CHAPTER 9

MEASURING & SUPPORTING LEARNING IN EDUCATIONAL GAMES

Valerie J. Shute, Ph.D., Florida State University
Seyedahmad Rahimi, M.A., Florida State University
Chen Sun, M.A., Florida State University

"The significant problems of our time cannot be solved by the same level of thinking that created them."
–Albert Einstein

We begin this chapter with a general description about current problems in U.S. education, then focus on the unfortunate side effects of standardized testing in K-12 schools. For instance, the increased frequency of administering high-stakes tests tends to (a) narrow the curricula, (b) have dire consequences (for students, teachers, and schools), and (c) exclude support for skills important for success in the 21st century, such as problem solving, critical thinking, creativity, persistence, and collaboration. We also describe specific advances in learning science, psychometrics, and technology that can be leveraged to create more effective and engaging assessments to support the aforementioned new competencies—*educational games*.

Advances in learning science have begun to identify factors affecting learners' growth in knowledge, skills, and other attributes. Additionally, advances in psychometrics allow us to measure complex, multi-dimensional constructs in an ongoing manner, and at a refined grain size. That information, then, can be used diagnostically as the basis for targeted and adaptive learning support provided to students, teachers, and/or back to the digital learning environment (e.g., educational game) to activate learning, reflection, and self-regulation, provide information for teachers to figure out subsequent instructional support, and appropriately alter the environment or task, respectively.

Technological advances can be leveraged toward education in general (and assessment in particular) to accomplish things that were impossible before, such as gathering and analyzing "big data" – broadly defined as massive quantities of student-related information that can be mined to make inferences about various attributes and outcomes. Big data and associated emerging analysis techniques represent the intersection of computer science, statistics, and ethics or policy (see 2015 NSF report: http://cra.org/wp-

content/uploads/2015/10/CRAEducationReport2015.pdf). Such accumulated data comprises evidence that can be used to improve students' learning and performance by providing accurate, actionable, and informative assessment information.

Status Quo

The United States is one of the world leaders in spending money for education. For example, in 2011–12, the U.S. spent $621 billion for public elementary and secondary schools (U.S. Department of Education, National Center for Education Statistics, 2015). However, the return on investment is not commensurate with these expenditures. According to the *Program for International Student Assessment* (Organization for Economic Co-operation and Development, 2013), U.S. students were outperformed in mathematics literacy proficiency by students in 29 countries, in science literacy by 22 countries, and in problem solving by 18 countries (the average score of these countries was statistically higher than the U.S.'s average score). These results were based on the average scores per test from students in 65 participating countries (i.e., mathematics, science, and problem-solving). Moreover, based on a report from the Economist Intelligence Unit (2012), which combines different international data sets (e.g., PISA, Trends in International Mathematics and Science Study—TIMSS, and Progress in International Reading Literacy Study—PIRLS) together with each country's academic data (e.g., literacy and graduation rates), U.S. education is ranked 14th out of forty countries. Why is this a problem? We live in a world with complex problems (e.g., climate change, nuclear proliferation, cyber security, racial and religious intolerance, and so on). We need people who are able to think critically, creatively, systemically, and work collaboratively with others to help solve these complex problems (Shute, 2011). Based on the aforementioned reports, we are falling behind other countries in terms of producing competent individuals and this can slow down the growth of our country over time (Hanushek & Woessmann, 2012). We need to boost our students' 21st century competencies (e.g., problem solving, creativity, systems thinking) to effectively compete internationally in the near future (Shute, 2007) and accelerate the growth of our country in the long run.

To improve students' 21st century competencies, we first need to accurately assess them. We no longer can rely solely on the old types of assessment (e.g., standardized tests with multiple-choice items) to measure these new competencies. However, we can still use standardized assessments when they are appropriate (e.g.,

when the learning outcome relates to the memorization of facts and equations). Two main reasons for not using the old assessment types to measure the new competencies are: (1) these tests are not capable of measuring the complex new competencies we need in the 21st century (e.g., multidimensional constructs like problem solving skill cannot easily or accurately be measured using multiple choice questions), and (2) they can have negative, unintended consequences.

Educational researchers have been examining the unintended consequences of standardized testing for years (e.g., Amrein & Berliner, 2002; Jones, 2007; Madaus & Russell, 2010; Smith, 1991). Some salient consequences of increased reliance solely on standardized testing include the adverse impacts on instruction and learning goals such as narrowing of curricula in schools, inciting cheating and corruption on high-stakes tests, and generally increasing students' anxiety (Madaus & Russell, 2010). Additional unintended consequences include the impact on teacher evaluations and the accountability of schools, based mainly on student test performance rather than on what teachers do in classrooms.

Negative effects on instruction. Increased emphasis on standardized testing in schools (also known as "teaching to the test") produces a narrowing of curricula (Abrams, Pedulla, & Madaus, 2003; Amrein & Berliner, 2002; Jones, 2007; Madaus & Russell, 2010). And if the tests were all encompassing, that would be fine; but tests are necessarily limited in scope. Educators (i.e., teachers, administrators, boards of education) note what is being assessed in the standardized tests, and then focus instruction towards those subjects, such as reading expository text and writing the 5-paragraph essay. As a result, applied work and interdisciplinary problem solving as well as other competencies and subjects (e.g., social science, art, physical education and even science in the elementary grades) are treated as less important—if not removed altogether from the curriculum (Madaus & Russell, 2010). In short, standardized tests tend to prioritize the subjects taught in school, and instructional goals tend to prepare students to succeed on the standardized tests—not to be creative problem solvers, critical thinkers, and equipped with other 21st century skills (Shute, Leighton, Jang, & Chu, 2016).

Negative effects on learning goals. Another unintended consequence of standardized testing is that these tests can change students' goal orientation from learning to performance (Jones, 2007; Shute, 2008). For instance, focusing on test performance becomes more important than on how much one actually learns. As Jones (2007) observed, high-stakes tests externally motivate students (i.e., supporting a performance goal orientation) rather

than internally motivate students (i.e., supporting a learning goal orientation). However, if students develop learning (as opposed to performance) goals, other positive characteristics may be internalized as they spend their time in school (Jones, 2007). Students with learning goals tend to be more persistent when confronting difficult problems and experiencing failure, use more complex learning strategies, and work on challenging tasks voluntarily (Shute, 2008). In short, students with learning goals are better prepared for living in our increasingly complex world.

Cheating and corruption. Another by-product of focusing too much on high-stakes tests is that this can influence teachers, students, and other stakeholders to cheat (Madaus & Russell, 2010). That is, when the accountability of teachers, quality of schools, students ranking, and even students' future academic paths depend on the results from standardized tests, the likelihood increases that some schools' personnel, students, and other stakeholders will cheat on tests (Berliner, 2011). For example, tests administered in schools may increase the time limits of tests, help students to answer questions, provide the questions to the students before the tests, or even change students' test scores in the system (Amrein & Berliner, 2002). Some students also cheat by impersonation on tests like the GRE or TOEFL. As the Guardian journal reported (Yuhas, 2015), fifteen students with fake Chinese passports were charged with impersonating other students to take GRE, TOEFL, and SAT tests.

Student anxiety. Testing can often create a stressful environment for students (Jones, 2007) that is detrimental to learning. For example, Wheelock, Bebell, and Haney (2000) asked students to draw a picture of themselves when they take a high-stakes test. Not surprisingly, the students illustrated themselves as anxious, frustrated, bored, hopeless, and gloomy. In another study, Culler and Holahan (1980) found that anxiety is negatively related to how well students perform on a test. They pointed out that more anxious students had significantly lower grades than less anxious students, and poorer study skills.

These are just a few of the unintended consequences of high-stakes tests in our educational system. However, we are in a position to address many of these problems. In the next section, we present a vision of what education can be like (compared to the status quo). Spoiler alert—this vision involves no tests -- at least none of the high-stakes, norm-referenced types of tests that are currently the norm in education today.

Vision

How would you feel—as a student—if you were told that there would be no high stakes tests throughout the school year? How would you feel—as a teacher—if you did not have to teach to and administer high stakes tests anymore? How would you feel—as a parent—if you saw your child develop new knowledge and skills and show excitement about learning? The vision presented here is based on the ideas presented in Shute, Leighton, Jang, and Chu (2016). In a nutshell, next-generation assessments can measure (and in some cases, support) students' growth in important cognitive and non-cognitive competencies. High-stakes tests will no longer be the primary means of assessing learning, thus the time that was used to prepare and administer tests can be reallocated to substantive student learning activities.

Advances in technology (e.g., ever-increasing computational power, virtual reality, wearable devices, and social networks) have created a world in which we produce many digital footprints, as data. Students can potentially learn a lot through interacting with well-designed digital environments and the data generated from these interactions can be used for assessing different knowledge and skills. We envision students being assessed unobtrusively, while they are learning (i.e., seamlessly within the digital environment). Further, data can be collected almost continuously and across various contexts (i.e., ubiquitously), providing over time, reliable and valid evidence about students' levels of targeted competencies. This type of unobtrusive and ubiquitous assessment of important knowledge, skills, and other attributes (e.g., dispositions) can provide teachers with rich information to help them guide their students, as well as information to help students develop competencies on their own. In short, we believe that assessment should: (a) support the learning process for learners and not undermine it; (b) provide ongoing formative feedback to students during the learning process (with perhaps summative feedback at the end); and (c) reflect current theories about how people learn, both from a general and developmental perspective.

The envisioned unobtrusive and ubiquitous aspects of assessment will require some changes to the current education system. The traditional way of teaching in classrooms today involves providing lectures and giving tests in class, then having students complete homework at home (without guidance when students may become stuck). An alternative pedagogical approach is the "flipped classroom" where students first examine and interact with content on their own (e.g., playing a particular game at home), then in class, students would apply their new knowledge and skills while instructors and peers are there to support and guide their work (see

Bergmann & Sams, 2012). Having a flipped classroom frees up class time for hands-on work and discussion, and permits deep dives into the content. Students learn by doing and asking questions, and they can also help each other—a process that benefits a majority of learners (Strayer, 2012).

To accomplish this vision, there are several obstacles that must be surmounted. Following are four issues that need more research.

1. Quality of Assessments

The first hurdle relates to variability in the quality of assessments within current and future educational games and other digital learning environments. Because schools are under local control, students in a given state could engage in sundry games and learning environments during their educational years. Teachers, publishers, researchers, and others will be developing these digital environments, but with no standards in place (or a survival-of-the-fittest mechanism to weed out inadequate products), they will likely differ in curricular coverage, difficulty of the material, scenarios and formats used, and many other ways that will affect the adequacy of the digital environment, tasks, and inferences on knowledge and skill acquisition that can justifiably be made. More research is needed to figure out how to equate educational games or create common measurements (i.e., standardized) from diverse environments. Towards that end, there must be common models employed 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 differ (e.g., playing alone vs. playing with another student; playing with learning goals vs. playing with trolling goals vs. playing with entertainment goals).

2. Interpreting Different Learning Progressions

The second hurdle involves accurately capturing and making sense of students' learning progressions. That is, while educational games 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 that can be applied across different learning activities and contexts (Shavelson & Kurpius, 2012). However as Shavelson and Kurpius pointed out, there is no single absolute order of progression as learning in educational games 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 and assessments of learning progressions in educational games need to be actively researched and validated to realize their potential.

3. Expanded Educational Boundaries

The third problem to resolve involves impediments to moving toward the idea of new contexts of learning (e.g., flipped classrooms). 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, 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 and data plan that can meet the requirements of online activities. In addition, some critics 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., "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. Moreover, use of analytics and big data will require the development of new methods and tools beyond traditional empirical research—seamless and context based.

4. Privacy/Security

The fourth hurdle involves figuring out a way to resolve privacy, security, and ownership issues regarding students' information. The privacy/security issue relates to the accumulation and aggregation 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 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. And because of the often high-stakes consequences associated with tests, many parents and other stakeholders fear that the data collected could later be used against the students.

Despite these hurdles, as well as others not included, 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 (when the assessment only samples broadly, not deeply from the curriculum). Given the importance of time on task (i.e., engaged time) 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 (and more recently computer-based adaptive), high-stakes tests for students, this shift toward ensuring that students have acquired "essential" skills fits with the envisioned future of assessment.

What would it take to accomplish this vision of the future of educational assessment? The next sections describe a solution to the problem using stealth assessment within well-designed educational games.

Stealth Assessment

The Partnership for 21st Century Learning (P21) has developed a framework providing a blueprint for particular outcomes that students should achieve (2015). In addition to content knowledge like math and science, students need to develop essential skills, as mentioned earlier, such as creativity, problem-solving, and collaboration to succeed in the 21st century. The framework points out the importance of using technology-based assessments to measure and support these skills. Moreover, the Office of Educational Technology (2015) suggested using technology to develop assessment tools to measure the processes of learning, to provide formative feedback, and to systematically gather and analyze information about learning so that teachers and

schools can make full use of the data. Technology-based assessments can facilitate learning processes and outcomes by engaging students in learning environments.

Stealth assessment (see Shute, 2011; and for a related idea called situated seamless assessment, see Young, Kulikowich, & Barab, 1997) is such an assessment technology. It involves embedding assessment(s) directly and invisibly into an immersive learning environment (such as a well-designed educational game). This appears to be a viable solution to resolving the rift between learning and testing caused by standardized, traditional tests (Dori, 2003; Hickey & Zuiker, 2012). That is, stealth assessment is intended to be formative, ongoing, and dynamic (Shute & Ventura, 2013). It can also be used to support learning. For example, as a student progresses through an educational game in which the stealth assessment is embedded, the estimates of competency states provided by the stealth assessment can be used as the basis to provide targeted instructional support as well as other types of adjustments to the game which can prepare students' for future learning (Schwartz & Martin, 2004). In this way, stealth assessment can comprise an invisible but strong bond linking teaching and learning (Wilson & Sloane, 2000).

Using educational games as a vehicle for assessment and support of learning is justified by the literature (see Van Eck, Shute, & Rieber, in press). For instance, learning is at its best when it is active, goal oriented, contextualized, and interesting (e.g., Bransford, Brown, & Cocking, 2000; Bruner, 1961; Vygotsky, 1978). Instructional environments should thus be interactive, provide ongoing feedback, grab and sustain attention, and have appropriate and adaptive levels of challenge—i.e., the features of well-designed games. So a well-designed game can fulfill the achievement of learning objectives by providing incentives to motivate learners, engaging learners via four types of interaction (i.e., cognitive, affective, behavioral, and sociocultural), encouraging a player to learn from failure, and adapting to players' responses (Plass, Homer & Kinzer, 2015). Plass and his colleagues argued for "accurate and ongoing assessment" for learning in games (p. 266), which is what stealth assessment delivers. In short, stealth assessment is a powerful tool to: (1) enhance acquisition of both content knowledge and essential skills required in today's world, (2) integrate learning and assessment in a natural and seamless way, (3) enable valid and reliable inferences of competencies being targeted, and 4) reveal/visualize learning processes and outcomes—to the student for reflection, to the teacher for support, and to the system for further adjustments and

instructional support. The mechanism for ensuring the reliability and validity of a stealth assessment is evidence-centered design (Mislevy, Steinberg, & Almond, 2003).

Evidence-Centered Design

Evidence-centered design (ECD) is an assessment framework underpinning the development of assessment tasks. Its strength lies in basing competency estimates on a chain of evidence that is grounded in task performances. That is, ECD directly connects valid claims of competency states to particular performance data, thus ensuring the validity of the assessment (Mislevy & Haertel, 2006; Reese, Tabachnick & Kosko, 2015). There are several main models in ECD that work in concert: (1) competency model (CM), (2) evidence model (EM), and (3) task model (TM).

The competency model (CM) clarifies what needs to be assessed (i.e., knowledge, skills, and other attributes). It delineates the variables that characterize the targeted knowledge and skills and allows for the inference of students' levels on those competencies (see Almond & Mislevy, 1999). ECD is especially powerful in assessing multivariate competencies (although also suited for unidimensional constructs). The instantiation of the CM in an assessment situation creates the *student model*, a term that originated in the intelligent tutoring system literature (see Shute & Psotka, 1996). The student model is like a profile or report card of students' current knowledge and skill states (and trajectories), but can present estimates at a finer grain size than summative types of assessment.

The evidence model (EM) defines particular behaviors (or "indicators") that reveal the targeted competencies as well as the relationship(s) among those behaviors to the competency variables. That is, specific student behaviors (and the scoring thereof) constitute the *evidence rules*, while the statistical connections established between the behaviors and the CM variables constitute the *statistical model*. Evidence rules specify the identification and scoring of particular actions taken within the game, thus comprising weighted evidence. Statistical models set values to the specified evidence, accumulate the evidence (i.e., observable variables), then statistically link the observables to the competency variables (i.e., unobservables). The statistical model can employ simple dichotomous models (e.g., correct/incorrect; present/absent) but also graded models (e.g., low,

medium, high) used in Bayesian Networks (see Shute & Ventura, 2013). The EM entails ongoing accumulation of evidence, and continuous updating of the CM variables across tasks.

The task model (TM) specifies the features of tasks (e.g., difficulty level and format) that can elicit particular behaviors to be used as evidence. That is, the goal of the TM is to produce assessment tasks that are constructed explicitly to elicit evidence that is aligned with targeted competency variables. Overall, a TM contains a wide collection of tasks and task types (see Almond, Kim, Velasquez & Shute, 2014). The EM serves as the glue between the TM and CM. Together, the CM, EM, and TM form a dynamic system that is the backbone of stealth assessment's functionality.

How Does Stealth Assessment Work?

Stealth assessment, using ECD for its assessment design, aligns the embedded assessment tasks with targeted competencies. The main purpose of stealth assessment is to make valid inferences about competency levels based squarely on collected evidence. Using stealth assessment in educational games enables one to directly link actions/behaviors in the game to the targeted competencies without interrupting students' learning (Shute, 2011). Players interact with the tasks/levels in a game (the TM), and behaviors are captured in a log file and analyzed according to the scoring rules in the EM. Results of the scored observables are processed statistically in the EM then entered into the student model (i.e., the player's CM). As the interaction continues, the student model keeps receiving data and updating claims about competency levels in the form of probabilities reflecting real-time estimates of learners' competencies. The estimations can be used to adapt tasks to meet learners' current level (e.g., choosing suitable difficulty level, or providing prompts/hints). Such support during gameplay is important to engage learners (Shute & Wang, in press; Walkington, 2013), and to facilitate learning. A recent meta-analysis by Wouters and Oostendorp (2013) has shown that games using adaptivity show an effect size of .34.

In the next section, we illustrate the implementation of stealth assessment in a digital game to assess a particular competency—problem-solving skill.

Application of Stealth Assessment

In a recent study in our lab, we embedded a stealth assessment of problem solving skills in a popular game called Plants vs. Zombies 2 (PvZ2, Electronic Arts). PvZ2 presents players with situations where they must

select plants (with different characteristics) to protect their home base from being overrun by zombies (which also differ in terms of their characteristics). A player chooses appropriate plants to form both a defense as well as offense in the lawn, set up in a grid like a chess board. To guard against zombies making it through to the home base, players need to engage their problem-solving skills and come up with different strategies to deal with different situations. Problem-solving skill was thus the targeted competency (see Shute, Moore & Wang, 2015; Wang, Shute & Moore, 2015).

Establishing the competency model of problem-solving skills began with an extensive literature review. This resulted in the identification of four main facets: (1) analyze givens and constraints, (2) plan a solution pathway, (3) use tools effectively and efficiently, and (4) monitor and evaluate progress. To align the game levels with the targeted competency variables, we specified particular indicators (observables) to be used as evidence for problem-solving skills. Next, we iteratively evaluated each indicator in terms of its relevance to the CM as well as its feasibility of automatic collection and scoring from gameplay (Shute et al., 2015). This yielded a total of 32 indicators aligned to the four facets (see Table 9.1 for some examples). The first and second facets have 7 indicators each. The third contains 14 and the last includes 4.

After identifying indicators and creating scoring rules, we assigned statistical relationships between indicators and competency variables. Next, we categorized each indicator into discrete levels (e.g., poor, ok, good, very good). Bayes nets (BN) processed the data when an indicator was demonstrated, calculating the probability per level (e.g., low, medium, high) and per facet of problem-solving skill. We constructed a BN for each level in the game to ensure that each level had its own specific indicators and rubrics (not all indicators were applicable in each level).

For example, in PvZ2, iceberg lettuce is a defensive plant that can slow down the zombies by freezing them, while a snapdragon is an offensive plant that breathes fire to burn/kill zombies. When these two plants are placed next to each other, their powers cancel each other out (e.g., if an iceberg lettuce freezes a zombie, and a proximal snapdragon breathes fire on the zombie, it will reanimate the zombie). Placing these two types of plants close to one another (e.g., within 2 spaces) comprises evidence of ineffective tool use. For the scoring rule, we set the ratio as the number of iceberg lettuces planted in the range of a snapdragon, divided by the total number of iceberg lettuces planted (see indicator #37 in Figure 9.1). Smaller ratios are better in this case.

During gameplay, actions that players take are scored in real-time relative to specific indicators. The BNs are continuously updated regarding the current estimates of problem solving skill, overall and at the level of the facets. As seen in Figure 1, there is a probability of .61 that the player is currently estimated to be "low" on the "tool use" facet. Moreover, other variables change as well. Overall, BNs graphically portray the relationships between the main competency, its facets, and the associated indicators. BNs are used to ensure dynamic communication between the data and the beliefs of certain competencies (Reese et al., 2015).

To validate the stealth assessment in PvZ2, we selected two external measures of problem-solving skills: Raven's Progressive Matrices (RPM; Raven, 1941) and MicroDYN (Wüstenberg, Greiff, & Funke, 2012). The former tests learners' inductive ability (i.e., rule identification), while the latter requires the application of existing information to solve problems (rule application). The scores from the two external tests significantly correlated with the estimates from the stealth assessment ($p < .01$) (Shute et al., 2015). Thus, the stealth assessment embedded in the game shows both internal and external validity (convergent validity).

After the stealth assessment is embedded into the game and the states of player's competencies are assessed and validated, the next logical step is to use this information to provide adaptive learning support (Shute, Ke & Wang, in press) to the player. In this case, adaptive support can refer to personalized feedback (regarding the competency or game level), increasing the difficulty level based on estimates of the player's abilities, and so on.

Defining Adaptivity

Csikszentmihalyi (1997) claimed that learners learn best when they are fully engaged in some process, or in the state of flow. Inducing a state of flow involves the provision of clear and unambiguous goals, challenging yet achievable levels of difficulty, and immediate feedback (Cowley, Charles, Black, & Hickey, 2008; Csikszentmihalyi, 1997). Based on flow theory, a task that is too difficult can be frustrating while a task that is too easy may be boring, thus the optimal state (of flow) resides between the two. Similarly, Vygotsky's Zone of Proximal Development (ZPD; 1978) suggests that learning is at its best when the learning materials are just beyond students' existing level of understanding and ability (Vygotsky, 1978). Considering these two aspects of deep learning—facilitating the state of flow and providing materials compatible with learners' ZPDs—adaptive learning environments can be used to facilitate both via adapting to learners' current competency state(s).

Adaptivity generally refers to the ability of a person or device to alter its behavior according to changes in the environment (Shute & Zapata-Rivera, 2012). Some common examples of adaptive devices include thermostats and cruise control in many cars. In the context of instructional environments, adaptivity can help to provide personalized instruction for different learners with varying ZPDs and facilitate the state of flow throughout the learning process. An adaptive learning environment should monitor various (and often evolving) characteristics of learners then balance challenges and ability levels to improve learning (Shute & Zapata-Rivera, 2012).

Adaptivity in Educational Games

When people play well-designed games, they often lose track of time (i.e., experience the state of flow). Teachers try to engage students with learning materials, but the engagement is usually not comparable to that experienced with good video games (Gee, 2003; Gee, 2005; Prensky, 2001). Over the past couple of decades, there has been growing interest in designing and developing educational games as a way to fully engage students in learning, and also add learning opportunities into games (Kickmeier-Rust & Albert, 2010). Again, adaptive educational games can help maintain players' state of flow (Csikszentmihalyi, 1997; Vygotsky, 1978) and ultimately improve their learning (Andersen, 2012) by keeping players within their ZPD.

One way to include adaptivity in educational games is to use *micro-adaptation* (Kickmeier-Rust & Albert, 2010; Shute, Graf, & Hansen, 2005). This approach entails monitoring and interpreting the learner's particular behaviors, as with stealth assessment. Micro-adaptivity then may provide the learner with appropriate educational supports and/or adjust various aspects of the game (e.g., level difficulty) based on the student model estimates without disrupting the state of flow (Kickmeier-Rust & Albert, 2010). Adaptive games can adapt challenges to the current estimated levels of player's knowledge and skills (Csikszentmihalyi, 1997; Vygotsky, 1978) and provide formative feedback (Shute, 2008) and other types of support in unobtrusive ways (Peirce, Conlan, & Wade, 2008).

Putting It All Together

Currently, while the U.S. educational system is doing a good job producing test-takers, we really need to refocus our attention to producing creative and critical problem solvers. Towards this end, we need to consider using (a) new approaches (e.g., stealth assessment) for measuring 21st century competencies, as well as (b) adaptive technologies within learning environments (e.g., educational games) to promote these competencies.

To effectively adapt to students' current competency states, one needs to first accurately estimate their current state. Shute, Ke, and Wang (2017) describe nine steps towards designing and developing stealth assessment in interactive environments (e.g., educational games). They suggested that the next logical step is to use the information from the stealth assessment to adapt to the learners' ability to maintain their ZDP and the state of flow (Csikszentmihalyi, 1997; Vygotsky, 1978). With learning environments that can accurately measure students' skills and competencies and then adapt to their skill level, we can help students improve their learning processes and outcomes, which will better prepare them for future learning.

Conclusion & Future Directions

In this chapter, we illustrated some of the problems we have in our educational system relative to 21st century needs, described a future with few if any high stakes tests and with better teaching approaches than now exist, provided an example of embedding stealth assessment into a video game, and elaborated on how the information gained from stealth assessments can help to create learning environments such as educational games that can effectively adapt to students' abilities and other attributes.

This chapter is intended to ignite ideas and research streams that can help to move the vison towards reality. To achieve this goal, we suggest that researchers who want to tackle some of these problems: (1) focus on research questions that can address any of the four hurdles we face to accomplish the presented vision, (2) explore ways to make the output of stealth assessment more user friendly (e.g., a dashboard of students' states by which teachers can make relevant instructional support decisions), (3) examine the when, what, and how to adapt without disrupting the flow in educational games to achieve the outcomes we are looking for, and (4) investigate the questions of how to make the components of stealth assessment (i.e., CMs and EMs) recycled into other games in a plug-and-play manner.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into practice*, *42*(1), 18-29.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education policy analysis archives*, *10* (18), 1-74.

Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). How task features impact evidence from assessments embedded in simulations and games. *Measurement: Interdisciplinary Research and Perspectives, 12*(1-2), 1-33.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 223-237.
Andersen, E. (2012). Optimizing adaptivity in educational games. *Proceedings of the International Conference on the Foundations of Digital Games,* 279-281.
Bauer, K. N., Brusso, R. C., & Orvis, K. A. (2012). Using adaptive difficulty to optimize videogame-based training performance: The moderating role of personality. *Military Psychology, 24*(2), 148.
Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day.* International Society for Technology in Education.
Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*(3), 287-302.
Bransford, J., Brown, L. L., & Cocking, R. R. (2000). *How People Learn: Brain, Mind, Experience, and School (expanded edition)*, Washington: National Academies Press.
Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21-32.
Brusilovsky, P. (1998). Methods and techniques of adaptive hypermedia. *Adaptive hypertext and hypermedia* (pp. 1-43) Springer.
Culler, R. E., & Holahan, C. J. (1980). Test anxiety and academic performance: the effects of study-related behaviors. *Journal of educational psychology*, *72*(1), 16.
Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment (CIE), 6*(2), 20.
Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life.* New York: Basic Books.
Dori, Y. J. (2003). From nationwide standardized testing to school-based alternative embedded assessment in Israel: Students' performance in the matriculation 2000 project. *Journal of Research in Science Teaching, 40*(1), 34-52. doi:10.1002/tea.10059
Economist Intelligence Unit (2012). *The learning curve. Lessons in country performance in education*. London: Pearson. Retrieved from: http://thelearningcurve.pearson.com/the-report
Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE), 1*(1), 20-20.
Gee, J. P. (2005). *Why video games are good for your soul: Pleasure and learning.* Melbourne, Australia: Common Ground Publishing.
Guardian journal (2015). *Chinese nationals charged with cheating by impersonation on US college tests*. Retrieved from http://www.theguardian.com/us-news/2015/may/28/china-nationals-cheating-college-tests.
Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes. *Journal of Economic Growth*, *17* (4), 267-321. doi:10.1007/s10887-012-9081-x
Hickey, D., & Zuiker, S. (2012). Multilevel assessment for discourse, understanding, and achievement. *Journal of the Learning Sciences, 21*(4), 522-582. doi:10.1080/10508406.2011.652320
Jones, B. D. (2007). The unintended outcomes of high-stakes testing. *Journal of applied school psychology*, *23*(2), 65-86. doi: 10.1300/J370v23n02_05
Kickmeier-Rust, M. D., & Albert, D. (2010). Micro-adaptivity: Protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning, 26*(2), 95-105.
Madaus, G., & Russell, M. (2010). Paradoxes of high-stakes testing. *The Journal of Education*, *190*(1/2), 21-30.
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-initiativeinbloom-gates-and-carnegie.html
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, *1*(1), 3-62.
Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.
New Media Consortium. (2015). *The NMC Horizon Report: 2015 K-12 Education Edition,*
Office of Educational Technology (2015). *Assessment: Measure What Matters*. US department of education. Retrieved from http://tech.ed.gov/netp/assessment-measure-what matters/

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

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior, 24*(5), 2415-2433.

Partnership for 21st Century Learning (2015). *Framework for 21st-century learning*. Retrieved from http://www.p21.org/storage/documents/P21_framework_0515.pdf

Peirce, N., Conlan, O., & Wade, V. (2008). Adaptive educational games: Providing non-invasive personalised learning experiences. *Digital Games and Intelligent Toys Based Education, 2008 Second IEEE International Conference On,* 28-35.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist, 50*(4), 258-283. doi:10.1080/00461520.2015.1122533

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the Horizon, 9*(5), 1-6.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology, 19*(1), 137-150.

Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2015). Video game learning dynamics: Actionable measures of multidimensional learning trajectories. *British Journal of Educational Technology, 46*(1), 98-122. doi:10.1111/bjet.12128

Reigeluth, C. M., & Karnopp, J. R. (2013). *Reinventing schools: It's time to break the mold* R&L Education.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129.

Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. *Learning progressions in science* (pp. 13-26) Springer.

Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. *ETS Research Report Series, 2006*(1), 1-49.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

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

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

Shute, V. J., & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist, 38*(2), 105-114.

Shute, V. J., Ventura, M., Small, M., & Goldberg, B. (2013). Modeling student competencies in video games using stealth assessment. In R. Sottilare, X. Hu, A. Graesser & H. Holden (Eds.), *Design recommendations for adaptive intelligent tutoring systems: Learning modeling* (pp. 141-152). Washington, DC: Army Research Laboratory.

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for sighted and visually-disabled students. In L. PytlikZillig, R. Bruning & M. Bodvarsson (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 169-202). Greenwich, CT: Information Age Publishing.

Shute, V. J., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59-78). New York, NY: Springer.

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

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 570-600). New York, NY: Macmillan.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, Massachusetts: The MIT Press.

Shute, V. J. & Wang, L. (in press). Assessing and supporting hard-to-measure constructs. To appear A. Rupp, & J. Leighton (Eds.), *Handbook of cognition and assessment*. New York, NY: Springer.

Shute, V. J., & Zapata-Rivera, D. (2008). Adaptive technologies. *Handbook of Research on Educational Communications and Technology,* 277-294.

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. *Adaptive technologies for training and education* (pp. 7-27). New York: Cambridge University Press.

Snow, R. E. (1991). The concept of aptitude. In R. E. Snow, & D. E. Wiley (Eds.), *Improving inquiry in social science* (pp. 249-283). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning Environments Research, 15*(2), 171-193.

U.S. Department of Education, National Center for Education Statistics. (2015). *Fast Facts: Expenditure*. Retrieved August 25, 2005, from http://www.ed.gov/about/overview/fed/10facts/index.html

Vandewaetere, M., Desmet, P., & Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behavior, 27*(1), 118-130.

Van Eck, R. N., Shute, V. J. & Rieber, L. P. (in press). Leveling up: Game design research and practice for instructional designers. In R. Reiser & J. Dempsey (Eds.), *Trends and issues in instructional design and technology (4th ed.)*. Upper Saddle River, NJ: Pearson Education, Inc.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34-41.

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology, 105*(4), 932-945. doi:10.1037/a0031882

Wang, L., Shute, V., & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations*, *7*(4), 66-87. doi:10.4018/IJGCMS.2015100104

Wheelock A, Bebell D, Haney W (2000). What can student drawings tell us about high stakes testing in Massachusetts? *Teachers College Record*, November 2

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208. doi:10.1207/S15324818AME1302_4

Wouters, P., & van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education, 60*(1), 412-425. doi:10.1016/j.compedu.2012.07.018

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving — more than reasoning? *Intelligence. 40*(1), 1-14.

Young, M. F., Kulikowich, J. M., & Barab, S. A. (1997). The unit of analysis for situated assessment. *Instructional Science, 25*, 133-150.

Yuhas, A. (2015). Chinese nationals charged with cheating by impersonation on US college tests. *The Guardian*: http://www.theguardian.com/us-news/2015/may/28/china-nationals-cheating-college-tests.

CHAPTER 9 TABLES

Facet	Example Indicators
Analyzing Givens & Constraints	▪ Plants > 3 Sunflowers before the second wave of zombies arrives ▪ Selects plants off the conveyor belt before it becomes full
Planning a Solution Pathway	▪ Places sun producers in the back, offensive plants in the middle, and defensive plants up front ▪ Plants Twin Sunflowers or uses plant food on (Twin) Sunflowers in levels that require the production of X sun
Using Tools and Resources Effectively	▪ Uses plant food when there are > 5 zombies in the yard or zombies are getting close to the house (within 2 squares) ▪ Damages > 3 zombies when firing a Coconut Cannon
Monitoring and Evaluating Progress	▪ Shovels Sunflowers in the back and replaces them with offensive plants when the ratio of zombies to plants exceeds 2:1

Table 9.1. Competency model and examples of indicators (from Shute et al., 2015)

CHAPTER 9 FIGURES

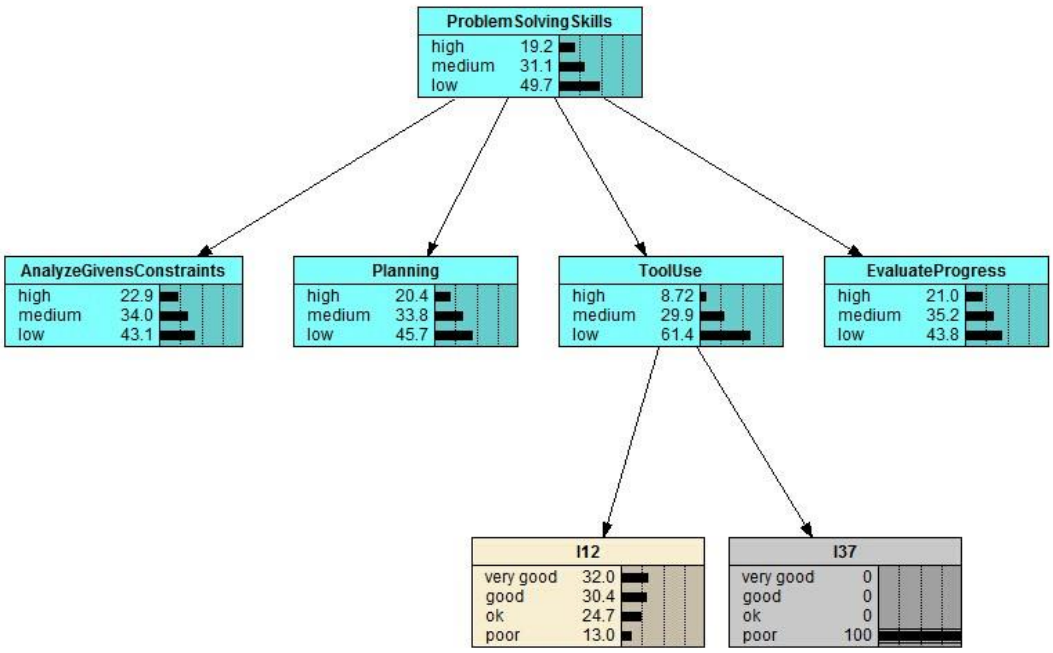

Figure 9.1. An example of a BN with data for indicator #37 entered (poor use of iceberg lettuces)