The History of Stealth Assessment and a Peek into Its Future

Valerie J. Shute, Florida State University, USA

ABSTRACT

Interactions with digital environments can yield huge amounts of electronic data—in terms of what was done, when, where, and how. What's known today as stealth assessment arose from a single question four decades ago: How can we mine digital data for meaningful information about a person's cognitive and noncognitive states, then use it to enhance learning? This query undergirded early work in intelligent tutoring systems but was handicapped by the technology of the time. This chapter describes the evolution of the concept and functioning of stealth assessment across the decades—how it emerged and where it may go in the future.

Keywords: Stealth Assessment, Evidence-Centered Design, Competency Model, Evidence Model, Task Model, Game-Based Learning, Game-Based Assessment

THE HISTORICAL ROOTS OF STEALTH ASSESSMENT

1984-1986—*Smithtown*

My first real foray into designing intelligent tutoring systems (ITSs; i.e., systems that could measure and respond to students' computer-based learning) came when I was a post-doctoral fellow at the Learning Research & Development Center (LRDC), University of Pittsburgh, working with Bob Glaser. My main project was to design, develop, and evaluate a microworld environment (also known as a guided-discovery world) to support students' learning of microeconomics (e.g., the laws of supply and demand) as well as scientific inquiry skills. I named the program "Smithtown" after Adam Smith, and it ran on state-of-the-art (at the time) AI systems—Xerox 1108/1186 Lisp machines. *Smithtown* was designed to be a highly interactive and intelligent simulation. Students posed what-would-happen-if questions and conducted experiments within the computer environment, testing and enriching their knowledge of functional relationships among variables by manipulating various economic factors. For example, in this hypothetical town, one could change the population, per capita income, quantity of resources, and so on, then see ramifications in the market on substitute and complementary goods and services.

There was no fixed curriculum. The student—not the system—generated problems and hypotheses. For instance, after generating a hypothesis (e.g., "Does increasing the price of coffee affect the demand for tea?"), the student could test it by executing a series of actions, such as changing the value of variables (e.g., two different prices of coffee) and observing the ensuing plots, relative to the demand for tea. This series of actions, or behaviors, for creating, executing, and following-up a given experiment, defined a student solution to their self-generated problem (for more on the resources available in the game, see Shute, Glaser, & Raghavan, 1989).

Despite having no curriculum, *Smithtown* did have the instructional goal of enhancing inquiry skills. These inquiry skills were determined from the results of an exploratory study by Shute and Glaser (1991) who identified more and less effective behaviors for interrogating and inducing information from a new domain. In the example above, making only a tiny change in the price of coffee would make it quite difficult to see any change in demand for the substitute good—e.g., tea. That particular action (i.e., making a very small price change) would provide evidence for a sub-optimal aspect of inquiry skills. In stealth assessment jargon, I would call all the facets comprising the targeted skills the "competency model" for inquiry skills. Information about each constituent competency variable was subsequently coded into rules that the system monitored in conjunction with a learner's actual behaviors. Thus, the system knew of sequences of good behaviors and

sequences of ineffective, or "buggy" behaviors. The *Smithtown* system would leave a student alone if he or she was performing adequately in the environment. But if the system determined that a student was floundering or demonstrating buggy behaviors, *Smithtown*'s coach would intervene, offering assistance on the specific problematic behaviors. For example, if a student persisted in changing many variables at one time without first collecting baseline data in the online notebook, the rule that would be invoked looked like the following:

> *If:* The student changes more than two variables at a time prior to collecting baseline data for a given market, and it is early in the session where the experiment number is less than four;
>
> *Then:* Increment the "Multiple Variable Changes" bug count by 1 and pass the list to the coach for possible assistance.

If this rule fired and the number of times it was invoked surpassed some threshold (e.g., four times), then the coach would appear and say (via text), "I see that you're changing several variables at the same time. A better strategy would be to enter a market, see what the data look like before any variables have been changed, then just change *one* variable while holding all the others constant."

In addition to the rules monitored by the system, we also developed a list of performance measures, called learning indicators, that enabled us to determine what type of actions or behaviors yielded better performance in this type of environment. We created a range of indicators, from low-level, simple counts of actions (e.g., total number of actions taken within *Smithtown*) to higher-level, complex behaviors (e.g., number of times a manipulation to an independent variable was made that showed an obvious change in the dependent variable). These indicators are shown in Appendix 1.

The system kept a detailed digital history of all student actions, grouping them into (i.e., interpreting them as) behaviors and solutions. Students' success in *Smithtown* was based on two main statistics: (a) the number of times the student demonstrated a buggy behavior (errors of commission), and (b) the ratio of the number of times the student used a good behavior (e.g., change one variable at a time) over the number of times it was applicable (errors of omission). Digital coaching was based on the heuristic of first advising about buggy behaviors, then advising on any persistent errors of omission. Advice was always given in the context of a particular experiment, so it was context sensitive. For example, the coach might say, "You haven't graphed any data yet and I think you should try it out. This is often a good way of seeing your data. It lets you plot variables together and some surprising relationships may become apparent." However, the coach was unobtrusive. That is, after advice was given, there was no further coaching for some time.

Smithtown also knew about variable relationships that constituted economics principles (such as "Price is inversely related to quantity demanded"). If a student used the system's hypothesis menu and stated this relationship (e.g., "As price increases, quantity demanded decreases"), the student would be congratulated and told the name of the law just discovered (e.g., "Congratulations! You have just discovered what economists refer to as the Law of Demand").

I hypothesized that such computer-based instruction on applying effective inquiry skills would ultimately lead to the acquisition of specific subject matter—in this case, microeconomics. Several papers describe the results of experiments evaluating *Smithtown.* For example, in one study (Shute, Glaser, & Raghavan, 1989) with 30 undergraduate students, we demonstrated that *Smithtown* fared very well when compared to traditional classroom instruction on economics. In short, we found that the two groups did not significantly differ on their posttest scores after holding pretest scores constant although the economics classroom group received more than twice as much instruction on the subject matter compared to the *Smithtown* group (i.e., about 11 hours vs. 5 hours, respectively). Moreover, the *Smithtown* system did *not* tutor economic knowledge directly.

Another experiment (Shute & Glaser, 1990) extended these findings using 530 Air Force recruits interacting with *Smithtown*. Learning indicators relating to hypothesis generation and testing (from Appendix 1, the indicators related to Use of Evidence, Effective Generalization, and Effective Experimental Behaviors) were the most predictive of successful learning of economic concepts in *Smithtown*, accounting for more of the variance in the learning criterion than did a measure of general intelligence.

Some questions remained after we completed several studies using *Smithtown*. Here are two that remain unanswered:

- (1) *How much flexibility of behaviors should be allowed before the coach intervenes?* Because there was not just one way to use the tools optimally to conduct experiments to induce regularities, what should the criteria be for saying that a given student is floundering or behaving unsystematically? In a largely discovery environment like *Smithtown*, how do we disambiguate learners who are not active because of cognitive shortcomings versus personality or other non-cognitive attributes?
- (2) *How generalizable are these findings regarding the delineated "effective" and "less effective" inquiry skills?* Does training on these skills transfer to successful performance in different domains?

In summary, almost four decades ago I began my work with *Smithtown*, and some of the fruits of that labor are present in today's version of stealth assessment (e.g., conceptualizing relevant competency models at the outset of the design, and engaging in real-time measurement of behaviors or "learning indicators" with associated instruction delivered in an unobtrusive manner). However, I did not, at the time, have any kind of systematic framework in which to develop and deploy all necessary models related to targeted competencies, evidence, and tasks, nor the computational and statistical machinery to quickly accumulate and make sense of this deluge of information in real time. That wouldn't come until a decade later—when I went to work at ETS in 2001 and learned about evidence centered design (e.g., Mislevy, Steinberg, & Almond, 2003). In the meanwhile, during the 1990s, I focused on developing other intelligent tutoring systems—primarily to use as vehicles to research important issues related to *enhancing learning for all*—which was and remains my main goal.

1986-2000—Intelligent Tutoring Systems and Associated Tools

The next stop on my journey towards stealth assessment technology was a research position at the Air Force Human Research Laboratory, where I worked for 13 years. While there, I developed a number of ITSs measuring and supporting various content, such as: Pascal programming skills (e.g., Shute, 1991), principles of electricity (i.e., Ohm's and Kirchhoff's laws; see Shute, 1993), flight engineering, (e.g., Shute & Gawlick, 1995), and statistics—descriptive statistics and probability (e.g., Shute, Gawlick-Grendell, Young, & Burnham, 1996). Again, my goal was to use these ITSs as research vehicles to test the best ways to optimize learning. I paid close attention to individual differences—cognitive, affective, personality, demographic, and so on—and used those variables to either control for or to adapt instruction. Next, I briefly provide one example of an ITS study, then detail work done with a second ITS – "Stat Lady" – which included several of the fledgling technologies used in today's stealth assessment.

Electricity Tutor. In one experiment (see Shute, 1993 for details), I created two versions of an electricity ITS. The computer-based environment was interactive with the learner actively involved in taking measurements (e.g., voltage, current) on various circuits (series and parallel) using online tools such as voltmeters and ammeters. The two versions of the ITS were identical, differing only in the feedback provided to the learner. That is, after working on a problem, whether correctly or incorrectly answered, one environment directly stated the relevant principle, and the learner applied it in the solution of related problems. The other environment required the learner to induce the relevant principle, providing only the variables involved in the rule but not their relationship(s). Both versions of the ITS (rule applied vs. rule induced) served as a complex

but controlled learning environment where a person learned 26 principles of electricity relative to Ohm's and Kirchhoff's laws.

The study allowed me to investigate the relationship between different types of learners (based on actions taken in the environment—specifically explorers vs. non-explorers) and two environments—rule-application vs. rule induction. I had predicted and found a disordinal interaction. That is, findings showed that when a learner type was matched to an ITS version (i.e., explorers assigned to the rule-induction condition and non-explorers assigned to the rule-application condition), learning was superior compared with mismatched learner types by condition. The findings generally supported the idea that personalizing instruction can optimize learning—one of the cornerstones of stealth assessment.

Stat Lady. After developing and testing various ITSs for several years, one thing that struck me was that while students did learn content, the ITS environments were rather tedious. However, we know that increased engagement results in increased learning (e.g., Lee, 2014). This prompted me to design and develop *Stat Lady* to teach statistics. I created a series of *Stat Lady* tutors, one for learning probability (Shute, Gawlick-Grendell, Young, & Burnham, 1996), and one for descriptive statistics (Shute, 1995; Shute & Catrambone, 1996). To make the ITS more engaging, the main character (*Stat Lady*) was patterned after Dana Carvey who played "church lady" on Saturday Night Live back in the late 1980s/early 1990s. She was very irreverent (saying things like "Isn't that special!" when a student messed up) which was an intentional design decision to keep things surprising and interesting.

As with most of my computer-based instruction work, the design of *Stat Lady* reflected the theoretical postulates that learning is a constructive process, enhanced by experiential involvement with the subject matter that is situated in real-world examples and problems. The *Stat Lady* tutor I focus on for this chapter instructed descriptive statistics and existed in two forms: with and without a student model named SMART, for Student Modeling Approach for Responsive Tutoring (Shute, 1995). The two versions of the tutor were the same, except for the presence or absence of the student model, and they shared a lot of features, including: (a) a humorous and experiential interface; (b) an identical set of curriculum elements, or CEs—there were 77 CEs altogether (see Shute, 1995 for a full listing); (c) a pool of playful problem sets, per CE; and (d) a three-level feedback design. Nevertheless, the versions differed in two important respects. First, in the non-SMART version, learners had to complete all 77 CEs (i.e., *Stat Lady* problems); but in the SMART version, learners only had to complete CEs that were identified by the system as being not fully mastered, based on pretest and performance data. Second, in the non-SMART version, learners had to solve at least two (out of five) CE-related problems before mastery was presumed and they could proceed to the next CE. In the SMART version, however, learners may have solved as few as zero (when the corresponding pretest items suggested mastery) but as many as five or more problems to demonstrate mastery of a given CE. Following performance assessment on a particular problem, learners were either advanced to a new CE, received remedial instruction, or continued to receive practice on the current CE.

In general, SMART operated as follows: (a) Each CE was initially evaluated following completion of the online pretest, and these initial knowledge/skill estimates were passed to the student model for initialization; (b) Each CE was subsequently evaluated during problem solution within the tutor; (c) CEs were linked to other CEs in inheritance hierarchies (i.e., parent, siblings, children) to provide student model updating information based on related CE values; and (d) because each CE knew its exact location in the tutor (i.e., where it was instructed and evaluated), remediation was precise and efficient. For a full listing of CEs and an example of a hierarchy, see Appendix A in Shute (1995). These activities mapped onto the four main routines that drove the student model, whereby the data were managed by a straightforward array of records, and each array element mapped onto an individual CE. All information for that CE was maintained within the log file, such as its outcome type and location in the tutor where it was instructed and assessed.

Once the student model was initialized with the pretest values, a learner was placed in *Stat Lady*'s curriculum at a CE that they did not know or only partly knew. The mastery criterion could be set at the outset of instruction to be greater than some value (e.g., > 0.83). Then, any CEs with values falling below this threshold would become candidates for instruction. In *Stat Lady*, learning about a particular CE involved instruction, and then relevant problem solving. Following the introduction of the CE content, a learner could solve a problem without any assistance from the tutor (i.e., on the first try). This is called level-0 assistance. Alternatively, the learner may have required various degrees of assistance (i.e., level-1, level-2, or level-3), which was provided to the learner in response to erroneous inputs, not explicitly requested. The simple presumption was that the more help required by the learner, the less they understood the current CE; and hence, the lower the associated probable degree of mastery (or p(CE) value). After a problem was completed, the appropriate CEs within the student model were updated.

Throughout the SMART version of *Stat Lady*, CEs with values below a pre-set mastery criterion were instructed, evaluated, and remediated, if necessary. The diagnostic part of the student model was driven by a series of regression equations based on the level of assistance the computer had to give each person, per CE. Remediation on a given element occurred when a learner failed to achieve mastery during assessment. Again, remediation was precise because each element knew its location within the tutor where it was instructed and assessed.

I examined the following issues in the SMART and non-SMART versions of *Stat Lady* (for more details on the architecture of SMART and additional results, see Shute, 1995): (a) predictive validity, (b) individual differences in learning from *Stat Lady*, and (c) effects of condition (SMART vs. non-SMART versions) on learning. Predictive validity related to how well the computed student model values, p(CE)s, predicted outcome scores. To test this relationship, I computed a stepwise multiple regression analysis with Posttest score as the dependent variable and the following independent variables: Pretest score, p(CE) data, aptitude factor score, education (years of school), and gender (male or female). Results showed that the first (i.e., strongest) variable to enter the equation was p(CE), with a multiple $R = 0.73$ (i.e., 54% of the unique outcome variance was explained by this variable alone). Next to enter the equation was aptitude, increasing the multiple R to 0.81 (accounting for an additional 11% of unique outcome variance). On the third and final step, pretest data entered the equation, accounting for an additional 4% variance, and increasing the multiple R to 0.82. None of the other variables reached the criterion for inclusion in the equation. The finding that education and gender failed to predict outcome performance suggested that the system did its job quite well, reducing the impact of influential sources of individual differences in learning. The last question examined learning gains between participants in one of two versions of the tutor: with and without SMART enabled. Findings showed that learners in the non-SMART version showed impressive learning outcome scores (i.e., 2 standard deviation pretest-to-posttest improvement). Their final posttest scores were 74.9%, on average. Learners in the SMART version, however, showed even higher gain scores—their average posttest scores were 82.1%. An analysis of covariance was computed on the posttest data with pretest as a covariate and version as a between-subjects variable. Results showed that there was a significant difference in learning outcome due to version: $F(1, 199) = 4.16$; $p < .05$, with superior outcome performance evidenced by participants in the SMART-enabled condition (Shute, 1995).

The last thing to mention regarding SMART is that while I originally used regression equations to estimate students' current competency states per CE, I also tested an alternative approach using a Bayesian network. That is, at any given time, there is a given p(mastery) for a particular CE. For each of the four possible outcomes (i.e., levels of help) per CE, the ratio of likelihoods of that outcome given mastery vs. non-mastery may be defined. For example, we may postulate that level-0 behavior is 5 times as likely given mastery than given non-mastery, and that the ratios are 2, 0.5, and 0.2 for level-l, level-2, and level-3 actions, respectively. We can then view each behavioral outcome for a problem during the tutoring phase as an observation, applying Bayes' Rule to update p(mastery). The Bayes net approach to deriving the curves has the advantages that the form of the curves follows from some theoretical assumptions, and each curve is characterized by a

single, interpretable parameter (the likelihood ratio). In a post-hoc analysis of these curves' validity (relative to the original ones) using data from the SMART study described above, I recomputed students' data (n = 104) using the new values to compute student model values. Results showed that the values arising from the new curves, alone, can account for even more of the unique outcome variance compared to the original curves (i.e., new curves $R^2 = 0.70$ while original curves $R^2 = 0.54$).

In summary, throughout the 1990s, I focused on identifying factors that enhanced learning by creating and testing various ITSs, creating different conditions within each system. But the components for developing the content (competency model), measuring performance, making inferences about competency states, providing instructional support, and so on were still separate entities—i.e., with some communication but little theoretical glue to bind them. At the beginning of the new century, I went to work at Educational Testing Service (ETS) and was fortunate to find a perfect framework in which to integrate all the individual components on which I had been working. This stitching together of the assessment and instructional elements into a coherent framework was a necessary next step in the development of stealth assessment.

2001-2008—Evidence Centered Design (ECD), ACED, Formative Feedback

So just as I was in search of a framework for the emerging stealth assessment technology, it turns out that Bob Mislevy, Linda Steinberg, and Russell Almond had just begun publishing on "evidence-centered assessment design" (or ECD for short). After reading all the available publications on ECD (particularly Mislevy, Steinberg, & Almond, 2003), I'd finally found a valid framework in which I could pull my components together to "model and support" students' learning.

Evidence-Centered Design

Evidence-centered design (ECD) is an assessment framework underpinning the development of valid assessment tasks. Its strength lies in basing competency estimates on a chain of evidence that is grounded in task performance. That is, ECD directly connects valid claims of competency states to learner performance data, thus ensuring the validity of the assessment (Mislevy & Haertel, 2006). 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 learners' competency levels on specific competency variables (see Almond & Mislevy, 1999). 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; discussed earlier). The student model, then, is like a profile or report card of each learner's current knowledge and skill states (and trajectories), presenting estimates at a finer grain size than summative types of assessment.

The evidence model (EM) defines specific behaviors (or "learning indicators") that reveal the targeted competencies as well as the relationship(s) among those behaviors to the competency variables. That is, 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 actions taken within the given activity, thus comprising weighted evidence. Statistical models set values to the specified evidence, accumulate the evidence (i.e., observable variables), then statistically link the evidence (observables) to the competency variables (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 two important processes: evidence identification (EI, identifying observables using the rules of evidence from the log data) and evidence accumulation (EA, accumulating the observables generated from the EI process using a statistical model and updating the student model). For more on the EM, see Almond, Shute, Tingir, and Rahimi (2020).

The task model (TM) specifies the features of tasks (e.g., difficulty level and format) that can elicit the behaviors to be used as evidence. That is, the goal of the TM is to produce assessment tasks (which can be game levels) that are constructed explicitly to elicit evidence that is aligned with targeted competency variables. Overall, a TM contains a wide collection of task types which provide the basis for developing specific tasks (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. An example follows.

ACED

My first ECD-based assessment-for-learning system was called ACED (Adaptive Content with Evidence-based Diagnosis; see Shute, Hansen, & Almond, 2008). This adaptive assessment for learning system, with its different types of feedback, represented a move towards creating an engaging environment for learning—in this case, about geometric sequences. For example, one problem asked the following: "*Emily receives an email message which states that she'll have a "very lucky day" if she sends it out within one hour to exactly 3 people who, in turn, send it out to exactly 3 people, and so on. Emily forwards the email and everyone she sends it to participates in the chain mail. How many emails would be sent at the 4th hour? Enter your answer here*: _____." If a student entered an incorrect answer (e.g., "27") to the question above, the elaborated feedback would indicate the following, "*That's not correct. Three times as many emails go out every hour. That means 3 emails go out in the first hour, 9 go out in the second hour, and 27 emails (your response) go out in the third hour. The question asks about the number of emails in the fourth hour, which would be* $3 \times 3 \times 3 \times 3 = 81$." Thus, with elaborated feedback, the accuracy of the answer is indicated, along with an explanation about how to solve the problem and the correct answer. For the verification feedback condition, the student was simply informed if their answer was correct or incorrect.

ACED allowed us to test the effects on learning of (a) linear vs. adaptive sequencing of items, as well as (b) elaborated vs. verification feedback. The key issue we examined was whether the inclusion of the feedback into the system (a) impaired the quality of the assessment (relative to validity, reliability, and efficiency), and (b) enhanced student learning. Results from a controlled evaluation testing 268 high-school students across a 2-hour period showed that the quality of the assessment (i.e., validity and reliability) was unimpaired by the provision of feedback. Moreover, students using the various versions of ACED system showed significantly greater learning of the content compared with a control group (i.e., "business as usual"). These findings suggest that assessments in other settings (e.g., state-mandated tests) might be augmented to support student learning with instructional feedback without jeopardizing the primary purpose of the assessment. In short, we found that while there was no main effect of adaptivity (i.e., differential sequencing of tasks) on learning, there was a significant effect of type of feedback provided, with a strong advantage for those learning with the elaborated feedback. These findings were important for stealth assessment which uses formative feedback to bolster learning.

Game-Based Learning

At this point (2007) I had moved from ETS to Florida State University and had aggregated enough of the basic elements and framework to move forward with conceptualizing stealth assessment (Shute, Ventura, Bauer, & Zapata-Rivera, 2009). In that paper, we used a popular videogame called *Elder Scrolls IV: Oblivion* as a case study to measure creative problem-solving skills based on actions taken in the game. The shift to using games as a vehicle for learning was a no-brainer. That is, many children are not particularly motivated by school (classroom learning) but are highly motivated by playing games. So, the idea was to combine school material with games to increase learning, especially for lower performing, disengaged students. In the Shute et al. (2009) paper, we illustrated how games and assessment may be effectively conjoined by employing (1) ECD, and (2) Bayesian networks (e.g., Pearl, 1988) to monitor and support learning in the context of gaming environments. This was the first published use of the term "stealth assessment" although I had previously

used the term during earlier conference presentations—AERA and Serious Games (Shute & Underwood, 2005; Shute, Ventura, & Bauer, 2007).

2009-2022—Stealth Assessment Is Born and Tested, Tested, Tested

What Is Stealth Assessment?

Generally, stealth assessment represents an evidence-based approach that discreetly assesses students' learning progression while they are engaged with highly interactive and immersive environments, like digital games, then provides assistance, as needed. Stealth assessment aims to blur the boundaries between gameplay, learning, and assessment using unobtrusive methods (e.g., log files and eye tracking) to continually collect student data and examine their progression of cognitive and non-cognitive variables throughout the game. In such cases, assessment is part of gameplay (Mayer, 2018).

Ongoing advances in educational and psychological measurement allow more accurate estimation of student competencies than before (e.g., the 1990s work with ITSs described earlier), enabling us to extract ongoing, multifaceted information from a learner, and react in immediate and helpful ways, as needed. When embedded assessments are so seamlessly woven into the fabric of the learning environment that they are virtually invisible, it's stealth assessment. It is accomplished via automated scoring and machine-based reasoning techniques to infer things that would be too hard for humans (e.g., estimating values of evidence-based competencies in real time, across a network of skills).

How Does Stealth Assessment Work?

Stealth assessment, using ECD for its design of the assessment models, aligns assessment tasks (e.g., a game level) with targeted competencies. Using stealth assessment in educational games enables one to directly link actions/behaviors in the game to the targeted competencies in a very engaging way 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 version of the 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 the most suitable difficulty level) or to provide prompts, hints, or more elaborate learning supports like short videos. Such support during gameplay is important to engage learners and to facilitate learning.

Funding Support and Evolution of Stealth Assessment

After conceptualizing all the parts necessary to create stealth assessment, I received my first funding support from the Gates Foundation (2011-2013) to develop a proof-of-concept. As part of this effort, I decided to start from scratch—i.e., to build a game that would house the stealth assessment machinery (the game was originally called "Newton's Playground" and later changed to "Physics Playground"). Rather than just building one stealth assessment in the game to measure learning of Newtonian physics, I decided to build out *three* stealth assessments to run concurrently in the game to measure physics understanding, as well as creativity and conscientiousness (see Shute & Ventura, 2013). Note that these early stealth assessments were built as purely measurement tools. They were analyzed relative to their reliability and validity, and later for any learning that arose from gameplay. The results from an early experimental study (reported in Shute, Ventura, & Kim, 2013) showed that the psychometric qualities of the stealth assessments were good and learning physics did occur just from gameplay.

Encouraged by the measurement success of the stealth assessments, we received additional Gates funding (2012-2014) to model the relationships among affective states, engagement, and conscientiousness as they influence learning in our game. In Shute et al. (2015), we reported the results of a study that investigated the relationships among incoming knowledge, persistence, affective states, in-game progress, and learning outcomes for students using *Physics Playground*. We used structural equation modeling to examine these relations and obtained a model with good fit to the data. We found evidence that both the pretest and the in-game (stealth) measures of student performance (e.g., number of levels solved) significantly predicted learning outcome, while the in-game measure of performance was predicted by pretest data, frustration, and engaged concentration. Moreover, we found evidence for two indirect paths from engaged concentration and frustration to learning, via the in-game progress measure. The paper concluded with a call for the design of effective learning supports within game environments.

Before we began work designing and developing effective (and unobtrusive) learning supports—the formative feedback part of stealth assessment—we spent some time examining whether the stealth assessment technology could be employed in commercial games. We received some MacArthur foundation funding to test out stealth assessment in the popular game *Portal 2* (see Shute, Ventura, & Ke, 2015). Results revealed that participants who were assigned to play *Portal 2* showed a statistically significant advantage over those assigned to play *Lumosity* (another popular game that claims to support cognitive skills) on each of the three composite measures—problem solving, spatial skill, and persistence.

Around the same time, we also received funding from the Glasslab to test out the inclusion of stealth assessment of problem-solving skills in another popular game (i.e., a slightly modified version of *Plants vs. Zombies 2* called Use Your Brainz; see Shute, Wang, Greiff, Zhao, & Moore, 2016). We began by developing a problem-solving competency model based on a review of the relevant literature. We then identified in-game indicators that would provide evidence about students' levels on various problem-solving facets. Our problem-solving model was then implemented in the game via Bayesian networks. To validate the stealth assessment, we collected data from students who played the game-based assessment for three hours and completed two external problem-solving measures (i.e., *Raven's Progressive Matrices* and *MicroDYN*). Results indicated that the problem-solving estimates derived from the game significantly correlated with the external measures, which suggested that our stealth assessment was valid. These successful insertions of stealth assessment into existing games, in addition to its inclusion in new games being built from scratch suggests that it's an extensible technology.

From about 2016 to the present, we have received numerous NSF and IES grants that have enabled us to examine several remaining questions/issues more deeply. For example, we wanted to further test the effects of adaptivity on learning in the game *Physics Playground* (Shute, Almond, & Rahimi, 2019), compared to linear and free choice versions of the game. According to the findings reported in Shute et al. (2020)—the three versions did not differ in outcome; however again, like we found with the ACED research reported earlier, feedback (learning supports) significantly influenced learning but adaptivity of the content did not matter.

More recently, we tested the design of many different types of learning supports—for usability and in support of knowledge and skill acquisition (Kuba et al., 2021). We also examined the effects of a new game-based incentive system on performance and outcomes (Rahimi et al., 2021). We are currently testing the value-added of affective supports over our cognitive supports, as well as the value-added of model-based delivery of supports over just a fixed-delivery schedule. We are also modeling collaborative problem solving in *Physics Playground* (Sun et al., 2020; Sun et al., 2022). Throughout this stealth assessment journey, we are continuing to make all our software freely available to educators and researchers alike. So, what does the future look like regarding stealth assessment?

PEEK INTO THE FUTURE OF STEALTH ASSESSMENT

Let me start with my vision, originally presented in Shute, Leighton, Jang, and Chu (2017). 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 of 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 just administering assessments more frequently (e.g., each week, each day) but rather continually collecting data as students interact with digital environments, both inside and outside of school.

Currently, the science of assessment and its core principles are grounded in paper-based, discrete knowledge-based items, most of which are divorced from the performance-based learning environments that make up the contexts that students encounter today. My view of the future of educational assessment, writ large, includes designing assessments that begin by asking: (1) What are we trying to measure? (2) Is this important to kids' future success? and (3) What does the student need to do/say to provide evidence of this knowledge/skill/other attribute? That's exactly where stealth assessment comes into play—developed specifically to pull out relevant information from observed student behaviors (as well as biometric data) to make inferences about various competencies and their associated facets for diagnoses and support.

In addition to measuring domain knowledge and skills (e.g., Calculus; see Smith, Shute, & Muenzenberger, 2019), stealth assessment can measure hard-to-measure constructs such as creativity (e.g., Shute & Rahimi, 2021), problem solving (e.g., Shute, Wang, Greiff, Zhao, & Moore, 2016), systems thinking (Shute, Masduki, & Donmez, 2010), persistence (Ventura, Shute, & Small, 2014), and so on. Moreover, assessments need to measure learning processes, and make inferences on outcomes via evidence identification (to automatically score actions) and evidence accumulation (to accumulate evidence using a statistical or AI-based model).

In conclusion, I can easily see a world without traditional tests, like the prevalent multiple-choice tests, delivered by paper and pencil or computer. Such tests are limited regarding the type of knowledge that can be assessed as well as the duration of the assessment. Furthermore, such old-school tests often produce anxiety in test takers, which can negatively impact outcomes. We still need to ensure that all new assessments are reliable, valid, and fairly measure the targeted outcomes. Stealth assessment is intended to do just that—provide accurate measurement and learning support for all.

I can also envision use of general CMs that are applicable across different contexts. We have done some preliminary testing of this idea measuring collaborative problem solving in two different games—*Physics Playground* and *Minecraft* using the same CM (see Sun et al., 2020). And while most assessment design frameworks (like ECD) are theoretical, I see the benefits of combining such top-down, theory-based approaches with exploratory, bottom-up approaches (e.g., educational data mining; machine learning) to identify even more relevant indicators to provide evidence for various outcomes.

Despite the various attractive features of using stealth assessment in games to support learning in school and at home, there are some hurdles to surmount before it can become mainstream. Assessment design frameworks, like ECD, represent a design methodology but not a panacea, so more research is needed to figure out how to create common measurements from diverse environments. That is, it's 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). Another hurdle involves figuring out a way to resolve privacy, security, and ownership issues regarding students' information. The privacy/security issue relates to the accumulation of student data from disparate sources. 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 being aware 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.

In any case, constructing the envisioned ubiquitous and unobtrusive stealth assessments across multiple learner dimensions, with data accessible by diverse stakeholders, could yield 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 larger 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. Traditional assessment practices in school can lead to assessing students in a manner that may 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 based on acquired competencies.

It's time to derive and deploy new methods, like stealth assessment, to measure and support learning. This has become possible given the increased availability of computer technologies that make it easy to capture the results of routine student work—in class, at home, or wherever. I imagine that 10-20 years from now, technology will be so well integrated into students' day-to-day lives that they are unaware of its presence and learning and assessment are finally, effectively blurred.

REFERENCES

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. doi:10.1080/15366367.2014.910060 PMID:30344456

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 223–237. doi:10.1177/01466219922031347

Almond, R. G., Shute, V. J., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in game-based assessments. In R. Lissitz & H. Jiao (Eds.), *Innovative psychometric modeling and methods* (pp. 163–192). Information Age Publishing.

Elder Scrolls lV: Oblivion. [Digital game]. (2006). Bethesda Softworks/ZeniMax Media.

Kuba, R., Rahimi, S., Smith, G., Shute, V. J., & Dai, C.-P. (2021). Using the first principles of instruction and multimedia learning principles to design and develop in-game learning support videos. *Educational Technology Research and Development*, *69*(2), 1201–1220. doi:10.100711423-021-09994-3

Lee, J. S. (2014). The relationship between student engagement and academic performance: Is it a myth or reality? *The Journal of Educational Research*, *107*(3), 177–185. doi:10.1080/00220671.2013.807491

Mayer, R. E. (2018). Educational Psychology's past and future contributions to the science of learning, science of instruction, and science of assessment. *Journal of Educational Psychology*, *110*(2), 174–179. doi:10.1037/edu0000195

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. doi:10.1207/S15366359MEA0101_02

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Kaufmann.

Rahimi, S., Shute, V. J., Kuba, R., Dai, C.-P., Yang, X., Smith, G., Alonso Fernández, C., & Fulwider, G. (2021). The use and effects of incentive systems on learning and performance in educational games. *Computers & Education*, *165*, 104135. doi:10.1016/j.compedu.2021.104135

Shute, V. J. (1991). Who is likely to acquire programming skills? *Journal of Educational Computing Research*, *7*(1), 1–24. doi:10.2190/VQJD-T1YD-5WVB-RYPJ

Shute, V. J. (1993). A comparison of learning environments: All that glitters.... In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 47–74). Lawrence Erlbaum Associates.

Shute, V. J. (1995). SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction*, *5*(1), 1–44. doi:10.1007/BF01101800

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. doi:10.3102/0034654307313795

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). Information Age Publishers.

Shute, V. J., Almond, R. G., & Rahimi, S. (2019). Physics playground (Version 1.3) [Computer software]. Retrieved from https:// pluto.coe.fsu.edu/ppteam/pp-links/

Shute, V. J., & Catrambone, R. (1996). Unified vs. tailored analogies: Effects on conceptual knowledge acquisition. In *Proceedings of the 1996 international conference on Learning sciences* (pp. 502–507). AACE.

Shute, V. J., D'Mello, S. K., Baker, R., Bosch, N., Ocumpaugh, J., Ventura, M., & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education*, *86*, 224–235. doi:10.1016/j.compedu.2015.08.001

Shute, V. J., & Gawlick, L. A. (1995). Practice effects on skill acquisition, learning outcome, retention, and sensitivity to relearning. *Human Factors*, *37*(4), 781–803. doi:10.1518/001872095778995553

Shute, V. J., Gawlick-Grendell, L. A., Young, R. K., & Burnham, C. A. (1996). An experiential system for learning probability: Stat Lady description and evaluation. *Instructional Science*, *24*(1), 25–46. doi:10.1007/BF00156002

Shute, V. J., & Glaser, R. (1990). Large-scale evaluation of an intelligent tutoring system: Smithtown. *Interactive Learning Environments*, *1*(1), 51–76. doi:10.1080/1049482900010104

Shute, V. J., & Glaser, R. (1991). An intelligent tutoring system for exploring principles of economics. In R. E. Snow & D. Wiley (Eds.), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach* (pp. 333–366). Lawrence Erlbaum Associates.

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and Individual Differences* (pp. 279–326). W.H. Freeman.

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

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*(2), 137–161.

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). Macmillan.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics educational game. *Computers in Human Behavior*, *116*, 1–13. doi:10.1016/j.chb.2020.106647

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kamikabeya, R., Liu, Z., Yang, X., & Sun, C. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in Physics Playground. *Journal of Computer Assisted Learning*, *37*(1), 127–141. doi:10.1111/jcal.12473

Shute, V. J., Torreano, L., & Willis, R. (2000). Towards an automated knowledge elicitation and organization tool. In S. P. Lajoie (Ed.), *Computers as cognitive tools* (Vol. 2, pp. 309–335). Lawrence Erlbaum Associates.

Shute, V. J., Torreano, L. A., & Willis, R. E. (1998). DNA – Uncorking the bottleneck in knowledge elicitation and organization. In B. P. Goettl, H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.), *Proceedings of the 4th International Conference on Intelligent Tutoring Systems* (pp. 146–155). Springer-Verlag. 10.1007/3-540-68716-5_20

Shute, V. J., Torreano, L. A., & Willis, R. E. (1999). Exploratory test of an automated knowledge elicitation and organization tool. *International Journal of AI and Education*, *10*(3–4), 365–384.

Shute, V. J., & Underwood, J. S. (2005, April). *Diagnostic assessment in math problem in solving evidence is key*. Paper presented at AERA, American Educational Research Association, Montreal, Canada.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. The MIT Press. doi:10.7551/mitpress/9589.001.0001

Shute, V. J., Ventura, M., & Bauer, M. I. (2007, May). *Melding the power of serious games and stealth assessments to foster learning: Flow and grow*. Paper presented at the Conference on Serious Games: Learning, Development and Change, USC Annenberg School for Communication, Los Angeles, CA.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and floster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Routledge, Taylor and Francis.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education*, *80*, 58–67. doi:10.1016/j.compedu.2014.08.013

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*(6), 423–430. doi:10.1080/00220671.2013.832970

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. doi:10.1016/j.chb.2016.05.047

Slater, S., Bowers, A., Kai, S., & Shute, V. J. (2017). A typology of players in the game Physics Playground. *Proceedings of the 2017 DiGRA International Conference*, 1–12.

Smith, G., Shute, V. J., & Muenzenberger, A. (2019). Designing and validating a stealth assessment for calculus competencies. *Journal for Applied Testing Technology*, *20*(S1), 1–8.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 1–17. doi:10.1016/j.compedu.2019.103672

Sun, C., Shute, V. J., Stewart, A. E. B., Beck-White, Q., Reinhart, C. R., Duran, N., & D'Mello, S. (2022). The relationship between collaborative problem-solving processes and objective outcomes in a game-based learning environment. *Computers in Human Behavior*, *128*, 1–14. doi:10.1016/j.chb.2021.107120

Ventura, M., Shute, V. J., & Small, M. (2014). Assessing persistence in educational games. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), *Design recommendations for adaptive intelligent tutoring systems: Learner modeling* (Vol. 2, pp. 93–101). U.S. Army Research Laboratory.

KEY TERMS AND DEFINITIONS

Competency Model (CM): This addresses the "what to measure" part of ECD, explicitly specifying the knowledge, skills, and other attributes (i.e., the competencies) to be measured by the assessment.

Curriculum Elements (CEs): Collectively, CEs define the knowledge/skill map of the content to be instructed/supported. In intelligent tutoring systems, this is often referred to as the "expert model," and in stealth assessment, this is the "competency model." In both cases, the student model represents a learner's knowledge and progress in relation to the knowledge/skill map.

Evidence Model (EM): The EM deals with the "how to measure" part of ECD and involves: (a) delineating the scoring of performance data (evidence identification), and (b) specifying how the scores will be aggregated (evidence accumulation).

Evidence-Centered Design (or Evidence-Centered Assessment Design; ECD): ECD represents an assessment framework underpinning the development of valid assessment tasks. Its strength lies in basing competency estimates on a chain of evidence that is grounded in task performance, thus ensuring the validity of the assessment. There are several main models in ECD that work in concert: competency model (CM), evidence model (EM), and task model (TM).

Learning (or Performance) Indicators: These represent specific actions (observable) performed by a student related to an outcome variable (unobservable). Typically, indicators consist of at least two main elements—action (verb) and content (referent). This allows us to make claims about competencies, with some degree of certainty (e.g., If a learner does X, then she knows Y, with $p = .95$). Learning indicators are the link between evidence and student model variables.

Stealth Assessment: Stealth assessment refers to evidence-based assessments that are woven directly and invisibly into the fabric of the learning environment (especially well-designed games). During gameplay, learners naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess.

Student Model: The student model represents the current, estimated state of a learner's knowledge and skill, thus serving as the basis for decisions about what to do next (e.g., advance the learner to a more difficult level of the same concept, provide remediation on a particular skill, etc.).

Task Model (TM): The TM outlines the types of items or tasks, including all relevant features, that need to be developed to elicit the competencies of interest from the learner.

APPENDIX 1. LIST OF INDICATORS FOR SCIENTIFIC INQUIRY SKILLS

ACTIVITY LEVEL

1) Total number of actions
2) Total number of experiments
3) Number of changes to the price of the goods

EXPLORATORY BEHAVIORS

4) Number of markets investigated
5) Number of independent variables changed
6) Number of computer-adjusted prices
7) Number of times market sales information was viewed
8) Number of baseline data observations of market in equilibrium

DATA RECORDING

9) Total number of notebook entries
10) Number of baseline data entries of market in equilibrium
11) Entry of changed independent variables

EFFICIENT TOOL USAGE

12) Number of relevant notebook entries divided by total number of notebook entries
13) Number of correct uses of table package divided by number of times table used
14) Number of correct uses of graph package divided by number of times graph used

USE OF EVIDENCE

15) Number of specific predictions divided by number of general hypotheses
16) Number of correct hypotheses divided by number of hypotheses rendered

CONSISTENT BEHAVIORS

17) Number of notebook entries of planning menu items
18) Number of notebook entries of planning menu items divided by planning opportunities
19) Number of times variables were changed that were specified beforehand in the planning menu

EFFECTIVE GENERALIZATION (Event counts)

20) An experiment was replicated
21) A concept was generalized across unrelated goods
22) A concept was generalized across related goods
23) The student had sufficient data (at least three data points) for a generalization

EFFECTIVE EXPERIMENTAL BEHAVIORS (Event counts)

24) A sufficiently large (greater than 10% of possible range) change to an independent variable
25) One of the experimental frames was selected

26) The prediction menu was used to specify an event outcome
27) A variable was changed (per experiment)
28) An action was taken (per experiment)
29) An economic concept was learned (per session)