Stealth assessment in video games

Val Shute
Florida State University, USA

Val Shute is the Mack and Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Before coming to FSU in 2007, she was a principal research scientist at Educational Testing Service, where she was involved with basic and applied research projects related to assessment, cognitive diagnosis, and learning from advanced instructional systems. Her general research interests hover around the design, development and evaluation of advanced systems to support learning, particularly related to 21st-century competencies. Her current research involves using games with stealth assessment to support learning of cognitive and non-cognitive knowledge, skills and dispositions. Her research has resulted in numerous grants, journal articles, books, chapters in edited books, a patent, and a couple of recent books, including *Measuring and supporting learning in games: Stealth assessment* (Shute & Ventura, The MIT Press, 2013) and *Innovative assessment for the 21st century: Supporting educational needs* (Shute & Becker, Springer-Verlag, 2010).

Abstract

Games can be powerful vehicles to support learning, but their success in education hinges on getting the assessment part right. In this presentation, I will explore how games can use stealth assessment to measure and support the learning of competencies critical for the future. I will discuss what stealth assessment is, why it is important, and how to develop and accomplish it. I will also provide examples within the context of a game called *Physics Playground* that I designed and developed with my team. I'll share what has been learned by recent research on stealth assessments in games, including:

- Does stealth assessment provide valid and reliable estimates of students' developing competencies, including qualitative understanding of physics, persistence, and creativity?
- Can students actually learn anything as a function of gameplay?
- Are games designed with stealth assessment capabilities still fun?

Preparing our kids to succeed in the future requires fresh thinking on how to design new kinds of assessments that overcome the limitations of traditional assessments, such as multiple-choice tests and self-report questionnaires, and also support learning. Traditional assessments are often too simplified, abstract, and decontextualised to suit current education needs. Alternatively, we can dynamically assess students in engaging, situated environments (like well-designed games) rather than having students fill in bubbles on a standardised test form. We can also provide immediate, ongoing feedback to support learning.

A century ago, traditional assessments were fine because a person who acquired basic reading, writing and maths skills was considered to be sufficiently literate. The goal was to prepare young people for production jobs, because 90 per cent of students were not expected to seek or hold professional careers. But when faced with highly technical and complex problems in today's world, we need to re-examine the nature of educationally valuable skills. Except in rare cases, our current education system neither teaches nor assesses these new competencies, despite a growing body of research showing that skills and dispositions such as persistence, flexibility, creativity, self-efficacy, critical thinking, systems thinking, openness, problem-solving and teamwork (to name a few) can positively impact student academic achievement and other aspects of life.

Games, assessment and learning: A new approach

Increasingly, research shows that digital games can support learning. However, this is usually shown using pre-test–game–post-test designs, where the pre- and post-tests measure content knowledge. Such traditional assessments don't capture and analyse the dynamic and complex performances that inform modern competencies. How can we both measure and enhance learning in real time? I believe that a performance-based approach to assessment is needed. The main assumptions underlying this new approach are that: (a) learning by doing (required in gameplay) improves learning processes and outcomes, (b) different types of learning and learner attributes may be verified and measured during gameplay, (c) strengths and weaknesses of the learner may be capitalised on and addressed, respectively, to improve learning, and (d) feedback can be used to further support student learning.

In a typical digital game, as players interact with the environment, the values of different game-specific variables change. For instance, getting injured in a battle reduces health, and finding treasure or other objects increases your inventory of goods. In addition, solving really hard problems in games permits players to gain rank or 'level up'. One could say that these are all 'assessments' in games: of health, personal goods and rank. But now consider monitoring educationally relevant variables at different levels of granularity via games. In addition to checking health status, players could check their current levels of, for example, systems-thinking skill and teamwork, where each of these competencies is further broken down into constituent knowledge and skill elements (for example, teamwork may be broken down into cooperating, negotiating and influencing skills). If the values of those competencies got too low, the player would likely feel compelled to take action to boost them.

One main challenge for educators who want to employ or design games to assess and support learning is making valid inferences — about what the student knows, believes and can do — at any point in time, at various levels, and without disrupting the flow of the game. One way to increase the quality and utility of an assessment is to use evidence-centred design, which informs the design of valid assessments and yields real-time estimates of students' competency levels across a range of knowledge and skills. Accurate information about the student can be used as the basis for delivering timely and targeted feedback. This information can also be used for presenting a new task or quest that is right at the cusp of the student's skill level, in line with Csikszentmihalyi's flow theory and Vygotsky's zone of proximal development. Given the goal of using educational games to support learning, we need to ensure that the assessments are valid, reliable, and also pretty much invisible (to keep engagement intact). That's where 'stealth assessment' comes in.

Overview of stealth assessment

Very simply, stealth assessment refers to evidence-based assessment that is woven directly and invisibly into the fabric of the learning or gaming environment. During gameplay, students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess. Evidence needed to assess the skills is thus provided by the players' interactions with the game itself (that is, the processes of play). These can be contrasted with the product of an activity, which is the norm for assessment in educational environments.

By analysing a sequence of actions within a problem or quest (where each response or action provides incremental evidence about the current mastery of a specific fact, concept or skill), stealth assessments within game environments can infer what learners know and don't know (or can and can't do) at any point in time. Now, because we typically want to assess a whole cluster of skills and abilities from evidence coming from learners' interactions within a game, methods for analysing the sequence of behaviours to infer these abilities are not as obvious. As suggested above, evidence-based stealth assessments can address these problems.

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 and ongoing, to support learning and to 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. Repeat 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. The remainder of this short paper will briefly describe evidence-centred design (which undergirds stealth assessment), and present a short example of a game that has three stealth assessments running within it.

Stealth assessment and evidence-centred design

Stealth assessment uses an assessment design framework referred to as 'evidence-centred design', formalised 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 and can do, and to what degree (collectively referred to as 'competencies' in this paper). Accurate inferences of competency states support instructional decisions that can promote learning. Evidence-centred design 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.

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 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 competency model — 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 competency model.

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

Task model. The third model addresses the kinds of tasks or situations that should be created to elicit those behaviours that comprise the evidence. A task model provides a framework for characterising 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 or problems during the solution process, they are providing a continuous stream of data that is analysed by the evidence model. The results of this analysis are data (such as scores) that are converted to probabilistic estimates of competency state, which are then passed on to the competency model which updates the claims about relevant competencies. In short, evidence-centred design provides a framework for developing assessment tasks that are explicitly linked to claims about personal competencies via an evidentiary chain (for example, valid arguments that serve to connect task performance to competency estimates), and are thus valid for their intended purposes.

Brief example of stealth assessment

Physics Playground is the name of a computer-based game with two-dimensional 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 coloured 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- to 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 to particular aspects of Newton's laws of motion.

For instance, conscientiousness was modelled with four main facets: persistence, perfectionism, organisation, and carefulness. For the persistence facet, we defined a set of observables (behaviours 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 analysed by the stealth assessment machinery. The difference between answering self-report questions about persistence (for example, '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

Our current capacity to assess students is often limited as 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 students' learning data. More accurate assessments enable us to better support student learning across a range of important educational areas.

References

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher*, 32(2), 13–23.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439–483.

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.

Shute, V.J., Leighton, J.P., Jang, E.E. & Chu, M-W. (In press). Advances in the science of assessment. To appear in *Educational Assessment*.