at the Sixth International Conference on Educational Data Mining, Memphis, TN.

DiCerbo, K. E., Frezzo, D. C., & Deng, T. (2011). Substantive validity of a simulation-based game. *Research and Practice in Technology-Enabled Learning, 6*(3), 161–185 Retrieved from http://apsce.net/RPTEL/RPTEL2011NovIssue-Article2_pp161-185.pdf.

DiCerbo, K. E., Bertling, M., Stephenson, S., Jie, Y., Mislevy, R. J., Bauer, M., & Jackson, T. (2015). The role of exploratory data analysis in the development of game-based assessments. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 319–342). New York: Springer.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In *World conference on educational multimedia, hypermedia and telecommunications* (pp. 939–944). Waynesville, NC: Association for the Advancement of Computing in Education.

Gaydos, M., & Squire, K., (2011). *Validating embedded assessment strategies in game-based learning environments: An expert-novice study.* Paper presented at the American Education Researchers Association, New Orleans, LA.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*(1), 111–143.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*, 251–270.

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice, 34*(4), 14–20.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. (2000). Assessing learning in a technology-supported genetics environment: Evidential and systemic validity issues. *Educational Assessment, 6*, 155–196.

Kamenetz, A. (2015, April 20). Anti-test 'opt-out movement makes a wave in New York State. Retrieved from http://www.npr.org/blogs/ed/2015/04/20/400396254/anti-test-opt-out-movement-makes-a-wave-in-new-york-state

Kim, Y. J. (2014). *Search for the optimal balance among learning, psychometric qualities, and enjoyment in game-based assessment*. Dissertation Thesis, Florida State University, Tallahassee, FL.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education, 87*, 340–356.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Mislevy, R. J., Oranje, A., Bauer, M. I., vonDavier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K. & John, M. (2014). *Psychometric considerations in game-based assessment* [White Paper]. Retrieved from Institute of Play website: http://www.instituteofplay.org/work/projects/glasslab-research/

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

Nelson, H. (2013). *Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time*. New York: American Federation of Teachers.

Purcell, K., Heaps, A., Buchanan, J., & Friedrich, L. (2013). *How teachers are using technology at home and in their classrooms*. Washington, DC: Pew Research Center.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.

Rideout, V. J., Foehr, U. G., & Roberts, D. F. (2010). Generation M2: Media in the lives of 8-to 18-year-olds. Menlo Park, CA: The Henry J. Kaiser Family Foundation. Retrieved from http://kff.org/other/poll-finding/report-generation-m2-media-in-the-lives/

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in Medicine, 21*, 3431–3446.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

Shute, V. J. (2009). Simply assessment. *International Journal of Learning, and Media, 1*(2), 1–11.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Shute, V. J. & Moore, G. R. (in press). Consistency and validity in game-based stealth assessment. To appear in H. Jiao & R. W. Lissitz (Eds.). Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective. Charlotte, NC: Information Age Publisher.

Shute, V. J., & Zapata-Rivera, D. (2010). Intelligent systems. In E. Baker, P. Peterson, & B. McGaw (Eds.), *Third edition of the international encyclopedia of education* (pp. 75–80). Oxford, UK: Elsevier Publishers.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it – or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education, 18*(4), 289–316.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research, 106*, 423–430.

Shute, V. J., Moore, G. R., & Wang, L. (2015). Measuring problem solving skills in Plants vs. Zombies 2. *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*, June 26–29, 2015, Madrid, Spain.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment, 21*(1), 1–27.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 329–348). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Slota, S. T. (2014). *Project TECHNOLOGIA: A game-based approach to understanding situated intentionality* (Doctoral dissertation). Retrieved from http://digitalcommons.uconn.edu/dissertations/638/

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*, 231–240.

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D. P., Way, W. D., & Sweeney, K. (2010). *Automated scoring for the assessment of common core standards*. Princeton, NJ: Educational Testing Service.

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling, 17*, 66–81.

Kristen DiCerbo is the Vice President of Education Research at Pearson. She leads a team focused on the implementation of learning science and formative assessment in curricula and digital tools. Her research program centers on interactive technologies, particularly the use of evidence from learner activity in games and simulations to understand what learners know and can do. She has also engaged with teachers to understand how to best communicate information about student performance to inform instructional decisions. Prior to joining Pearson, Kristen provided research support to the Networking Academies at Cisco and was a school psychologist in a local school district in Arizona. Kristen received her master's degree and Ph.D. in Educational Psychology at Arizona State University

Valerie Shute is the Mack and Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Her general research interests hover around the design, development, and evaluation of advanced systems to support learning – particularly related to twenty-first century competencies. Her current research involves using games with stealth assessment to support learning – of cognitive and noncognitive knowledge, skills, and dispositions. Her research has resulted in numerous grants, journal articles, books, chapters in edited books, and a patent

Yoon Jeon "YJ" Kim is a Research Scientist at the MIT Teaching Systems Lab. Yoon Jeon's research centers on the design and development of learning and assessment in technology-rich environments, particularly video games and simulations. She also has been working closely with teachers, co-designing curricula that incorporate emerging technologies within STEM domains for the purpose of supporting "21st century skills" such as systems thinking and science inquiry