

Advances in the Science of Assessment

Valerie J. Shute^a, Jacqueline P. Leighton^b, Eunice E. Jang^c, and Man-Wai Chu^b



^aFlorida State University; ^bUniversity of Alberta; ^cUniversity of Toronto

ABSTRACT

Designing, developing, and administering assessments has remained fairly unchanged across the past century. However, recent developments in instructional technology, learning science theory, and advances in the design of assessments necessitate a newfound perspective on assessment. The objective of the present article is to review the topic of assessment in depth—past, present, and future. Specifically, we focus on the use of technologically rich learning environments that have spurred advances in student assessment, new methods and procedures from these advances, and consequently the need to consider implementing comprehensive assessment systems that provide rigorous and ubiquitous measurement of the whole student learning experience.

Standardized achievement tests used for high-stake decisions are on the rise. They are being used around the globe to take account of publicly funded primary and secondary educational systems and, in some cases, especially in the United States, to inform debate and decisions about teacher competency standards, school funding, and the forecast of economic growth (Hanushek & Woessmann, 2012; Koretz & Hamilton, 2006; see also the U.S. Department of Education, 2009). In relation to the latter, Hanushek and Woessmann (2012) demonstrated that a country's gross domestic product growth could be predicted from student scores on large-scale standardized achievement tests such as the *Programme for International Student Assessment* (Organization for Economic and Co-operation Development, 2013). Other standardized achievement tests, such as the College Board's SAT and high school exit examinations (e.g., Alberta Education, 2013), are intended to measure a variety of academic outcomes depending on the test, including acquisition of knowledge and skills and college readiness. Whatever the objective, test performance is consequential as it can dictate high school graduation, the college or university in which a student is ultimately able to matriculate, and the subject major the student can expect to acquire during postsecondary study. Therefore, the consequences of student assessment are unmistakable for individuals, their families, and society.

In this article, we consider how achievement testing and the science of assessing learning are progressing with advances in technology. These advances have implications for stakeholders—students, teachers, parents, policymakers, and industry leaders. In the first section of the article, we provide a brief review of the controversies and changing landscape of standardized measures of learning. In the second section, we present a new feature of this landscape, technologically rich environments (TRES; see Bennett, Persky, Weiss, & Jenkins, 2007) and the pedagogical opportunities they afford for learners and teachers. In addition, we discuss the need for innovative assessments to measure and support learning within TRES. In the third section, we present three principled assessment design frameworks and show how they can provide increasingly accurate and sensitive measures of learners' cognitive (e.g., knowledge and skill acquisition) and noncognitive (e.g., dispositional and emotional) attributes, at one

CONTACT Valerie J. Shute  vshute@fsu.edu  Florida State University, 1114 W. Call Street, 3210 Stone Building, Tallahassee, FL 32306-4453.

Color versions of one or more figures in this article can be found online at www.tandfonline.com/heda.

© 2016 Taylor & Francis Group, LLC

point in time and across time. That is, these assessment design frameworks can direct the development of innovative assessments that meet psychometric standards and possess psychological rigor. In the final section, we call for the reconceptualization of current assessments with innovative assessments that capitalize on technological advances.

Section 1—Controversy and promise: the historical landscape for student assessment large-scale standardized measures of achievement

Finding efficient yet accurate methods to assess learning outcomes has been a persistent objective for maintaining the accountability of public educational systems. However, efficient assessment methods have traditionally relied on standardized, selected response test item formats such as multiple-choice tests. Selected response formats were preferred in order to avoid subjectivity associated with human scoring, whereas standardization in testing allowed for comparability of test scores among test takers across contexts. Although such traditional tests still proliferate as an efficient way to evaluate learning and educational systems, they have a controversial history. The controversy is in part because these methods are rooted in intelligence testing.

In the early 20th century, psychologists such as Thorndike, Terman, Yerkes, and Spearman championed the use of group-administered multiple-choice tests to evaluate human intelligence. In 1918, intelligence testing en masse was introduced with the U.S. Army Alpha test composed of multiple-choice questions designed to identify the “mental fitness” of army draftees expecting to be given jobs and serve in World War I (Leahey, 1992). In the United States at this time, the use of group-administered intelligence tests was distinct from the individually administered intelligence tests that Alfred Binet conducted with French children.

From 1917 to 1940, the apparent ease of evaluating human intelligence with group-administered tests opened the door to some questionable ethical practices based on test results (Leahey, 1992). For example, IQ tests during that time were used to screen new immigrants as they entered the United States, and the results were often used to make broad generalizations about entire populations. This, not surprisingly, led some “intelligence experts” to urge Congress to pass stricter immigration restrictions (Kamin, 1995). Nonetheless, the efficiency of the methods used to administer group intelligence tests was considered desirable and transferred to the measurement of learning outcomes, thereby helping to usher in modern-day standardized achievement tests in primary and secondary school, as well as college admission/readiness tests (Zwick, 2004). One case in point is the College Board’s Achievement test, a college admission or readiness test that back in 1900 was delivered in essay format. However, the essay could not be reliably scored (i.e., the scores varied quite a bit from year to year). To counter such concerns, Carl Brigham, a psychometrician who had worked on the U.S. Army Alpha tests, helped the College Board create the multiple-choice SAT in 1926. Following the creation of the multiple-choice SAT, several other major testing products followed, such as the SAT achievement (subject) tests in 1937, machine scored answer sheets in 1939, and the normalization of the SAT in 1941 to ensure that all future SATs would be comparable in content and difficulty (Lawrence, Rigol, Van Essen, & Jackson, 2003).

The SAT especially, but also other large-scale standardized tests (e.g., ACT), have been criticized on a number of fronts (e.g., Sackett, Borneman, & Connelly, 2008; Zwick, 2004). Sackett et al. (2008) detailed several criticisms against large-scale achievement testing as follows: the tests predict badly, they do not measure all the relevant determinants of important criteria related to achievement and learning, they are subject to coaching and do not measure genuine ability and classroom achievement, they are biased against members of racial and ethnic minority groups, and they are subject to motivational differences among students. To be fair, most of these critiques do not take issue with the care and rigor of the technical frameworks underwriting large-scale tests but rather take aim at the out-of-scope significance that large-scale test results have acquired in the hands of inexperienced users.

Since the 1960s, large-scale achievement testing has played a dominant and consequential role in the assessment of student learning outcomes. This role has come with a predictable set of rules and

by-products. For example, the rules have included designing test items manually, imposing rigorous content and statistical checks to ensure psychometric defensibility, administering in paper-and-pencil format, and rank ordering students. The by-products of this focus have included the systematic assessment of students, comparability of learning outcomes across grades, years, and jurisdictions, but in some cases measuring lower level (superficial) skills in reading, math, and science, and providing ineffective forms of feedback to teachers, students, and policymakers.

The science of large-scale achievement testing has grown to such a level of sophistication, imposing what seems like a gigantic “footprint” in the assessment of learning, that it has systemically dwarfed classroom assessments. Classroom assessments—unable to be developed with the technical infrastructure to ensure psychometric defensibility—continue to have significant influence in providing students and teachers with valuable, albeit limited, information about learning outcomes. However, classroom assessments lack the technical rigor to satisfy as the final say on learning outcomes; thus, in many cases, classroom assessments have gained prominence as tools for preparing students to perform on the large-scale tests that really matter. To be sure, both large-scale tests and classrooms assessments provide valuable information. However, at least two conditions (i.e., advances in the science of learning and technology) now exist that indicate that the rules and effects that have come to characterize assessments may be changing, and incentivizing testing specialists and psychometricians to consider significant changes in how assessments—large-scale and classroom—are conceptualized, designed, administered, and interpreted (Mayrath, Clarke-Midura, Robinson, & Schraw, 2012; Shute & Becker, 2010).

Advancements in learning sciences and technology

The two conditions spurring us to rethink and revamp assessments include advances in the learning sciences and technology. First, advances in the learning sciences indicate that acquiring and demonstrating new knowledge and skills occurs within an environment or pedagogical context, which includes (a) learners with specific cognitive and emotional profiles, and (b) tools to promote and evaluate learning (Pellegrino, Chudowsky, & Glaser, 2001). Central to this understanding is recognition that teaching and learning are not strictly cognitive activities but also emotional ones for both teachers and students (Sternberg & Horvath, 1995). The affective attributes of individual learners and teachers must be considered when assessments are developed; administered; and, most important, used to evaluate and communicate feedback to stakeholders. The latter is vital to ensure that formative feedback is viewed as relevant and used by students (Jang & Wagner, 2014; Leighton, Chu, & Seitz, 2013; Shute, 2008).

Second, technology has dramatically changed the environments and processes by which students learn and communicate, teachers instruct, and assessments are designed and administered. Paper-and-pencil tests are slowly becoming a thing of the past as assessments are now increasingly being designed as adaptive and delivered online (e.g., computer adaptive testing, with computer-based testing), employing dynamic and interactive tasks and simulations (e.g., Gierl & Haladyna, 2012). Items for large-scale tests are increasingly created and assembled automatically by sophisticated computer algorithms that can produce not only items in more cost-effective ways but also enough of them to address security concerns (Gierl & Haladyna, 2012).

This wave of innovation, ushered in by advances in the learning sciences and technology, has revolutionized the science of assessment, permitting greater ecological validity and feedback to students related to the breadth and depth of knowledge and skills learned in-situ, including so-called 21st-century skills (e.g., critical thinking, creativity, collaboration, and problem solving). That is, advances in technologies and their integration with assessment systems have allowed for the assessment of multidimensional learner characteristics (cognitive, metacognitive and affective) using authentic digital tasks (e.g., games and simulations).

In the next section, we focus on technologically rich environments and the pedagogical opportunities they afford for learners and teachers. We also discuss the need for innovative assessments to measure learning in technologically rich environments.

Section 2—Technologically rich environments: A current landscape for student assessment

TREs can be broadly defined as any environment or context that involves and encourages concentrated interaction of an individual with technology. For example, the environment could be a classroom with access to SMART boards (i.e., computerized whiteboards), iPads, or online digital games. However, regarding TREs discussed in this article, it is critical to note that these environments do not passively “house” technological devices. Rather, the devices are intended to be interactive and promote pedagogically relevant as well as socially and emotionally meaningful learning situations for students.

TREs can take many different forms, and an exhaustive list of those forms is beyond the scope of this article. However, several online educational sites offer examples of TREs, such as MIT’s Education Arcade (Massachusetts Institute of Technology, 2013), where digital and multiplayer online games (e.g., *The Radix Endeavor* and *Quandary*), simulations (e.g., *Molecular Workbench* and *StarLogo: The Next Generation*), and social networking (e.g., *Ning*, *Think.Com*, *Diigo*, *Panwapa*) are presented and described. Other sites include Michigan State’s Matrix: Centre for Humane Arts, Letters, and Social Sciences (Michigan State University, 2013), *Arcademic Skill Builders* (www.arcademics.com), and *Gamasutra* (UBM Tech, 2014).

In addition to these sites, other examples of TREs currently used to study learning and assessment include (a) *Crystal Island* (see Rowe, Shores, Mott, & Lester, 2011), designed for students in middle school microbiology; (b) *BioWorld* (see Lajoie, Lavigne, Guerrero, & Munsie, 2001), designed for learners to acquire and demonstrate knowledge about diseases through solving specific patient cases; (c) *Physics Playground* (see Shute & Ventura, 2013; Shute, Ventura, & Kim, 2013), designed to assess and support students’ conceptual physics understanding, creativity, and persistence; (d) the *Digital Deteriorating Patient Activity* (Blanchard, Wiseman, Naismith, & Lajoie, 2012), designed as a real-life educational simulation that prepares medical students to effectively approach emergency situations through role-play; and (e) *Metatutor* (Azevedo, Johnson, Chauncey, & Burkett, 2010), designed to examine the effectiveness of several human scaffolding conditions in facilitating undergraduate students’ learning about the circulatory system.

A defining feature shared by many TREs currently used to study learning is their stimulating problem-solving richness. This richness is achieved by the visual, auditory, and interactive realism of the TRE, permitting learners to engage actively with the learning environment to solve problems by accessing a variety of resources in real time. Depending on the quality of the technology, TREs elevate or transform an otherwise static learning opportunity (e.g., lectures in a classroom) into a dynamic occasion approximating the vividness of learning in vivo as if the student were “really there” solving the task. For example, [Figure 1](#) illustrates two screenshots of the commercial, off-the-shelf game *Civilization III* (Take-Two Interactive Software, 2010), considered to be a high-quality educational game for teaching history (the latest version is *Civilization IV*).



Figure 1. Two screenshots of *Civilization III* (Take-Two Interactive Software, 2010).

Designed for the PC and considered one of the most complex games for history and social studies instruction, *Civilization III* has its players create a society that comes to power through a variety of strategies, including diplomacy, cultural impact, and military strength. By playing the game, which can span the full range of human history, students learn about political alliances, trade, diplomatic negotiations, resource management, geography, and technology.

Facer (2003) indicated that the vitality of learning experiences evoked by TREs is deliberately expected to encourage and tap specific cognitive and emotional regulation skills, such as active engagement in the learning process, rapid information processing, discrimination of relevant from irrelevant variables, and parallel processing from different sources. In addition, TREs enable learners to experience nonlinear methods of exploring information, access to a wide array of information (e.g., imagery and text), communication networks that are not geographically bounded, and the awareness that playing is a form of problem solving (see also Jenkins, Purushotma, Weigel, Clinton, & Robinson, 2009). For example, in *Crystal Island* (e.g., Rowe et al., 2011), middle school students are challenged to understand an unexplained illness that has afflicted a team of researchers on an island. From a first-person perspective, learners must explore the camp, gather data about patient symptoms and associated diseases, formulate hypotheses about transmission, employ virtual lab equipment and resources (e.g., online texts about microbiology concepts), and report conclusions. There are multiple problem-solving paths learners can take in *Crystal Island*, and the challenge is for students not only to solve a mystifying problem but also to make key decisions about how their solutions are going to unfold. According to Facer, successful digital games are expected to produce for students what Csikszentmihalyi (1990) described as *flow*, that is, the full and energetic absorption within an activity due to its strong but age-appropriate set of challenges and intrinsically motivating objectives.

In immersive TREs, involving digital games, online tasks, and/or simulations, learners are situated in electronically enhanced 2D or 3D contexts. One of the objectives of these digitally enhanced contexts is to create for learners the experience or sensation of authenticity as they solve challenging tasks. However, TRE experiences are not arbitrary. This sensation is expected to induce in learners not only a stronger investment of attention, interest, and flow but also a more accurate set of conditions than is normally possible in static classroom-based environments for inspiring learning and measuring its outputs, including the cognitive, emotional, and kinaesthetic processes educators wish to probe in their learners. As Klopfer, Osterweil, Groff, and Haas (2009) of the MIT Education Arcade have emphasized, the potential of digital games, as an example of TREs, for teaching core learning skills is high.

Given that playing through such immersive games like *Civilization* and *Crystal Island* requires a substantial investment of time, we can imagine a type of flipped classroom scenario where kids are assigned the game as homework (or “homeplay”) across several weeks, and then the class can enjoy lively discussions of successes, failures, and strategies moderated by the teacher. We revisit this idea in Section 4 of this article.

TREs and 21st-century skills

One of the promises of using TREs—in addition to providing enhanced contexts to support learning and the assessment of core knowledge and skills in reading, math, and science—is the opportunity they afford for acquiring and demonstrating 21st-century skills (e.g., higher order thinking and problem-solving skills across domains, and learning-to-learn skills). For example, TREs provide opportunities for students to engage in “on-the-fly” complex problem solving, which often includes searching for information, discriminating between distinct data sources, planning strategies, coordinating and collaborating with others, hypothesizing about consequences to courses of action, testing ideas, receiving feedback directly, synthesizing multiple informational streams, modifying and/or revising and re-executing strategies, patience, perseverance, cognitive flexibility, creativity, and cooperation. However, we are not yet at a point where TREs are being widely used in schools, and students are not acquiring these valuable skills. For example, a 2004 analysis paper by the American Diploma Project, a U.S.-based advocacy group for college and career readiness, conveyed that approximately 60% of U.S.

employers have serious doubts not only about students acquiring 21st-century skills but also about the education students are generally receiving in core subject areas such as reading, writing, and math. Remedial training in core subjects costs states as much as \$40 million a year.

According to the Partnership for 21st Century Skills (2008), a coalition initiated in 2002 between the U.S. Department of Education and eight founding organizations (including major industries such as Apple Computer Inc., Cisco Systems Inc., Dell Computer Corporation, and Microsoft Corporation), 21st-century skills are a necessity in today's information-rich and technology-intense economy. These skills specifically include (a) learning and innovation requiring creativity (e.g., brainstorming, elaboration of ideas, openness to diverse perspectives, viewing failure as part of a cyclical process of learning and opportunity to learn), critical thinking (e.g., analysis of how parts of a whole interact, evaluation of evidence, arguments), problem solving (e.g., asking clarification questions that lead to better solutions), communication and collaboration (e.g., listening effectively to fully understand intended meanings); (b) information, media, and technology literacies (e.g., managing the flow of information from a variety of different sources); (c) life and career know-how (e.g., incorporating feedback effectively, adapting to various roles and responsibilities, working well in ambiguous, ill-defined environments); and (d) reading, writing, science, and mathematics.

In cooperation with the Partnership for 21st-Century Skills, Duncan (U.S. Department of Education, 2009) identified four key objectives for remedying the poor preparation students receive for entering the job market. The objectives included (a) setting clear educational goals and having core standards for schools to work toward; (b) facilitating federal resources, such as Race to the Top funding, to support state proposals seeking innovative local solutions; (c) evaluating educational programs that work for raising academic performance; and (d) building on and sharing successful programs across jurisdictions. In response to the challenge, but independently of the federal government, the National Governors Association Center for Best Practices and the Council of Chief State School Officers in 2010 brought forward the Common Core State Standards in collaboration with teachers, researchers, and content experts. With the goal of aligning curricula across the states and specifying student outcomes, state governments decided, separately, whether to adopt the standards to guide educational process and procedures.

Advances for the assessment of 21st-century skills in TREs

The value placed on teaching 21st-century skills is accompanied by increased awareness and effort to properly assess these skills. Arne Duncan highlighted this need in his 2009 speech to the Chamber of Commerce's Education and Workforce Summit:

We have to get a lot smarter about how we evaluate our students—and how we measure success. Instead of setting arbitrary proficiency levels we need to look at growth—and the science of measuring growth has to continue to evolve. We need to invest in the science of testing and measurement and find ways to do it better—without simply doing more of what we are currently doing. None of us like the overemphasis on testing, so we have to find a practical way to measure progress. . . . The testing industry must reform and develop college-ready assessments in an array of subjects. States must invest in new data systems. Local districts must retrain teachers and administrators. This won't happen overnight, but it must happen over time. (Duncan, 2009)

Responsive, comprehensive, and balanced assessment procedures, which measure higher order thinking skills and learning-to-learn skills in students, are an essential aspect of effective teaching in the classroom today (Brookhart, 2011). However, most teachers in traditional classroom environments do not routinely employ such assessments, relying instead on past classroom assessment practices that provide a limited view of student learning. Therefore, it is not surprising that the question of how to properly administer responsive, comprehensive, and balanced assessments within TREs is a new and open question.

To address the need for new types of assessment procedures in K-12 classrooms, two U.S.-based federally funded assessment consortia have been organized with the goal to develop innovative or

“next-generation” assessments in English and math by 2014–15 in accordance with the Common Core State Standards. The first is the Partnership for Assessment of Readiness for College and Careers (PARCC), an association of 18 states plus the District of Columbia and the U.S. Virgin Islands. The second is the Smarter Balanced Assessment Consortium, an association of 25 states. Both share the same policy goals for summative standardized testing in conjunction with classroom-based formative assessment. They differ somewhat in their foci and methods. For instance, in relation to the assessment of English literacy, PARCC outlines advances in the use of “authentic” texts and a sequence of questions that draw students deeper into the text. Further, in relation to the assessment of mathematics, PARCC describes the use of multistep conceptual and application problems that require students to delve deeper into a given domain instead of covering many concepts superficially. Both consortia are devising formative and summative computer-based assessments that move away from traditional paper-based, multiple-choice, and constructed-response items and make greater use of technology-enhanced items, reflecting a variety of question types and features (e.g., drag-and-drop, multiple select, text highlighting, and equation builder), response options, and interactivity. For instance, PARCC’s computer-based through-course (nonsummative) assessments along with end-of-year assessments include constructed response items and performance-based tasks, some of which are computer enhanced, and involve automated scoring and human scoring. Moreover, Smarter Balanced Assessment has designed computer-based online adaptive tests that tailor a variety of item types to student ability levels. Both PARCC and Smarter Balanced Assessment are grappling and making inroads with how to administer and measure student skills using primarily computer-based assessments.

Computer-based assessments reflect a relatively modest deployment of technology for testing purposes at the present time. Many computer-based assessments, including adaptive tests, capitalize on innovations along key dimensions: item format, response action, media inclusion, level of interactivity, scoring, and communication of test results (Parshall, Spray, Kalohn, & Davey, 2002). Innovations along these key dimensions might suggest improvements in the measurement of higher order thinking skills. However, the empirical evidence for this in terms of stronger validity arguments for higher order skills, reliability of ability estimates, and efficiency has been found to be surprisingly limited (Jodoin, 2003; Sireci & Zenisky, 2006; Wan & Henly, 2012). For example, the construct equivalence of essay/short-answer items, similar to what one would expect with innovative item types, and traditional multiple-choice items for subjects such as mathematics, science, and computer science (e.g., Bennett, Rock, & Wang, 1991; DeMars, 1998) suggests equivalence. However, Wan and Henly (2012) argued that this finding could be attributed to the unspecified, lower level of cognition invoked by both items and/or the similarity of item design. In other words, validating constructed-response item types (e.g., essay questions) by comparing them to traditional selected-response item types may not be the appropriate route for making a strong case for improved measurement (Linn, Baker, & Dunbar, 1991).

Wan and Henly (2012) further pointed out that although testing specialists are discussing and creating prototypes for innovative item types (especially in terms of improving construct representation of the knowledge and skills being measured; see Bennett, Morley, & Quardt, 2000; Davey, Godwin, & Mittelholtz, 1997; Luecht, 2001; Parshall et al., 2002; Shute, 2007; Sireci & Zenisky, 2006), empirical studies of the psychometric and cognitive demands of innovative items are few and limited. According to Wan and Henly, these studies tend to be informal and small scale without the benefit of large-scale operational response data. However Jodoin (2003) provided an exception as he compared the precision of measurement (reliability) associated with two types of innovative formats (i.e., a drag-and-connect and create-a-tree) for the Microsoft Certified Systems Engineer Certification Program with traditional multiple-choice. Using large-scale operational data from actual examinees, he found that the innovative item types provided more information about students’ knowledge and skills but less information per time unit. In addition, in an investigation of the reliability, efficiency, and construct validity of two innovative item types involving figural response and constructed response in a K-12 computerized science assessment, Wan and Henly’s results were similar to Jodoin’s. Specifically, Wan and Henly found that the constructed-response items provided more information overall but less information per minute of testing time than the traditional selected-response item types; however, the reliability and

efficiency of the information yielded by of the figural response items was equivalent to multiple-choice items. Clearly more research of this type is needed in relation to innovative assessments.

Technologically enhanced item types

Many item or task formats have been offered to capitalize on advances in technology and the learning sciences. Scalise and Gifford (2006) described a taxonomy of 28 item types for e-learning or computer-based assessment. As shown in Figure 2, the taxonomy categorizes item types along two dimensions—constraint and complexity. The constraint dimension identifies seven levels: (a) multiple-choice (being the most constrained, with the least amount of choice for novel responses), (b) selection/identification, (c) reordering/rearrangement, (d) substitution/correction, (e) completion, (f) construction, and (g) presentation/portfolio (being the least constrained, with the most amount of choice for novel responses). Further, the complexity dimension shows that items within each level of constraint can be designed to be more or less elaborate up to four levels, although Scalise and Gifford do not specify the nature of the elaboration, that is, whether the elaboration is cognitive or technological. For example, completion item types, which are categorized at an intermediate level of constraint (Level 5 in Figure 2), can be designed to be less cognitively complex as exemplified by filling in a single numerical response (most basic shown in cell 5A) or more complex by requiring a full matrix of values to be completed (most elaborate shown in cell 5D). This variation in cognitive complexity might also have implications for the technology required to underwrite the different items.

Although Scalise and Gifford's (2006) taxonomy is useful for categorizing item types, the taxonomy is relatively traditionalist. Most of the item types shown in Figure 2 are conventional in format, albeit with allowances for computer delivery or "technological enhancement." For example, the most elaborate multiple-choice item type is the multiple-choice with new media distractors (see 1D in Figure 2) or the construction of an essay with automated editing (see 6D). Three added dimensions that could strengthen this taxonomy are as follows. First, expectancies about the cognitive demands exerted on learners from different item types could be outlined with relevant research or identified for further research. Research shows discrepancies between expected task difficulty and actual cognitive demands that learners experience (Cohen & Upton, 2007). For example, when summarizing tasks, anticipated to







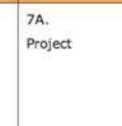

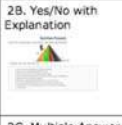


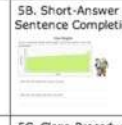
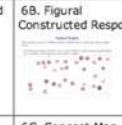
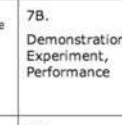
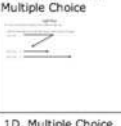





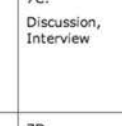






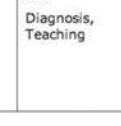
1. Multiple Choice	2. Selection/ Identification	3. Reordering/ Rearrangement	4. Substitution/ Correction	5. Completion	6. Construction	7. Presentation
1A. True/False 	2A. Multiple True/False 	3A. Matching 	4A. Interlinear 	5A. Single Numerical Constructed 	6A. Open-Ended Multiple Choice 	7A. Project 
1B. Alternate Choice 	2B. Yes/No with Explanation 	3B. Categorizing 	4B. Sore-Finger 	5B. Short-Answer and Sentence Completion 	6B. Figural Constructed Response 	7B. Demonstration, Experiment, Performance 
1C. Conventional Multiple Choice 	2C. Multiple Answer 	3C. Ranking and Sequencing 	4C. Limited Figural Drawing 	5C. Cloze-Procedure 	6C. Concept Map 	7C. Discussion, Interview 
1D. Multiple Choice with New Media Distractors 	2D. Complex Multiple Choice 	3D. Assembling Proof 	4D. Bug/Fault Correction 	5D. Matrix Completion 	6D. Essay and Automated Editing 	7D. Diagnosis, Teaching 

Figure 2. Scalise and Gifford's (2006) taxonomy of 28 item types for computer-based assessment.

be cognitively complex, appear at the end of a reading comprehension testlet, their difficulty levels are often lower than expected because learners acquired textual information from solving preceding tasks (Jang, 2009a). Leighton and Gierl (2007a, 2007b, 2011; see also Jang, 2009b) describe ways of investigating the cognitive-processing demands of assessment tasks to generate stronger validity arguments about student learning and achievement (see Kane, 2006, for validity arguments) and provide clearer diagnostic information about student knowledge and skill acquisition for remediation. Empirical evidence from learners' verbal descriptions of their thinking processes can help build greater understanding between how students think (cognition), how tasks elicit cognition (observation), and what inferences are made about students' ability (interpretation; National Research Council, 2001; see also Jang, 2014; Leighton, 2004).

Second, expectancies about the affective demands elicited by item types could be outlined with relevant research or identified for further research. At this time, we are not aware of any published scholarly research investigating the affective demands of the item types shown in Figure 2. Similarly to cognitive-processing demands, identifying the affective states of learners while they are engaged in assessment tasks can help identify sources of difficulties for learners and thus facilitate opportunities for formative feedback and intervention. For example, Conati (2002; see also Conati & Maclaren, 2009) investigated the design of an intelligent pedagogical agent for the game Prime Climb. This agent was tasked with gathering learner affective information—from sensors measuring learners' eyebrow position, skin conductance, and heart rate—during gameplay. The agent extended helpful interventions when the student made a mistake or in some way indicated a need for help. For example, if a student made a cognitive mistake, and displayed a high heart rate and furrowed eyebrows, the agent combined the affective and cognitive information and intervened before the student experienced complete frustration and lack of engagement (Conati & Maclaren, 2009). However, the agent reserved help if the student did not provide affective information indicating the need for an intervention.

Third, item types could be expanded to show methods of implementation. The presentation item type (Level 7 on the constraint dimension), especially the “demonstration, experiment, or performance” type (cell 7B), could be linked to a range of administration possibilities for embedding this item type within true-to-life learning environments such as TREs. Toward this end, DiCerbo and Behrens (2012a, 2012b) identified four different levels of technology-enhanced assessments. As shown in Figure 3, *Level 1—Computerized paper-and-pencil tests* represents the most basic integration of assessment within a TRE. Level 1 includes traditional types of items, often multiple-choice, administered to students via computer. According to DiCerbo and Behrens (2012a), “This level is characterized by the use of

	Level 1	Level 2	Level 3	Level 4
General description	Computerized paper and pencil	Integrated context and tasks	Natural digital activity with assessment features	Accumulation of information from variety of natural digital activities
Tasks	Discrete items	Complex tasks/problems	Complex tasks/problems	Variety of tasks/problems
Integration with Learning Activity	Low	Low	Moderate	High
Variety of Activity/Response	Low	Moderate	Moderate	High
Examples	Multiple choice end of chapter test	Simulation-based assessment	Stand-alone digital game	Dashboard of visualizations of student performance across tasks

Figure 3. Summary of DiCerbo and Behrens's (2012a) four different levels of technology-enhanced assessments.

relatively large numbers of discrete, independent items administered and scored digitally. The technology enhancement may include one or more of the following: innovative item types, automated scoring, adaptivity, and/or advanced feedback” (p. 6). *Level 2—Integrated contexts and tasks* offers students an alternative to discrete, traditional multiple-choice items by presenting technology-enhanced simulated performance-based tasks that require multiple opportunities for complex interactivity for assessment purposes. At Level 2, DiCerbo and Behrens (2012a) indicates, “The technology . . . allows for the observation of discrete actions taken by examinees within these tasks” (p. 8). It is important to note that at this level, as in Level 1, there continues to be a separation between instruction and assessment activities; the activities that define instruction are distinct from the activities that define assessment. *Level 3—Natural digital activity with assessment features* transitions from the previous levels by blurring the line between instruction and assessment. At Level 3, instruction and assessment are part of the same set of TRE activities such as simulation-based games (e.g., Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2008), and assessment is no longer presented as something that is done after the instruction is delivered. *Level 4—Accumulation of information from a variety of natural digital activities* is similar to Level 3 except that assessment is no longer confined, for example, to a single game or set of activities but is instead distributed across a variety of TREs, tasks, and experiences. In this way, an *assessment ecosystem* (DiCerbo & Behrens, 2012a) is created where assessment data are constantly being collected and analyzed as students engage with multiple types of TRE activities such as games, simulations, and digitally enhanced experiences (e.g., tablets in classrooms).

Benefits and limitations of assessments at levels 1–4

The assessments described in Levels 1 through 4 in [Figure 3](#) fall under the general class of e-assessments. Traditionally, e-assessments use computer and information technology to make the assessment process more efficient by automating functions that would otherwise require human assessors (Baker & O’Neil, 1995; Shute & Kim, 2012). The Joint Information Systems Committee (2007) defined e-assessments as

covering a range of activities in which digital technologies are used in assessment. Such activities include the designing and delivery of assessments, marking (by computers, or humans assisted by scanners and online tools), and all processes of reporting, storing, and transferring of data associated with public and internal assessments. (p. 6)

One of the advantages of e-assessments is that most, if not all, can be used *concurrently* (i.e., on demand and instantaneously), and many of them are *embedded* within a learning activity or environment. Some e-assessments are even *transformative* insofar as they not only measure traditional knowledge and skills but also provide a means to measure unique higher order thinking and learning-to-learn skills that push beyond the limits of what is traditionally measured (Binkley et al., 2010, 2012). Referring to [Figure 3](#), assessments at Levels 1 and 2 offer concurrent access, but only assessments at Levels 3 and 4 offer students, increasingly, the opportunities for embedded and transformative evaluation experiences given task complexity, integration with learning activities, and variety of activities and responses.

One of the benefits of assessments at Levels 3 and 4 is the extensive data collected from individuals so that a fine-grained and cohesive portrait can be developed about a particular student’s knowledge, skills, affective states, and so on. When information is collected from such a wide variety of digital activities (including technology-enhanced labs and e-portfolios) to form an assessment ecosystem for individual students, inferences about learner strengths and weaknesses can be better substantiated. However, this accumulation of data creates a practical problem of how to manage such a deluge of data, especially for assessments at Level 4. For example, methods must be developed for how to systematically mine the data, statistically summarize it, and logically present it to support inferences about student learning and, of importance, to be understood and utilized by stakeholders. We discuss this issue and additional obstacles, as well as some potential ways to surmount them, in Section 4.

Related to the extensive data collected, assessments at Levels 3 and 4 are increasingly personalized to help students better understand their areas of strength and areas for improvement. In this respect,

assessments at Levels 3 and 4 capitalize on one of the central goals of cognitive diagnostic assessments (CDA; Leighton & Gierl, 2007a). Although CDAs were originally conceived as assessments at Levels 1 and 2, CDA items were designed to measure, based on theories and models of cognition, specific knowledge structures and processing skills in students to provide information about their cognitive strengths and areas for improvement (Leighton & Gierl, 2007a). A central goal of CDA is to provide detailed formative feedback (see Shute, 2008) to students and teachers about areas of student strength and weakness in order to inform detailed and personalized instruction. Similarly, Jang (2009a), using a method of diagnostic skill classification called *reparameterized unified model*, has found a way to create fine-grained profiles of second-language learning skills for approximately 2,500 students based on their LanguEdge assessment performance. The LanguEdge is a large-scale assessment designed for second-language learners who aim to take the TOEFL. As part of the LanguEdge courseware (which provides support to instructors), the assessment is designed to evaluate student reading comprehension, such as searching, analyzing, and using contextual clues for deducing the meaning of a word; using lexical and grammatical cohesion devices to comprehend relations between sections of text; and distinguishing major and minor ideas in text by analyzing and evaluating the importance of information. Jang (2009a) found that classifications of students into categories of mastery were promising with reliabilities ranging from 88% to 90% depending on the skill category. This work illustrates the potential of personalized assessments to deliver much more information than a single overall grade on an exam to students and teachers.

Another way of personalizing assessments for students is to embed them into ongoing learning activities, thus enhancing the connectedness between learning and assessment. By embedding assessments into TREs, such as stand-alone digital games, the potential of the assessment to distract learners from the object of measurement may be minimized. For example, one type of Level 3 assessment achieves this goal by means of embedding *stealth assessments* (Shute, 2011) within a gaming environment such as *Physics Playground* for assessing and supporting conceptual physics understanding, persistence, and creativity. According to Shute and Ke (2012),

Stealth assessment refers to evidence-based assessments that are woven directly and invisibly into the fabric of the gaming environment. During gameplay, students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess. (p. 52)

In addition to providing valid measures, two other goals of stealth assessment are to stimulate engagement and sustain flow (Csikszentmihalyi, 1990) of learning. Stealth assessment is designed to emphasize the view for learners that assessment is a part of the natural sequence of interactions with the game or learning environment—exploring, observing, manipulating, testing, evaluating, synthesizing, and revising. Given the dynamic nature of stealth assessment, it is not surprising that it promises advantages—for example, measuring learner competencies continually, adjusting task difficulty or challenge in light of learner performance, and providing ongoing feedback. Examples of stealth assessment prototypes, designed to measure a range of knowledge and skills—from systems thinking to creative problem solving to causal reasoning—can be found in relation to the following games: *Taiga Park* (Shute, Masduki, & Donmez, 2010), *Oblivion* (Shute, Ventura, Bauer, & Zapata-Rivera, 2009), and *World of Goo* (Shute & Kim, 2011), respectively. In the game *Physics Playground* (formerly *Newton's Playground*; see Shute & Ventura, 2013), three stealth assessments were created and evaluated in relation to the validity and reliability of the assessments, student learning, and student enjoyment (see Shute et al., 2013). The stealth assessments correlated with associated external validated measures for construct validity and demonstrated reliabilities around .85 (i.e., using intraclass correlations among the in-game measures such as number of gold trophies received for various objects created). Furthermore, students (167 middle school students) significantly improved on an external physics test (administered before and after gameplay) despite no instruction in the game. Students also enjoyed playing the game (reporting a mean of 4 on a 5-point scale ranging 1 [*strongly dislike*] to 5 [*strongly like*]), and boys and girls equally enjoyed the game.

In the next section, assessment design principles are presented to guide sensible and defensible inferences about student learning based on technologically enhanced task types.

Section 3—Designs for principled student assessment

To draw reliable and valid inferences about student knowledge, skills, and other attributes, assessments must be developed or created using principled assessment design (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999; Kane, 2006). For example, Shute and her colleagues developed stealth assessment using Evidence-Centered Design (ECD; Mislevy, Almond, & Lukas, 2003), a principled assessment design framework that supports making explicit links between student skills of interest, behavioral evidence for the presence of those skills, and tasks that probe for the evidence needed. A poorly designed assessment can be manifest in an outdated paper-and-pencil test, but it can also be apparent in a technologically enhanced gaming environment. Caution must be exercised when making claims about student learning or achievement when the assessments undergirding learning activities are developed arbitrarily. A good assessment, regardless of format, needs validity arguments backed by empirical evidence.

In any activity used to assess students (e.g., traditional paper-and-pencil multiple-choice tests, or technologically enhanced stealth assessments delivered via computer), a distinction can be made between the external *expression* of the assessment and the internal *underlying framework* or rules employed to create the assessment. An assessment activity might appear on the surface to be an appropriate measure of student learning if the content of the activity matches the content of the learning domain. However, content can be deceptive. Content can be a serious distraction when evaluating assessments as appropriate measures of learning, especially if the real object of measurement is not knowledge of content per se, but how knowledge of the content is applied, manipulated, or processed.

When higher level skill sets are the real objects of measurement, it is necessary to evaluate assessment activities not by their surface similarities with learning domains but by their deep structural correspondences with intended learning outcomes; for example, does an activity measure the application of a skill or set of skills? To ensure that assessment activities yield useful data for making inferences about student learning beyond simple knowledge claims, principled assessment design must guide the development and structure of the assessment.

Principled assessment design can be viewed as a plan, comprising a visual or textual scheme, to guide the purpose, expression, development, internal structure, and defensibility of an assessment. For example, principled assessment design includes the sources of evidence needed to demonstrate that an assessment measures the knowledge and skills of interest in students, thus supporting and defending test-based inferences about student learning to stakeholders. It is important to note that principled assessment designs encompass more than test blueprints or specifications. Test blueprints provide “a detailed description for a test . . . and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices” (AERA, APA, & NCME, 1999, p. 183). Although a test blueprint is a central part of assessment design, it is not the only part, as it does not include the sources of evidence that need to be collected to defend the particular knowledge and skills tested, format of test items, and interpretation of test item performance.

We have identified three leading principled assessment designs that provide the most advantageous design systems by which to develop technologically enhanced assessment tasks: ECD, Cognitive Design System (CDS), and Assessment Engineering (AE). It is important to note, however, that these design systems do not differ substantially in terms of end goals but they do have differential foci, albeit nuanced, for arriving at those goals. Generally, ECD focuses on the empirical or evidentiary link between test-based interpretations about examinees and the evidence to support those claims, CDS focuses on the cognitive theoretical basis for test item design and interpretation, and AE focuses on the systematic translation between cognitive-based item features into item shells or “models.” In describing

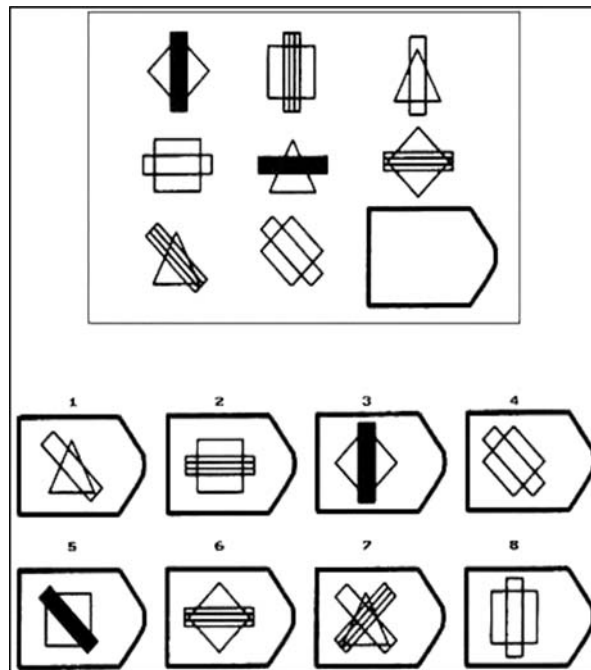


Figure 4. Example item from Raven's Progressive Matrices test. Reprinted with permission from Carpenter, P., Just, M., & Shell, P. (1990, July). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404–431. © American Psychological Association.

each assessment design, we use Raven's Advanced Progressive Matrices (APM) Test (Raven, Raven, & Court, 1998) as a running example to illustrate the differences among designs. The APM Test is a nonverbal 48-item multiple-choice test designed to measure reasoning ability by asking examinees to identify the missing element that completes a figural pattern. An example item is shown in Figure 4.

Evidence-centered design

Much has been written about and in relation to ECD. We therefore do not provide a comprehensive review and instead refer the reader to Mislevy, Steinberg, and Almond (2003) for a full development of concepts. As a summary, ECD reflects a set of procedures, premised on the creation of evidentiary arguments, for generating many kinds of assessments, from classroom assessments to technologically enhanced items and tests. ECD is based on addressing a series of questions posed by Messick (1994), such as the following, to structure and operationalize planning, designing, and making test-score inferences:

1. What knowledge, skills, and attributes (KSAs) do we want to assess?
2. What are the observable features of the KSAs that will allow measurement?
3. What criteria and rubrics can be developed to score KSAs?
4. What kinds of tasks elicit or probe observable features of KSAs?
5. What task specifications guide assessment assembly and administration?

Answers to these basic questions are found using ECD's Conceptual Assessment Framework (CAF; see Figure 5) and a Four-Process Architecture for assessment delivery systems. The CAF illustrates five models. All five models are relevant to answering the questions posed, but three of the five models are fundamental for defending the empirical basis of test-score inferences.

The first is the *student model*, which explicitly specifies the KSAs to be measured by the assessment so as to facilitate the operationalization of the construct with observable behaviors. The second model

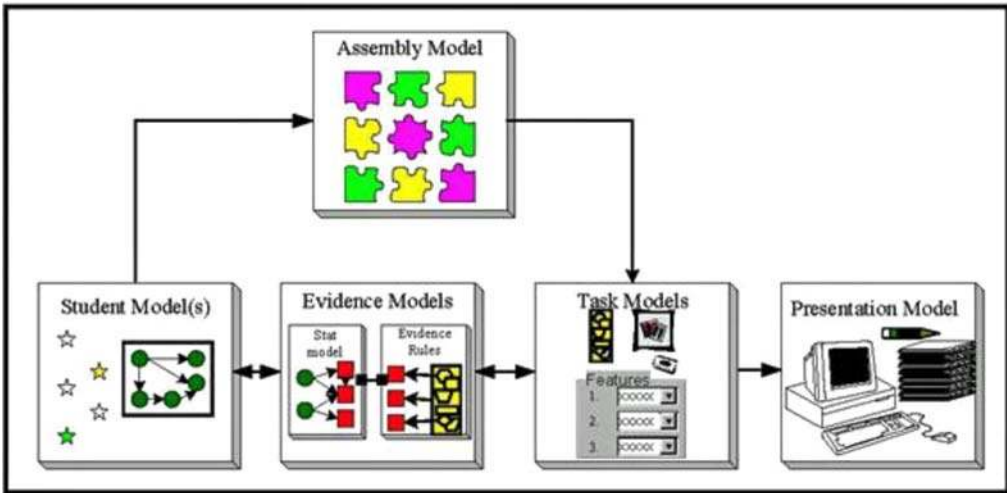


Figure 5. The Conceptual Assessment Framework, from Mislevy, Steinberg, and Almond (2003).

is the *evidence model*, which specifies the assignment of scores to student test performance such as whether dichotomous (i.e., an item response is assigned a value of 1 if correct, otherwise a 0) or polytomous (i.e., an item response is assigned values other than just 0 or 1 to show increasing performance quality) scoring will be used and how the scores will be aggregated. Finally, the third model is the *task model*, which outlines the types of items or tasks, including all features, requiring development to elicit the KSAs of interest from the student.

The CAF feeds into a Four-Process Architecture for organizing and implementing the delivery of assessments. The Four-Process Architecture involves four key decisions (see Figure 6). First, a decision must be made as to how tasks will be selected for the student (*Activity Selection Process*), for example, whether all tasks are presented or, alternatively, selected based on identified and scored (*Evidence Identification Process*), for example, whether overall item performance will be scored as 1 or 0 or at a finer level of granularity, with distinct scores at different points within the task. Fourth, a decision must be made about how a student’s performance, once identified, will be aggregated and summarized (*Evidence Accumulation Process*), for example, using classical test theory, item response theory, or Bayesian Networks.

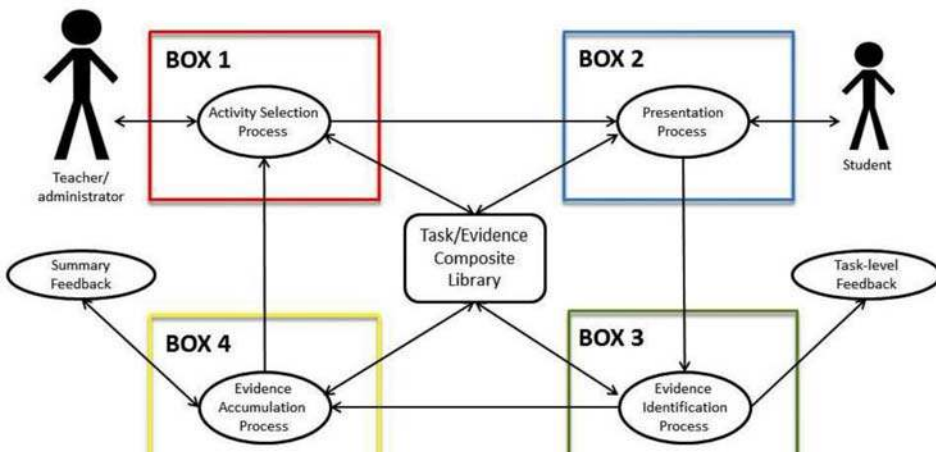


Figure 6. The Four-Process Architecture (from Almond, Steinberg, & Mislevy, 2002).

Had ECD been used to develop the APM Test, its design would focus on the evidentiary link between reasoning ability (i.e., KSAs) and the success of identifying missing elements to complete abstract patterns (i.e., tasks to probe KSAs). The CAF and the Four-Process Architecture (Almond, Steinberg, & Mislevy, 2002) would guide the development of individual items, including choice of patterns to complete, missing elements, and decisions about whether the APM Test should be multiple-choice or open-ended, dichotomously or polytomously scored for increasing sophistication of responses and how scores are aggregated—all in support of the claim that reasoning ability is measured by this particular pattern completion task. Although ECD provides a sophisticated and comprehensive guide for test design, Leighton and Gierl (2011) noted that it is relatively ambivalent on the need for the KSAs to have an empirical cognitive scientific basis. For example, Mislevy and Haertel (2006) indicated that a domain analysis can be based on “transdisciplinary” research on learning but not that it is a requirement (p. 7). Next, we present Embretson’s (1998) Cognitive Design System, which, as the name suggests, focuses on integrating advances in cognitive science with assessment design.

Cognitive design system

The CDS (Embretson, 1998) involves two parts—a conceptual and a procedural framework divided into approximately 10 stages. The *conceptual framework* serves to provide a theoretical and empirically based rationale for the measurement of specific behaviors that are intended to serve as indicators of KSAs (Cronbach & Meehl, 1955; Kane, 2006). The conceptual framework explicitly includes reference to theoretical and empirical research from the cognitive and learning sciences to support choices for (a) measuring behaviors associated with KSAs, (b) designing tasks with particular features to probe for behaviors, and (c) providing explanations for how task features elicit evidence for the KSAs of interest. The conceptual framework is translated into a procedural framework described by the stages shown in Table 1.

The first two stages in Embretson’s CDS are standard and can be found in almost all conventional test design plans (Schmeiser & Welch, 2006). However, Stages 3 through 6 are unique in that they emphasize information from the cognitive and learning sciences to guide task design. Particularly, in Stages 3 through 6, a cognitive model, informed by the empirical research literature, guides the design, evaluation, and psychometric analysis of the items or tasks.

Embretson (1998) applied CDS to generate items to measure reasoning ability similar to those in the APM test. She used Carpenter, Just, and Shell’s (1990) processing theory, which provides experimental evidence that the main cognitive processes involved in solving APM items are “generating and evaluating relationships across the rows and columns” (Embretson, 1998, p. 384). Moreover, working memory capacity is shown to be essential in carrying out these processing functions. Based on this cognitive scientific work, Embretson (1998) designed items that reflected systematic combinations of rules for generating and evaluating pattern relationships across rows and columns, administered the newly created items to a sample of participants, examined participants’ performance on the items with their performance on other tests designed to measure working memory such as the Armed Services Vocational Aptitude Battery, and evaluated the psychometric quality of the newly created items by fitting different item response theory (IRT) models to participants’ test data. This latter focus on psychometric quality is shown in Table 2.

As shown in Tables 1 and 2, the overall goal in CDS is to create a cognitive IRT model to summarize student test performance based on what is scientifically known about how students process information related to a construct of interest. Embretson (1998) defined cognitive IRT models as “jointly mathematical models of cognitive processes and IRT models of response patterns. These models contain parameters to represent the cognitive demands of items as well as the person’s ability” (p. 384).

CDS has not been used to design assessments of learning in TRES. Nonetheless, CDS may be well suited for this purpose because it can accommodate the inclusion of psychologically complex KSAs. CDS has exerted most of its influence in the development of cognitively rich items with predictable

Table 1. Embretson's cognitive design system—initial six stages (procedural framework).

Stage	Function	Question(s) Answered
1. Specify general goals of measurement	a. Construct representation b. Nomothetic span	a. What does the construct measured entail? b. Why is the construct significant?
2. Identify design features in item/task domain	a. Item/task-general features b. Item/task-specific features	a. What are the mode, format, and conditions under which the item will be presented? b. What cognitive skills are probed by specific item features?
3. Develop a cognitive model	a. Review theories b. Select or develop model for psychometric domain c. Revise model d. Test model	a. What are demonstrable behaviors associated with the construct? b & c. What is the best way to operationalize the construct given the purpose of the measurement? Can it be used to design items/tasks? d. Does the model developed stand up to empirical scrutiny?
4. Evaluate cognitive model for psychometric potential	a. Evaluate cognitive model plausibility on current test b. Evaluate impact of complexity factors on psychometric properties	a. Does this model align with the items/tasks included in the assessment? Can this model be used to guide item design? b. Does this model increase or decrease item difficulty of the construct as expected? Do the item features manipulated for each item based on the cognitive model lead to construct relevant increases (or decreases) in overall item difficulty?
5. Specify item distributions on cognitive complexity	a. Distribution of item complexity parameters b. Distribution of item features	a & b. Does this model lead to the expected overall distribution(s) of item difficulty based on manipulated item features?
6. Generate items to fit specifications	a. Artificial intelligence	

psychometric properties. For example, the explicit link established in CDS between cognitive scientific theory and item features has resulted in the formulation of precise and detailed *item models*. Item models provide directions to computer algorithms for generating large quantities of items at a fraction

Table 2. Embretson's Cognitive Design System—Final Four Stages (Procedural Framework).

Stage	Function	Question(s) Answered
7. Evaluate cognitive and psychometric properties for revised test domain	a. Estimate component latent trait model parameters b. Evaluate plausibility of cognitive model c. Evaluate impact of complexity factors on psychometric properties d. Evaluate plausibility of the psychometric model e. Calibrate final item parameters and ability distributions	a, c, d, e. Is the cognitively enhanced psychometric model working well to summarize student performance? b. Does the cognitive model (as a hypothesis of the variables that reflect student mastery of the construct) explain student performance?
8. Psychometric evaluation	a. Measuring processing abilities b. Banks the items by cognitive processing demands	a. What knowledge, skills, and attributes characterize strong versus weak student performance? b. How can items be best categorized by the knowledge, skills, and attributes measured in students?
9. Assemble test forms to represent specifications	a. Fixed content test b. Adaptive test	a. How can test forms be assembled if the same test content is delivered to all students? b. How can test forms be assembled if the test content is to be adapted to the ability level of students?
10. Validation: Strong program of hypothesis testing		a. Does the assessment measure what it set out to measure?

of the cost normally spent on human-generated test items (e.g., Daniel & Embretson, 2010; Embretson & Yang, 2006; Gierl & Lai, 2012; Gorin & Embretson, 2012).

Assessment engineering

AE (see Luecht, 2013) combines the core ideas of both ECD and CDS, but builds on these by providing a highly detailed operational process (Luecht, 2013) for manufacturing assessments (see Figure 7).

Similar to CDS's item model, but with greater detail, Luecht (2013) explained that a task model is a general frame or scheme that specifies n number of slots needing to be filled with values of variables for systematically creating items. For example, types of variables might include declarative and procedural knowledge components, relationships among components, content, and context, in order to create a family of items. Luecht (2013) strongly recommended that the variables chosen for task models be based on cognitive scientific research. In light of the APM test, AE would highlight the need to identify all the variables required to generate matrix items. For example, in addition to variables cuing for generation and evaluation of pattern relationships across rows and columns, variables such as pattern choice and form of presentation would be included to specify all conceivable aspects of generating a family of related items.

AE offers additional ideas for guiding the manufacturing of high-quality assessments. A key idea is that the construct measured by the test and subsequent claims about student performance should be connected at the outset and explicitly linked to an *ordered proficiency scale*. An ordered proficiency scale, shown with a number 1 in Figure 7, indicates what it means to be at increasingly higher levels along the construct of interest, including the performance-level descriptors that provide evidence for this position on the scale. AE underscores the need to show how test item content and cognitive complexity become more sophisticated at increasingly higher points along the scale. Applying this to the APM test would involve knowing how increasingly higher scores along the 48-item continuum translate to performance descriptions of enhanced reasoning ability. Luecht (2013) indicates the instrumentality of cognitive psychology to meet this end. Cognitive scientific research can specify the variables that have been shown to systematically increase cognitive load for human information processing (e.g., Carpenter et al., 1990; Kalyuga, 2011, 2012; Sweller, Ayres, & Kalyuga, 2011).

One of the reasons that Luecht (2013) emphasized attention on cognitive complexity, which is often neglected in traditional assessment design in favor of content, is to generate task model maps that are

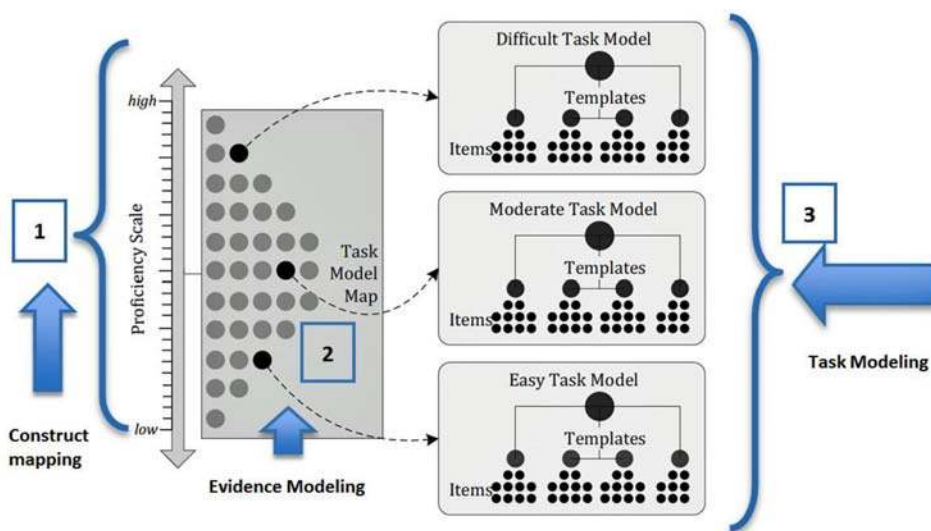


Figure 7. Luecht's (2013) development of task models based on the cognitive psychological literature.

grounded in empirical evidence. Shown with a number 2 in Figure 7, the evidence model becomes a repository of relevant empirically based task variables to guide the design of task model templates at different levels of difficulty along the ordered proficiency scale. This aspect of AE is shown with a number 3 in Figure 7. These task model templates are intended not only to formalize the content and cognitive specifications of items but also to automate their development. Automation can be achieved because task models delineate, within certain parameters, the slots that need to be filled with varying information to produce tasks at particular difficulty levels.

Having described the historical and current landscapes of student assessment, as well as three principled assessment design frameworks, we now conclude with our vision on where we can go from here regarding assessment design and development.

Section 4—The future of innovative assessment design

As discussed in the first section of this article, student testing—especially standardized summative assessment in North America—has a long and controversial history. Controversy has focused on how standardized test results are sometimes used to open or close opportunities for learners, some would argue unfairly. To ensure accuracy and fairness, and to avoid the potential for misuse, a rigorous science of assessment has developed and is continually enforced by professional organizations, government ministries, and university scholars. One issue that has dominated policy and research in the design and development of assessments is how to generate items and tasks that contain the necessary features to validly, reliably, and fairly measure the knowledge and skills we seek to measure in students. In addition, aligning assessments against the learning environments in which students are acquiring knowledge and skills is vital to enhance the suitability of the assessment for measuring specific knowledge and skills and therefore improve the validity of inferences made about student mastery. It would make little sense to teach students to drive a car in vivo—out on the road—and then administer a paper-based driving test that has little or no basis in the performance-based knowledge and skills they have acquired on the road. Likewise, technologically rich environments, the learning contexts in which many students find themselves, have created a need to reconsider quite dramatically the design and development of assessments.

We need technologically enhanced assessments to properly measure the knowledge and skills students are learning today and important competencies they will need to acquire in order to be prepared to meet 21st-century economic and labor force demands. However, the science of assessment and its core principles are grounded in paper-based, discrete knowledge-based items, most of which are somewhat divorced from the performance-based learning environments that make up the contexts students encounter today. New assessments must be designed, developed, and evaluated. Furthermore, core ideas about how to define and operationalize the standardization of tasks, task difficulty, task discrimination, reliability of performance, and validity arguments must be outlined relative to TREs. Toward this end, state-of-the-art innovative assessments need to be developed using *principled assessment design* to ensure the development of increasingly accurate and sensitive measures of learners' knowledge and skill acquisition. Several principled assessment designs that could direct the development of innovative assessments, and meet psychometric standards and possess psychological rigor, were presented in this article.

Principled assessment designs such as ECD, CDS, and AE provide comprehensive frameworks for designing tasks that are expected to evoke knowledge and skills of interest in students, thus generating evidence for claims about student learning of these knowledge and skills. These designs also blur the line between what can be considered to be formative or summative, because assessments created from these designs are expected to be equally informative to student learning irrespective of whether there is a grade or score attached to student performance. That is, instead of distinguishing between formative versus summative, we are evolving in our understanding to realize that a better way to begin to think of assessments generally is as *informative* to student learning; regardless of a specific grade or other high-stakes consequence, they can be used to provide needed information about what has been achieved

versus not, and what can be done now to address areas of weakness (DiCerbo & Behrens, 2012b; Leighton & Gierl, 2011; Shute et al., 2009). Principled assessment design may not yield assessments that specify individual learning plans post assessment (as this is something that would need to be coordinated by the teaching community), but they can orient learners—and possibly other stakeholders—as to the current status of the learner and to where attention should be focused for the next phase of learning.

The vision

To bring our vision of next-generation assessments into focus, imagine an educational system in which high-stakes tests are no longer the dominant means to drive educational reforms through accountability systems. Instead, students would progress through their school years engaged in different learning contexts, all of which capture and measure growth in valuable cognitive and noncognitive skills. This information would then be used to further enhance student learning.

In our complex, interconnected digital world, we are learning constantly and producing numerous digital footprints or data along the way. This vision does not refer to simply administering assessments more frequently (e.g., each week, each day) but rather continually collecting data as students interact with digital environments (Level 3, in Figure 3) both inside and, of importance, outside of school. When the various data streams coalesce (Level 4, in Figure 3), the accumulated information can potentially provide increasingly reliable and valid evidence about what students know and can do across multiple contexts. The vision involves high-quality, ongoing, unobtrusive assessments embedded in various TREs that can be aggregated to inform a student's evolving competency levels (at various grain sizes) and aggregated across students to inform higher level decisions (e.g., from student, to class, to school, to district, to state, to country). In short, our take on assessment is that it should (a) support, not undermine, the learning process for learners; (b) provide ongoing formative information (i.e., be part of a system of giving useful feedback during the learning process that informs and complements a summary judgment at the end); and (c) be responsive to what is known about how people learn, generally and developmentally.

This vision of assessment has its primary goal to improve learning (e.g., Black & Wiliam, 1998; Leighton & Gierl, 2011; Shute, 2009; Stiggins, 2002), and we believe that it's critical to support the kinds of learning outcomes and processes necessary for students to succeed in the 21st century. Again, most current approaches to assessment/testing are too disconnected from learning processes. That is, the typical classroom cycle is: Teach. Stop. Administer test. Go loop (with new content). But consider the following metaphor representing an important shift that occurred in the world of retail outlets (from small businesses to large department stores), suggested by Pellegrino et al. (2001, p. 284). No longer do these businesses have to close down once or twice a year to take inventory of their stock. Instead, with the advent of automated checkout and barcodes for all items, these businesses have access to a continuous stream of information that can be used to monitor inventory and the flow of items. Not only can businesses continue without interruption, but the information obtained is far richer, enabling stores to monitor trends and aggregate the data into various kinds of summaries, as well as support real-time, just-in-time inventory management. Similarly, with new assessment technologies, schools should no longer have to interrupt the normal instructional process at various times during the year to administer external tests to students. Instead, assessment should be continual and invisible to students, supporting real-time, just-in-time instruction and other types of learning support.

The envisioned ubiquitous nature of assessment will require a reconceptualization of the boundaries of the educational system. That is, the traditional way of teaching in classrooms today involves providing lectures and giving tests in class, then assigning homework to students to complete outside of class (usually more reading on the topic and perhaps answering some topical questions). Alternatively, consider a relatively new pedagogical approach called “flipped classrooms.” This involves a reversal of the traditional approach where students first examine and interact with a target topic by themselves at home and at their leisure (e.g., viewing an online video and/or playing an educational game), and then in class, students apply the new knowledge and skills by solving problems and doing practical work (see

Bergmann & Sams, 2012). The flipped classroom is already operational for core courses at some universities (e.g., Alberta) across North America. The teacher supports the students in class when they become stuck, rather than delivering the initial lesson in person. Flipped classrooms free up class time for hands-on work and discussion and permit deep dives into the content. Students learn by doing and asking questions, and they can help each other, a process that benefits a majority of learners (Strayer, 2012). It is important to note that our working hypothesis is that these efforts are more likely to engage a greater population of teachers and learners—including poor and unmotivated learners. If this wasn't a problem in the current educational landscape, then perhaps TREs might make little sense as a potential solution to a problem that didn't exist. However, flipped classrooms and TREs may provide affordances to engage, teach, and reach a greater population of learners than is currently possible with traditional pedagogical practices.

Limitations and future research

For this vision of the future of assessment—as ubiquitous, unobtrusive, engaging, and valid—to gain traction, there are a number of large hurdles to overcome. We describe four of the more pressing issues that need more research.

The first hurdle relates to variability in the *quality* of assessments within TREs. That is, because schools are under local control, students in a given state could engage in thousands of TREs during their educational tenure. Teachers, publishers, researchers, and others will be developing TREs, but with no standards in place, they will inevitably differ in curricular coverage, difficulty of the material, scenarios and formats used, and many other ways that will affect the adequacy of the TRE, tasks, and inferences on knowledge and skill acquisition that can justifiably be made from successfully completing the TREs. The principle design frameworks we presented in Section 3 represent a design methodology but not a panacea, so more research is needed to figure out how to equate TREs or create common measurements (i.e., standardize) from diverse environments. Toward that end, there must be common models employed, perhaps linked to the Common Core State Standards, across different activities, curricula, and contexts. Moreover, it is important to figure out how to interpret evidence where the activities may be the same but the contexts in which students are working are different (e.g., working alone vs. working with another student). To illustrate, Kim and Shute (2015) examined two versions of Physics Playground. The two versions of the game differed only in terms of linearity in gameplay. The linear version required students to play levels in a fixed order, and the nonlinear version permitted students to choose any level they wanted to play. Investigation of the assessment qualities—validity, reliability, and fairness—suggested that changing one game element (e.g., linearity) significantly influenced how players interacted with the game, thus changing the evidentiary structure of the in-game measures. The most salient difference between the two versions was found in the ways students thought about and interacted with the game. For example, imposing linearity influenced the perceived goals of gameplay. Students in the linear condition focused on just solving lots of levels, whereas students in the nonlinear condition spent more time per level, striving for more “elegant” solutions. Moreover, students who played the nonlinear version of the game showed significant improvement on qualitative physics understanding measured by the pretest and posttest, whereas the students assigned to the linear condition did not.

The second hurdle involves accurately capturing and making sense of students' *learning progressions*.¹ That is, although TREs can provide a greater variety of learning situations than traditional face-to-face classroom learning, evidence for assessing and tracking learning progressions becomes heterogeneous and complex rather than general across individual students. Thus there is a great need to model learning progressions in multiple aspects of student growth and experiences, which can be applied across different learning activities and contexts (Shavelson & Kurpius, 2012). However

¹We use the term *learning progressions* generally. We acknowledge that some investigators (e.g., Duschl, Maeng, & Sezen, 2011) have differentiated learning progressions in science education from what are called learning trajectories in mathematics education in terms of the length of teaching sequences and granularity of features.

as Shavelson and Kurpius pointed out, there is no single absolute order of progression, as learning in TREs involves *multiple* interactions between individual students and situations, which may be too complex for most measurement theories in use that assume linearity and independence. Clearly, theories of learning progressions in TREs need to be actively researched and validated to realize TREs' potential. For example, Jang, Wagner, and Xu (2014) developed a conceptual framework that models medical students' learning experiences and growth in the problem-based TRE called *Bioworld* (Lajoie et al., 2001) in which students solve clinical cases with interactive feedback. Based on the triangulation of multiple data sources including computer logs, self-reports, and think-aloud protocols, they examined interactive relationships among learner and task variables. Although the Jang et al. framework specifies the interconnected learner variables and mediating task variables, it assumes that individual learners take different learning pathways and can be used to fine-tune their progressions by encouraging students to identify immediate learning goals relevant to their current status. Learner (or data) dashboards (Verbert, Duval, Klerkx, Govaerts, & Santos, 2013) can be used to facilitate communication of learning progressions to students and teachers, but again, more research is needed on data dashboards and their optimal representations.

The third problem to resolve involves the need to expand educational boundaries and resolve impediments to moving toward the idea of the *flipped classroom*. One issue concerns the digital divide, where some students may not have access to a home computer. In those cases, students can be allowed to use library resources or a computer lab. Alternatively, the online components can be accessed via a cell phone, as many students who do not have computers or Internet at home do have a phone that can meet the requirements of online activities. In addition, some critics may argue that flipped classrooms will invariably lead to teachers becoming outdated. However, teachers become even more important in flipped classrooms, where they educate and support rather than lecture (i.e., serve as “guide on the side” rather than “sage on a stage”). This represents an intriguing way to take back some of the very valuable classroom time and serve as a more efficient and effective teacher. Much more empirical research is needed to determine how this pedagogical approach works relative to traditional pedagogies.

Finally, we have to figure out a way to resolve *privacy*, *security*, and *ownership* issues regarding students' information. The privacy/security issue relates to Level 4 assessment (i.e., the accumulation of student data from disparate sources). The recent failure of the \$100 million inBloom initiative (see McCambridge, 2014) showcases the problem. That is, the main aim of inBloom was to store, clean, and aggregate a wide range of student information for states and districts and then make the data available to district-approved third parties to develop tools and dashboards so that the data could be easily used by classroom educators. The main issue boils down to this: Information about individual students may be at risk of being shared far more broadly than is justifiable. Because of the often high-stakes consequences associated with tests, many parents and other stakeholders fear that the data collected (reminiscent of Orwell's “Big Brother”) could later be used against the students. Related to this hurdle are the “data dashboards” themselves, which can be intimidating and thus unused by classroom teachers. Almond, Shute, Underwood, and Zapata-Rivera (2009) described various graphical ways to present Bayesian network-based proficiency estimates from students to teachers, but much more research is needed in this area.

What would it take to implement the vision once the hurdles are surmounted? We use ECD to illustrate. In addition to ECD's ability to handle multivariate competency models (Mislevy, Steinberg, & Almond, 2003), it is able to accumulate evidence across disparate sources (e.g., homework assignment, in-class quiz on an iPad, high score on a video game). This is possible, as ECD provides assessment designers with processes that enable them to work through the design trade-offs that involve multiple competency variables—either within one assessment or across multiple assessments. The “alchemy” involves turning the raw data coming in from various sources into evidence. Evidence models will need to be able to interpret the results of all of the incoming data for the purposes of updating the student model. The rules of evidence must describe which results can be used as evidence, as well as any

transformation that needs to be done to those results (e.g., averaging, rescaling, setting cut scores; see Almond, 2010, for more on this process).

As mentioned in Hurdle 1 (i.e., differential assessment quality), we need to find the right probabilities (or other parameters) that will complete the evidence models. One solution is to use Bayesian logic on the parameters of the system. An assessment designer could complete a questionnaire for each task included in the system. This would help the designer define the observable(s) for each task and describe how they are related to competency variables. It would also ask questions about the strengths of the relationships, used to produce prior distributions for the parameters of the ECD models (see Almond, 2010). The prior parameters could then be used to immediately score the student interactions with the digital activity (e.g., a TRE, online quiz, or game). As sufficient data (i.e., outcomes from students' interactions with a collection of tasks) become available, Bayesian inference can be used to replace the prior distributions for parameters with posterior distributions. This should improve the quality of inferences that come from the system.

Despite the foregoing hurdles and limitations, constructing the envisioned ubiquitous and unobtrusive assessments across multiple learner dimensions, with data accessible by diverse stakeholders, could yield various educational benefits. First, the time spent administering tests, handling make-up exams, and going over test responses is not very conducive to learning. Given the importance of time on task as a predictor of learning, reallocating those test preparation activities into ones that are more educationally productive would provide potentially large benefits to almost all students. Second, by having assessments that are continuous and ubiquitous, students are no longer able to “cram” for an exam. Although cramming can provide good short-term recall, it is a poor route to long-term retention and transfer of learning. Standard assessment practices in school can lead to assessing students in a manner that is in conflict with their long-term success. With a continuous assessment model in place, the best way for students to do well is to do well every day.

The third direct benefit is that this shift in assessment mirrors the national shift toward evaluating students on the basis of acquired competencies. With increasing numbers of educators growing wary of pencil-and-paper high-stakes tests for students, this shift toward ensuring students have acquired “essential” skills fits with the idea of our envisioned future of assessment. Finally, with a slight change in pedagogical focus (from teacher centered to more student centered)—as manifest, for example, in flipped classrooms—this can address the serious shortfall of time that is needed to interact with particularly rich TREs. We can fix time and let outcomes vary, or we can fix outcomes (set to high standards) and let time vary. The problem in classrooms is not having enough time. Furthermore, as described earlier, current research comparing the quality of new versus traditional assessments on the Microsoft Certified Systems Engineer Certification Program (see Jodoin, 2003) showed that the innovative items provided more information about students' knowledge and skills but less information per time unit.

The time is now ripe for such assessments given the dire need for supporting new 21st-century skills and the increased availability of computer technology. New technologies make it easy to capture the results of routine student work—in class, at home, or wherever. It could be that 21st-century assessment will be so well integrated into students' day-to-day work that the students don't even know it is there. This represents quite a contrast to our current testing contexts.

Finally, although the benefits of using a seamless-and-ubiquitous model to run a business have been clear for more than four decades, applying this metaphor to education may require some adjustments, as we are dealing with humans, not goods. For instance, one risk associated with our vision is that students may come to feel as if they are constantly being evaluated, which could negatively affect their learning and possibly add stress to their lives. Another risk of our continuous assessment vision could result in teaching and learning turning into ways to “game the system” depending on how it is implemented and communicated. But the aforementioned hurdles and risks, being anticipated and researched in advance, can help to shape the vision for a richer, deeper, more authentic assessment (to support learning) of students in the future. How many current businesses would elect to return to pre-barcode days?

References

- Alberta Education. (2013). *General information bulletin: Introduction to the diploma examination program*. Edmonton, AB: Alberta Education. Retrieved from http://www.education.alberta.ca/media/6902694/03-dip-gib-2013-14_intro%20revisions_2013-08-01.pdf
- Almond, R. G. (2010). Using Evidence Centered Design to think about assessments. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs* (pp. 75–100). New York, NY: Springer-Verlag.
- Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning*, 50, 450–460.
- Almond, R. G., Steinberg, L. S., & Mislavy, R. J. (2002). Available from <http://www.jtla.org> Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 1(5).
- American Diploma Project. (2004). Ready or not: Creating a high school diploma that counts. (A partnership of Achieve Inc., The Education Trust, & Thomas B. Fordham Foundation). Retrieved from <http://www.achieve.org/files/ReadyorNot.pdf>
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2010). Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In M. Khine & I. Saleh (Eds.), *New science of learning: Computers, cognition, and collaboration in education* (pp. 225–247). Amsterdam, the Netherlands: Springer.
- Baker, E. L., & O'Neil, H. F. Jr., (1995). Computer technology futures for the improvement of assessment. *Journal of Science Education and Technology*, 4, 37–45.
- Behrens, J. T., Frezzo, D. C., Mislavy, R. J., Kroopnick, M., & Wise, D. (2008). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). New York, NY: Erlbaum.
- Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement*, 24, 294–309. doi:10.1177/01466210022031769
- 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, Research and Development Series* (NCES 2007-466). Washington, DC: National Center for Education Statistics.
- Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, 28, 77–92. doi:10.1111/j.1745-3984.1991.tb00345.x
- Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. Alexandria, VA: International Society for Technology in Education.
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2010). *Defining 21st-century skills*. (White Paper 1). Retrieved from Assessment & Teaching of 21st Century Skills website: <http://atc21s.org/wp-content/uploads/2011/11/1-Defining-21st-Century-Skills.pdf>
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Chapter 2: Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). New York, NY: Springer. doi:10.1007/978-94-007-2324-5_2
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice*, 5, 7–74.
- Blanchard, E. G., Wiseman, J., Naismith, L., & Lajoie, S. P. (2012). realistic digital deteriorating patient to foster emergency decision-making skills in medical students. In *12th IEEE International Conference on Advanced Learning Technologies* (pp. 74–76). Rome, Italy: IEEE Communications Society. doi:10.1109/ICALT.2012.44
- Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, 30, 3–12.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404–431.
- Cohen, A. D., & Upton, T. A. (2007). "I want to go back to the text": Response strategies on the reading subtest of the New TOEFL. *Language Testing*, 24, 209–250.
- Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, 16, 555–575. Retrieved from <http://www.cs.ubc.ca/~conati/my-papers/jaa102.pdf>
- Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19, 267–303. doi:10.1007/s11257-009-9062-8
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302. Retrieved from <http://psychclassics.yorku.ca/Cronbach/construct.htm>
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper and Row.
- Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, 34, 348–364. doi:10.1177/0146621609349801

- Davey, T., Godwin, J., & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement*, 34, 21–41. doi:10.1111/j.1745-3984.1997.tb00505.x
- DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education*, 11, 279–299. doi:10.1207/s15324818ame1103_4
- DiCerbo, K. E., & Behrens, J. T. (2012a, April). *From technology-enhanced assessments to assessment-enhanced technology*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.
- DiCerbo, K. E., & Behrens, J. T. (2012b). Implications of the digital ocean on current and future assessment. In R. Lizzitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age.
- Duncan, A. (2009). *Economic security and a 21st century education*. Retrieved from U.S. Department of Education website: <http://www.ed.gov/news/speeches/economic-security-and-21st-century-education>
- Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, 47, 123–182. doi:10.1080/03057267.2011.604476
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 300–396. doi:10.1037/1082-989X.3.3.380
- Embretson, S. E., & Yang, X. (2006). Multicomponent latent trait models for complex tasks. *Journal of Applied Measurement* 7, 335–350. Retrieved from https://smartech.gatech.edu/bitstream/handle/1853/34254/embretson_JAM_2006.pdf?sequence=1
- Facer, K. (2003). *Computer games and learning: Why do we think it's worth talking about computer games and learning in the same breath?* Future Lab Series. Retrieved from http://admin.futurelab.org.uk/resources/documents/discussion_papers/Computer_Games_and_Learning_discpaper.pdf
- Gierl, M. J., & Haladyna, T. M. (2012). Automatic item generation: An introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 3–12). New York, NY: Routledge.
- Gierl, M. J., & Lai, H. (2012). Using item models for automatic item generation. *International Journal of Testing*, 12, 273–298. doi:10.1080/15305058.2011.635830
- Gorin, J. S., & Embretson, S. E. (2012). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 136–156). New York, NY: Routledge.
- Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes. *Journal of Economic Growth*, 17, 267–321. doi:10.1007/s10887-012-9081-x
- Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying fusion model to LanguEdge assessment. *Language Testing*, 26, 31–73. doi:10.1177/0265532208097336
- Jang, E. E. (2009b). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, 6, 210–238. doi:10.1080/15434300903071817
- Jang, E. E. (2014). *Focus on assessment*. New York, NY: Oxford University Press.
- Jang, E. E., & Wagner, M. (2014). Diagnostic feedback in language classroom. In A. Kunnan (Ed.), *Companion to language assessment* (vol. 2, pp. 693–711). New York, NY: Wiley-Blackwell.
- Jang, E. E., Wagner, M., & Xu, Z. (2014, April). *Ecological assessment framework in computer-based learning environments*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Jenkins, H., Purushotma, R., Weigel, M., Clinton, K., & Robinson, A. J. (2009). *Confronting the challenges of participatory culture: Media education for the 21st century*. Cambridge, MA: MIT Press.
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40, 1–15. doi:10.1111/j.1745-3984.2003.tb01093.x
- Joint Information Systems Committee. (2007). *Effective practice with e-assessment: An overview of technologies, policies and practice in further and higher education*. London, England: Higher Education Funding Council for England. Retrieved from <http://www.jisc.ac.uk/media/documents/themes/elearning/effpraceassess.pdf>
- Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, 23, 1–19. doi:10.1007/s10648-010-9150-7
- Kalyuga, S. (2012). Interactive distance education: A cognitive load perspective. *Journal of Computing in Higher Education*, 24, 182–208. doi:10.1007/s12528-012-9060-4
- Kamin, L. J. (1995). The pioneers of IQ testing. In R. Jacoby & N. Glauberman (Eds.), *The bell curve debate: History, documents, opinions* (pp. 476–509). New York, NY: Times Books.
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.
- 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.
- Klopfer, E., Osterweil, S., Groff, J., & Haas, J. (2009). *Using the technology of today, in the classroom today: The instructional power of digital games, social networking, simulations and how teachers can leverage them*. Cambridge, MA: MIT, The Education Arcade. Retrieved from http://education.mit.edu/papers/GamesSimsSocNets_EdArcade.pdf

- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.
- Lajoie, S. P., Lavigne, N. C., Guerrero, C., & Munsie, S. D. (2001). Constructing knowledge in the context of Bioworld. *Instructional Science*, 29, 155–186. doi:10.1023/A:1003996000775
- Lawrence, I. M., Rigol, G. W., Van Essen, T., & Jackson, C. A. (2003). *A historical perspective on the content of the SAT* (College Board Research Report No. 2003-3, ETS RR-03-10). Retrieved from www.research.collegeboard.org
- Leahey, T. H. (1992). The mystical revolutions of American psychology. *American Psychologist*, 47, 308–318. doi:10.1037/0003-066X.47.2.308
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6–15.
- Leighton, J. P., Chu, M-W., & Seitz, P. (2013). Cognitive diagnostic assessment and the learning errors and formative feedback (LEAFF) model. In R. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 183–207). Charlotte, NC: Information Age.
- Leighton, J. P., & Gierl, M. J. (2007a). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, MA: Cambridge University Press.
- Leighton, J. P., & Gierl, M. J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26, 3–16. doi:10.1111/j.1745-3992.2007.00090.x
- Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment*. Cambridge, MA: Cambridge University Press.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Luecht, R. M. (2001, April). *Capturing, codifying and scoring complex data for innovative, computer-based items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Luecht, R. M. (2013). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59–76). New York, NY: Routledge.
- Massachusetts Institute of Technology. (2013). *Scheller teacher education program: The education arcade*. Cambridge, MA: Retrieved from <http://education.mit.edu/> Author.
- Mayrath, M. C., Clarke-Midura, J., Robinson, D. H., & Schraw, G. (2012). Introduction to technology-based assessments for the 21st-century skills. In M. C. Mayrath, J. Clark-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research* (pp. 13–54). Charlotte, NC: Information Age.
- McCambridge, R. (2014). Legacy of a failed foundation initiative: inBloom, Gates and Carnegie. *Nonprofit Quarterly*. Retrieved from <https://nonprofitquarterly.org/policysocial-context/24452-legacy-of-a-failed-foundation-initiative-inbloom-gates-and-carnegie.html>
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23. doi:10.3102/0013189X023002013
- Michigan State University. (2013). *Matrix: Center for digital humanities & social sciences*. East Lansing, MI: Author. Retrieved from <http://www2.matrix.msu.edu/>
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Rep. No. RR-03-16). Retrieved from Educational Testing Service website: <http://www.ets.org/Media/Research/pdf/RR-03-16.pdf>
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for educational testing. *Educational Measurement: Issues and Practice*, 25, 6–20.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
- National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards*. Washington, DC: Author. Retrieved from <http://www.corestandards.org/read-the-standards/>
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Organization for Economic Co-operation and Development. (2013). *PISA 2012 results in focus: What 15-year-olds know and what they can do with what they know*. Retrieved from <http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf>
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer-Verlag.
- Partnership for 21st-Century Skills. (2008). *21st-century skills, education & competitiveness: A resource and policy guide*. Retrieved from http://www.p21.org/storage/documents/21st_century_skills_education_and_competitiveness_guide.pdf
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Section 4, Advanced Progressive Matrices*. Oxford, UK: Oxford Psychologists Press.

- Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 21, 115–133. doi:10.3233/JAI-2011-019
- Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment. *American Psychologist*, 64, 215–227. doi:10.1037/0003-066X.63.4.215
- Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment*, 4(6). Retrieved from <http://files.eric.ed.gov/fulltext/EJ843857.pdf>
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: Praeger.
- Shavelson, R. J., & Kurlius, A. (2012). Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 13–26). Rotterdam, the Netherlands: Sense Publishers.
- Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 139–187). New York, NY: Erlbaum.
- Shute, V. J. (2008). Focus on formative assessment. *Review of Educational Research*, 78, 153–189. Retrieved from http://myweb.fsu.edu/vshute/pdf/shute%202008_b.pdf
- 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.
- Shute, V. J., & Becker, B. J. (2010). Prelude: Assessment for the 21st century. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs* (pp. 1–11). New York, NY: Springer-Verlag.
- Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 43–58). New York, NY: Springer.
- Shute, V. J., & Kim, Y. J. (2011). Does playing the World of Goo facilitate learning? In D. Y. Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 359–387). New York, NY: Routledge. Retrieved from <http://myweb.fsu.edu/vshute/pdf/goo.pdf>
- Shute, V. J., & Kim, Y. J. (2012). E-Assessment. In J. Balacheff, J. Bourdeau, P. Kirschner, R. Sutherland, & J. Zeiliger (Eds.), *TEL Thesaurus*. Retrieved from <http://www.tel-thesaurus.net/wiki/index.php/E-Assessment>
- Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments. *Technology, Instruction, Cognition, and Learning*, 8, 137–161. Retrieved from <http://myweb.fsu.edu/vshute/pdf/TICL2010.pdf>
- Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: MIT Press. Retrieved from http://myweb.fsu.edu/vshute/pdf/Stealth_Assessment.pdf
- Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of informal physics in Newton’s Playground. *The Journal of Educational Research*, 106, 423–430.
- Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). New York, NY: Routledge.
- Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Erlbaum.
- Sternberg, R. J., & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher*, 24(6), 9–17. doi:10.3102/0013189X024006009
- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, 83, 758–765.
- Strayer, J. (2012). How learning in an inverted classroom influences cooperation, innovation and task Orientation. *Learning Environments Research*, 15, 171–193.
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer.
- Take-Two Interactive Software. (2010). *Sid Meier’s civilization*. Retrieved from <http://www.civ3.com/civ3.cfm>
- UBM Tech. (2014). *Gamasutra: The art & business of making games*. Retrieved from <http://www.gamasutra.com/>
- U.S. Department of Education. (2009, November 9). *Economic security and a 21st century education: Secretary Arne Duncan’s remarks at the U.S. chamber of commerce’s education and workforce summit*. [Speeches: Archived information]. Retrieved from <http://www2.ed.gov/news/speeches/2009/11/11092009.html>
- Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learner Analytics Dashboard Applications. *American Behavioral Scientist*, 51, 1500–1509. doi:10.1177/0002764213479363
- Wan, L., & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25, 58–78. doi:10.1080/08957347.2012.635507
- Zwick, R. (2004). Is the SAT a “wealth test?” The link between educational achievement and socioeconomic status. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 203–216). New York, NY: Routledge Falmer.