STEALTH ASSESSMENT

When assessment is seamlessly woven into the fabric of the learning or gaming environment so that it's virtually invisible—blurring the distinction between learning and assessment—this is stealth assessment. It is intended to be invisible, ongoing, support learning, and remove (or seriously reduce) test anxiety while not sacrificing validity and consistency. A good way to describe stealth assessment is with a metaphor. Consider the way that businesses were run before the onset of barcodes in the mid 1970s. Before barcodes, businesses had to close down once or twice a year to take inventory of their stock. But with the advent of automated checkout and barcodes for all items, businesses today have access to a continuous stream of information that can be used to monitor inventory and the flow of items. Not only can a business continue without interruption, but the information obtained is far richer than before, enabling stores to monitor trends and aggregate the data into various kinds of summaries, as well as to support real-time, just-in-time inventory management.

Now think about approaches to assessment in schools today. They are usually divorced from learning where the typical educational cycle is: Teach. Stop. Administer test. Go loop (with new content). But with stealth assessment, schools would no longer have to interrupt the normal instructional process at various times during the year to administer external tests to students. Instead, assessment would be continual and invisible to students, supporting real-time, just-in-time instruction.

Relevance of Stealth Assessment

Why is stealth assessment relevant to education right now? Constructing such seamless and ubiquitous assessments across multiple learner dimensions, with data accessible by diverse stakeholders could yield several educational benefits. First, the time spent administering tests, handling make-up exams, and going over test responses is not particularly conducive to learning. Approximately 10% of class time is currently spent on assessment activities. Given the importance of time on task as a predictor of learning, reallocating that 10% into activities that are more educationally productive is a potentially large benefit that would apply to almost all students in all classes.

Second, by having assessments that are continuous and ubiquitous, students are no longer able to "cram" for an exam. Although cramming provides good short-term recall, it is a poor route to long-term retention and transfer of learning. Thus, standard assessment practices in school 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. And though this statement sounds tautological, it is not how most classes are structured. By moving students towards a model where they will retain more of what they learn, we are enabling them to better succeed in cumulative domains, such as mathematics and science, which are essential to our nation's economic health.

The third direct benefit is that this shift in assessment mirrors the national shift from evaluating students based on the number of years they have occupied seats at a desk to evaluating students on the basis of acquired competencies. A growing number of states are

requiring students to pass a high-stakes final exam in order to graduate high school. While increasing numbers of educators are growing wary of pencil and paper, high-stakes tests for which students must prepare, this shift toward ensuring students have acquired "essential" skills fits with the idea of continuous, stealth assessment.

The remainder of this entry overviews evidence-centered design (which undergirds stealth assessment), and describes briefly an example of a game that has three stealth assessments running within it.

Stealth Assessment and Evidence-Centered Design

Stealth assessment uses an assessment design framework referred to as "evidence-centered design," formalized by Robert Mislevy, Linda Steinberg, and Russell Almond in the late 1990s. In general, the primary purpose of any assessment is to collect information that will allow the assessor to make valid inferences about what people know, believe, can do, and to what degree (collectively referred to as "competencies" in this entry). Accurate inferences of competency states support instructional decisions that can promote learning. Evidence-centered design (ECD) defines a framework that consists of several conceptual and computational models that work in concert. The framework requires an assessor to: (a) define the claims to be made about learners' competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of tasks or situations that will elicit that evidence. Each of these models are now described.

Three Main Models in ECD

Competency Model

The first model in a good assessment addresses the question: What collection of knowledge, skills, and other attributes should be assessed? Variables in the competency model (CM) describe the set of personal attributes on which inferences are based. The term student (or learner) model is used to mean an instantiated version of the CM – like a profile or report card, only at a more refined grain size. Values in the learner model express the assessor's current belief about the level on each variable within the learner's CM.

Evidence Model

The second model is the evidence model which asks: What behaviors or performances should reveal those constructs identified and structured in the CM? An evidence model expresses how the student's interactions with, and responses to a given problem constitute evidence about competency model variables. The evidence model (EM) attempts to answer two questions: (a) What behaviors or performances reveal targeted competencies; and (b) What's the statistical connection between those behaviors and the CM variable(s)? Basically, an evidence model lays out the argument about why and how observations in a given task situation (i.e., student performance data) constitute evidence about CM variables.

Task Model

The third model addresses the kinds of tasks or situations that should be created to elicit those behaviors that comprise the evidence. A task model (TM) provides a framework for characterizing and constructing situations with which a learner will interact to provide evidence about targeted aspects of knowledge or skill related to competencies.

As learners interact with tasks/problems during the solution process, they are providing a continuous stream of data that is analyzed by the evidence model. The results of this analysis are data (e.g., scores) that are passed on to the competency model, which in turn updates the claims about relevant competencies. In short, the ECD approach provides a framework for developing assessment tasks that are explicitly linked to claims about personal competencies via an evidentiary chain (e.g., valid arguments that serve to connect task performance to competency estimates), and are thus valid for their intended purposes.

Brief Example of Stealth Assessment

Newton's Playground is the name of a new computer-based game with 2D physics simulations for gravity, mass, potential and kinetic energy, transfer of momentum, and so on. The goal of all 75 levels in the game is to guide a green ball over to hit a red balloon. Everything in the game obeys the basic rules of physics. Using the mouse, players draw colored objects on the screen, which "come to life" when drawn. These objects apply Newtonian mechanics to get the ball to balloon and they include simple machines such as levers, ramps, pendulums, and springboards.

Three stealth assessments are coded deeply into the game: measuring creativity, conscientiousness, and qualitative physics understanding. Competency and evidence models were created for each of the constructs. This entailed, per construct, about a 10-12 month literature review, then structuring the main competency variables into a model. Evidence was defined as the things a person did in the game that would provide information about particular competency variables. Task models provided a blueprint for creating all of the levels in the game. Levels increased in difficulty across the seven different playgrounds, and each level

focused on eliciting evidence related particular aspects of Newton's laws of motion.

For instance, conscientiousness was modeled with four main facets: persistence, perfectionism, organization, and carefulness. For the persistence facet, we defined a set of observables (i.e., behaviors in the game providing relevant evidence) that included the following: time spent on unsolved levels, number of restarts of a level, and number of revisits to unsolved levels. The game automatically tallies this information in log files that are then analyzed by the stealth assessment machinery. The difference between answering self-report questions about persistence (e.g., "I always try my hardest") and actually exerting substantial effort when trying to solve a hard problem in the game is a clear example of the expression: *Actions speak louder than words*. And they do.

Conclusion

In addition to the direct benefits to education described earlier, there are some indirect benefits as well. For example, our current capacity to assess students is often limited in that it is based on a relatively small number of test items. As we move to a seamless assessment model, we will be able to more accurately assess students since we will have access to a much broader collection of the student's learning data. More accurate assessments enable us to better support student learning across a range of important educational areas.

Valerie J. Shute

See also 21st Century Technology Skills.

FURTHER READINGS

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

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

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