

# Stealth Assessment

Valerie J. Shute and Seyedahmad Rahimi

November 2022



## Rapid Community Reports

Primers

# Stealth Assessment

*“If we think of our children as plants ... summative assessment of the plants is the process of simply measuring them. The measurements might be interesting to compare and analyze, but, in themselves, they do not affect the growth of the plants. On the other hand, formative assessment is the garden equivalent of feeding and watering the plants—directly affecting their growth.”*  
(Clarke 2001, p. 2)

## Authors

Valerie J. Shute, Florida State University, [vshute@fsu.edu](mailto:vshute@fsu.edu)  
Seyedahmad Rahimi, University of Florida, [srahimi@ufl.edu](mailto:srahimi@ufl.edu)

## Abstract

Traditional classroom assessments are usually summative, but they are constrained in terms of items and time, and can trigger anxiety. As such, they are not well-suited to accurately assess what’s been learned and they rarely engender learning. Stealth assessment is a type of formative assessment, which occurs while a person is actively engaged in learning activities, and is intended to resolve many of the problems related to summative assessments. In this primer, we briefly describe stealth assessment, provide lessons learned via research in this area, and discuss issues and future directions of this work. Resources are provided that further expand on aspects of this stealth assessment approach.

## Keywords

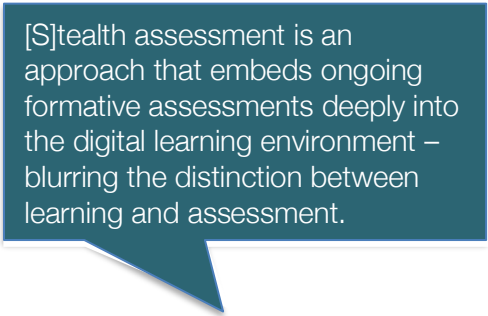
Stealth assessment, formative assessment, educational games, digital learning environments, learning analytics

Suggested Citation: Shute, V. J., & Rahimi, S. (2022). Stealth assessment. *Rapid Community Report Series*. Digital Promise and the International Society of the Learning Sciences.  
<https://repository.isls.org/handle/1/7671>

# Overview

Think back to your high school classes in history, algebra, and chemistry. How did your teachers determine how well you learned the content? Invariably, this was done using midterm and final exams—likely employing multiple-choice items. Such classroom assessments have remained the same across many decades. There are, however, two main types of classroom assessment – summative (like the example above) and formative. Summative assessment (assessment of learning, the more dominant of the two) usually takes the form of a test and represents the end point of learning. Formative assessment (assessment for learning) typically consists of quizzes delivered throughout a course of learning with the goal to provide support to the learner along the way.

There are drawbacks to summative assessments, which are often overlooked given how easy they are to deliver and score. They have a limited number of items and time, thus cannot fully assess what has been taught or learned in class. They also just measure learning at a single point in time summarized by an overall “score” with rarely any formative feedback. The consequences of such traditional or summative assessments can also leave lower-performing students demotivated. This is where stealth assessment comes in (Shute, 2011)—as a possible solution to these problems.



[S]tealth assessment is an approach that embeds ongoing formative assessments deeply into the digital learning environment – blurring the distinction between learning and assessment.

In general, stealth assessment is an approach that embeds ongoing formative assessments deeply into the digital learning environment – blurring the distinction between learning and assessment (Figure 1). Interacting with an immersive game or digital learning environment, students continually produce rich sequences of actions as data points which are captured in log files. The captured data are automatically scored by in-game rubrics, then aggregated in real-time by Bayesian

networks (or other statistical models), which show evolving mastery levels on targeted competencies. Shute, Lu, and Rahimi (2021) provide more information about the steps needed to develop a stealth assessment.

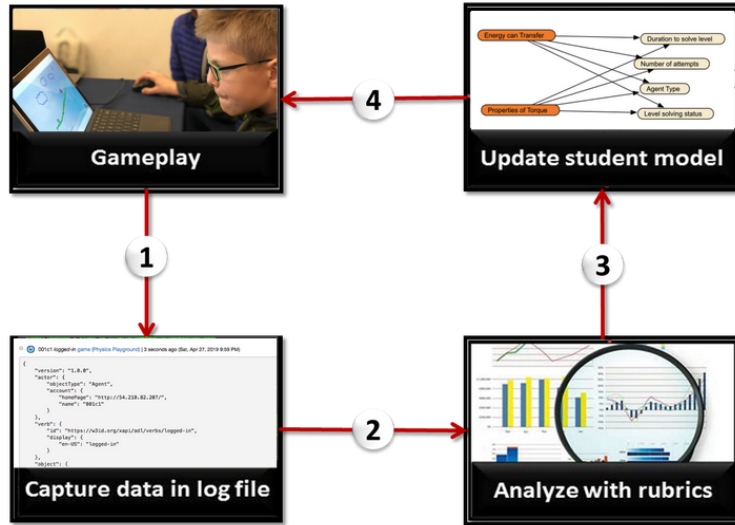


Figure 1: Stealth assessment process

Moving from log file data to valid inferences about competency states is an important yet difficult task. Stealth assessment addresses this issue by employing the evidence-centered design (ECD, see Figure 2) framework (Mislevy, Steinberg, & Almond 2003). ECD provides a way to reason about assessment design and student performance. It consists of conceptual and computational models working together. The three key models include the competency model, the evidence model, and the task model. The competency model delineates everything you want to measure during the assessment. The task model identifies the features of selected learning tasks needed to provide observable evidence about the targeted unobservable competencies. This is realized through the evidence model, which serves as the bridge between the competency model and the task model.

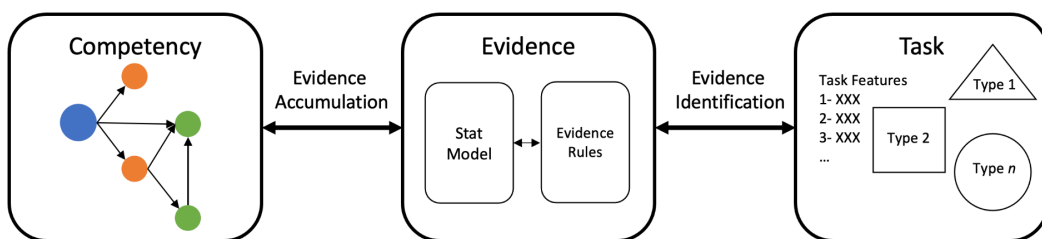


Figure 2. The three key models of ECD (adapted from Mislevy et al., 2003)

The connections between Figures 1 and 2 are as follows. First, Figure 1 shows the stealth assessment process in operation, once it's designed and implemented, while Figure 2 illustrates the assessment design process. Second, in Figure 1, arrow 2 maps to the "evidence identification" part of the ECD process shown in Figure 2 (i.e., automated collection and scoring of incoming performance data), while arrow 3 in Figure 1 represents the "evidence accumulation" process of

ECD (i.e., combining the scored data from the evidence identification process to estimate competency levels using a statistical model such as Bayesian networks). Finally, arrow 4 in Figure 1 goes beyond ECD as it uses current estimates of mastery levels to provide relevant learning support adaptively and unobtrusively so as to not interrupt gameplay/engagement.

In contrast with traditional assessment, digital game-based assessment methods, such as stealth assessment, have the following merits: (a) they are fun and engaging for most kids and can reduce test anxiety; (b) they allow for recording students' interactions in detail (i.e., via the accumulation of log data generated by keystrokes, mouse clicks, and choice patterns), which can be used to analyze students' learning progress and provide ongoing feedback; and (c) they can be designed to provide real-time learning supports (see Shute et al., 2020 for examples of such learning supports) which traditional assessment does not.

## Key Lessons



### Stealth Assessment is Valid

The first step in any stealth assessment project is to validate it—otherwise, you risk making unsupported claims. Over the last decade or so, dozens of research teams have designed and tested stealth assessments in various digital games, measuring a range of knowledge and skills. Some examples include systems thinking in *Taiga Park* (Shute et al., 2010), social skills in *Zoo U* (DeRosier, et al., 2012), persistence in *Poptropica* (DiCerbo, 2014), computational thinking in *Engage* (Min et al., 2015), creative problem-solving in *Oblivion* (Shute et al. 2009), causal reasoning in the *World of Goo* (Shute & Kim, 2011), data literacy in *Storylet* (Chin et al., 2016), risk-taking in *Spheres & Shield* (de-Juan-Ripoll et al., 2020), problem-solving in *Plants vs Zombies 2* (Shute et al., 2016), Newtonian physics, creativity, and persistence in *Physics Playground* (Shute & Ventura, 2013; Shute & Rahimi, 2021), and reading in *Word Knowledge E-Book* (Yang et al., 2021). In summary, the current evidence for game-based stealth assessment suggests that it is valid, reliable, and does support learning. There are, however, many areas to explore regarding the generalizability of this approach and ways to further improve learning.

In summary, the current evidence for game-based stealth assessment suggests that it is valid, reliable, and does support learning.

### Feedback is Crucial for Learning

Stealth assessment is intended as assessment *for* learning, so formative feedback—one of the most important parts of learning anything—should be used as part of the learning supports. However, figuring out the type and timing of feedback is critical yet non-trivial to accomplish. That is, more research is needed on the design/content of feedback as well as the best time to deliver it

(e.g., before a task as an advance organizer or after a task for reflection). For a summary of various types and timings of feedback and their effects on learning, see Shute (2008). Another tricky part of the process involves figuring out ways to not disrupt student engagement in the delivery of feedback. Finally, when designing such learning supports in games, make sure to pay attention to learning and instructional theories (e.g., the first principles of instruction, multimedia principles, and motivational theories) for the best results.

## Make Learning Fun for All

Test anxiety is real, engagement leads to learning, and current classroom tests are limited. But when using games with stealth assessment to measure and support learning, students will likely be engaged and learn. For instance, in one study (Shute et al., 2020) researchers designed and validated a stealth assessment of physics understanding in the game *Physics Playground* (Shute et al., 2019). Not only did the students learn physics as a function of gameplay, but they also enjoyed the experience, rating it on average a 4 on a 1–5 scale (1 = strongly dislike to 5 = strongly like). Moreover, there were no differences in terms of enjoyment by gender or ethnicity.

## Theoretical Foundation is Key

For both measurement and support of learning, develop competency models at the outset. A well-defined, elaborated competency model (created iteratively with the help of experts and deep literature reviews) can help improve the validity of any stealth assessment (for more, see Mislevy et al., 2003). Once a high-quality competency model is established, then associated “learning indicators” (evidence) and real-time scoring/updating methods (e.g., ECD for a top-down approach) can be developed. Later, exploratory methods can find additional learning indicators (e.g., educational data mining or other bottom-up approaches). More research, combining both top-down and bottom-up approaches, is needed to figure out how best to meld these methods. We believe that together these methods can yield even more accurate estimates of students’ competencies and support learning than either method alone.

## Issues



As mentioned, the largest benefits offered by stealth assessment embedded in well-designed games (or other types of digital learning environments, or DLEs) are that these are accurate, engaging assessments that can reduce test anxiety and bias while concurrently fostering the acquisition of important knowledge and skills.

But for this approach to assessment to become mainstream—as ubiquitous, unobtrusive, engaging, and valid—there are a number of hurdles to overcome. Following are some of the more pressing issues that need more research.

## Variability in the Quality of Assessments

Because schools are under local control, students in a given state could engage in hundreds (if not thousands) of DLEs during their educational tenure. Teachers, publishers, researchers, and others will be developing such environments. However, with no standards in place, they will inevitably differ in curricular coverage, difficulty of the material, scenarios and formats used, and many other ways that will affect the adequacy of the game/environment, tasks, and inferences on knowledge and skill acquisition that can justifiably be made from successfully completing activities in the environments. Assessment design frameworks, like ECD, represent a design methodology but not a panacea, so more research is needed to figure out how to equate DLEs or create common measurements from diverse environments. Moreover, it is important to figure out how to interpret evidence where the activities may be the same but the contexts in which students are working are different (e.g., working alone vs. working with another student).

## Accuracy of Students' Learning Progression

While DLEs can provide a greater variety of learning situations than traditional face-to-face classroom instruction, evidence for assessing and tracking learning progressions becomes complex rather than general across individual students. Thus, there is a need to model learning progressions in multiple aspects of students' growth and experiences, which can be applied across different learning activities and contexts (Shavelson & Kurpius, 2012). However, as Shavelson and Kurpius point out, there is no single absolute order of progression as learning in DLEs involves multiple interactions between individual students and situations, which may be too difficult for most measurement theories in use that assume linearity and independence. Clearly, theories of learning progressions in games and other DLEs need to be actively researched and validated to realize their potential.

## Privacy, Security, and Ownership of Student 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 would later be used against the students.

# Conclusion

In closing, we have identified some of the pitfalls of traditional, summative assessment in school as the primary way to evaluate student learning and offered an alternative—stealth assessment. Despite the foregoing issues described above, constructing stealth assessments across multiple learner dimensions, with data accessible by diverse stakeholders, could yield various educational benefits. First, the time spent administering tests, handling make-up exams, and going over test responses is not very conducive to learning. Given the importance of time on task as a predictor of learning, reallocating those test-preparation activities into ones that are more educationally productive would provide potentially 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's 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, suggested herein, mirrors the national shift toward evaluating students on the basis of acquired competencies (see Sturgis, 2014).

Having presented a description of what stealth assessment is, why it is valuable, and how it works, a logical next step includes figuring out how teachers fit into the story. For example, (1) *Use existing games with stealth assessment in the classroom*. As more researchers build educational games with stealth assessment, and if the games have built-in learning dashboards, teachers can see how each student is doing and intervene as needed. (2) *Co-design stealth assessments*. Teachers are valuable partners in designing stealth assessments. In some schools (e.g., Quest to Learn; q2l.org) teachers, instructional designers, and game developers together create games supporting a range of important competencies, like problem-solving and communication skills. As part of our own work, we employed several physics teachers in the design of our *Physics Playground* stealth assessment game.

We believe that it's time to derive and deploy new methods, like stealth assessment, to measure and support learning of not only content, but also important competencies like creativity, problem-solving, persistence, critical thinking, and so on. This has become possible given the increased availability of computer technologies. New technologies make it easy to capture the results of routine student work—in class, at home, or wherever. Such ongoing assessment, integrated into students' day-to-day lives, would make it virtually invisible, in stark contrast with current testing contexts. The aforementioned hurdles, being anticipated and researched in advance, can help to shape the vision for a richer, deeper, more authentic assessment (to support learning) of students in the future.



# References

- Chin, D. B., Blair, K. P., & Schwartz, D. L. (2016). Got game? A choice-based learning assessment of data literacy and visualization skills. *Technology, Knowledge and Learning*, 21(2), 195–210. <https://doi.org/10.1007/s10758-016-9279-7>
- Clarke, S. (2001). *Unlocking formative assessment*. Hodder & Stoughton.
- DeRosier, M. E., Craig, A. B., & Sanchez, R. P. (2012). Zoo U: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction*, 1–7. <https://doi.org/10.1155/2012/654791>
- de-Juan-Ripoll, C., Soler-Domínguez, J. L., Chicchi Giglioli, I. A., Contero, M., & Alcañiz, M. (2020). The spheres & shield maze task: A virtual reality serious game for the assessment of risk taking in decision making. *Cyberpsychology, Behavior, and Social Networking*, 23(11), 773–781. <https://doi.org/10.1089/cyber.2019.0761>
- DiCerbo, K. E. (2014). Game-Based assessment of persistence. *Journal of Educational Technology & Society*, 17(1), 17–28. JSTOR.
- Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Wiebe, E., Boyer, K. E., & Lester, J. C. (2015). DeepStealth: Leveraging deep learning models for stealth assessment in game-based learning environments. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (pp. 277–286). Springer International Publishing. [https://doi.org/10.1007/978-3-319-19773-9\\_28](https://doi.org/10.1007/978-3-319-19773-9_28)
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the structure of educational assessments, *Measurement: Interdisciplinary Research & Perspective* 1(1): 3–62. [https://doi.org/10.1207/S15366359MEA0101\\_02](https://doi.org/10.1207/S15366359MEA0101_02)
- Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.). *Learning progressions in science: Current challenges and future directions* (pp. 13–26). Sense Publishers.
- 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. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.
- Shute, V. J., Almond, R. G., & Rahimi, S. (2019). *Physics playground* (Version 1.3) [Computer software]. Tallahassee, FL. Retrieved from <https://pluto.coe.fsu.edu/ppteam/pp-links/>
- Shute, V. J., Lu, X., & Rahimi, S. (2021). Stealth assessment. In J. M. Spector (Ed.), *The Routledge Encyclopedia of Education* (pp. 1–9). Taylor & Francis.
- 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 educational games. *Journal of Computer-Assisted Learning*, 127–141. doi: [10.1111/jcal.12473](https://doi.org/10.1111/jcal.12473)
- 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., 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). Routledge, Taylor & Francis.
- Shute, V. J., & Kim, Y.-J. (2011). Does playing the World of Goo facilitate learning? In D. Y. Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 359–387). Routledge Books.
- 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.
- Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. The MIT Press.
- Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, 116, 1–13.
- Sturgis, C. (2014). Progress and proficiency: Redesigning grading for competency education. International Association for K-12 Online Learning (iNACOL).  
<https://files.eric.ed.gov/fulltext/ED561319.pdf>
- Yang, D., Zargar, E., Adams, A. M., Day, S. L., & Connor, C. M. (2021). Using interactive e-book user log variables to track reading processes and predict digital learning outcomes. *Assessment for Effective Intervention*, 46(4), 292–303.  
<https://doi.org/10.1177/1534508420941935>

## Resources

- Morris, A. (2020, December 29). Science needs a better way to study creativity—Video games might be the answer, new study. *Forbes Magazine*.  
<https://www.forbes.com/sites/andreamorris/2021/12/29/science-needs-a-better-way-to-study-creativityvideo-games-might-be-the-answer/?sh=597e08074ea4>

# Acknowledgments

This work was supported by the U.S. National Science Foundation (award number #037988, Shute PI) and the U.S. Department of Education (award number #039019, Shute PI). We would also like to thank the team members who helped throughout the past decade in various projects related to Shute’s lab—Russell Almond, Fengfeng Ke, Sidney D’Mello, Ryan Baker, Adam LaMee, Weinan Zhao, Ginny Smith, Seyfullah Tinger, Jiabei Xu, Yoon Jeon Kim, Lubin Wang, Gregory Moore, Curt Fulwider, Lukas Lui, Xi Lu, Chen Sun, Chi-Pu Dai, Renata Kuba, and Xiatong Yang.

Rapid Community Reports are supported by the National Science Foundation under grant #2021159. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.