Stealth Assessment

Valerie Shute, Xi Lu and Seyedahmad Rahimi
Florida State University

Abstract
Stealth assessment is intended to not only measure important competencies, but also to support their development during gameplay or within other types of immersive learning environments. It uses evidence-centered assessment design to create the models (Mislevy, Steinberg, and Almond 2003), in conjunction with a statistical scoring and accumulation approach and educational data mining techniques to capture and analyze real-time data and dynamically measure students' learning. This chapter describes the background and benefits of stealth assessment, reviews a growing body of empirical studies that support stealth assessment in different digital games, and describes how stealth assessment can be embedded in a game via a ten-step iterative process. Finally, the chapter discusses some of the hurdles and obstacles to overcome and current efforts on developing various types of learning supports to accompany stealth assessment in a game.

Keywords: Evidence-centred design; Stealth assessment; Competency model; Evidence model; Task model; Digital learning environments; In-game support.

> Assessment should not merely be done to students; rather, it should also be done **for** students, to guide and enhance their learning.
>
> NCTM (2000)

1. Introduction

Typical classrooms today, just like a century ago, continue to differentiate learning and assessment. Moreover, conventional classrooms are not keeping up with the rapid advances of technologies and the needs of a twenty-first-century education – requiring high-level skill sets such as critical thinking, creativity, collaboration, and communication (Binkley et al. 2010; P21 2015). Consequently, it is important for educators and researchers to find an effective way of assessing and supporting higher-level competencies and encourage deep learning to prepare students for success in the twenty-first century. Over the past few decades, game-based assessment and learning have arisen as possible solutions to this problem.

The idea that digital games can foster learning is becoming more broadly accepted as a number of studies have reported the benefits of using video games to promote knowledge,

skills, and other personal attributes (e.g., Ke and Shute 2015; Shute 2011; Van Eck et al. 2017). Moreover, because twenty-first-century skills are hard to measure through traditional assessment (Binkley et al. 2010; Shute and Wang 2016), well-designed digital games are deemed a viable alternative for assessment of such skills (Shute 2011; Tobias, Fletcher, and Wind 2014). Also, reducing testing time and investing more time on engaging, instructive digital activities can provide more reliable and valid assessment for learning (DiCerbo, Shute, and Kim 2017).

Traditional classroom assessments (e.g., tests and quizzes) have several limitations. For example, they have constraints (e.g., limited number of items, time, and space), and thus cannot fully assess what has been taught in class. In addition, traditional assessment mostly measures learning at a single point in time and does not provide ongoing, formative feedback (DiCerbo, Shute, and Kim 2017). In other words, traditional assessment is like the parable of three blind men touching an elephant – revealing only partial information about students' learning. In addition, the consequences of traditional assessment tend to leave lower-performing students less motivated.

In contrast with traditional assessment, digital game-based assessment methods have the following merits: (a) they are fun to most kids and can reduce test anxiety (Ifenthaler et al. 2012); (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 (Shute and Rahimi 2017; Snow et al. 2015); and (c) they can be designed to provide real-time learning supports which traditional assessment does not (Shute and Wang 2016).

Alongside the benefits of game-based assessment and learning, a critical challenge is how to validly and reliably assess students' knowledge and skills in games (Shute et al. 2009). This is where stealth assessment (Shute 2011) comes in. In a nutshell, stealth assessment is a technology that embeds ongoing formative assessments into the digital learning environment – blurring the distinction between learning and assessment (Shute 2009; 2011; Shute and Kim 2013).

The ensuing sections review the origin and development of stealth assessment, discuss relevant work as part of this research stream, and explain how stealth assessment and learning can be linked together in a digital learning environment. The final section presents key challenges and opportunities facing stakeholders (i.e., researchers, teachers, students, and school administrators).

2. Stealth assessment: background

The term and ideas of 'stealth assessment' were initially presented in 2005 during an AERA symposium on diagnostic assessment (see Shute 2011). Generally, stealth assessment is an evidence-based approach that unobtrusively assesses students' learning progression while they are engaged with highly interactive and immersive environments (e.g., digital games or other digital learning environments – DLEs) (Shute and Kim 2013). Stealth assessment aims to blur the boundaries between game play, learning, and assessment (Shute 2015; Tobias, Fletcher, and Wind 2014) using unobtrusive methods (e.g., eye tracking and log files) to continually collect student data and examine their progression of both cognitive and non-cognitive competencies throughout the game (Taub et al. 2017). In such cases, assessment is part of gameplay (Mayer 2018).

Compared with traditional assessment, stealth assessment has a lot to offer. For instance, stealth assessment: (a) can measure both easy-to-measure things (e.g., declarative knowledge related to a topic) as well as hard-to-measure competencies (e.g., processes and products

related to creativity and problem-solving skills); (b) does not trigger test anxiety because the assessments are invisible and part of an engaging environment; (c) provides ongoing feedback, adaptive support, and personalized assessments to each student; and thus (d) can engender a flow experience during game play/assessment. That is, the data captured and analyzed are used to estimate students' evolving status of knowledge and skills, which can serve as the basis for providing real-time support and adjustments (e.g., hints, rewards, and changing task difficulty and/or order) to move students forward (Shute 2011; Shute and Kim 2013; Shute and Ventura 2013; Shute et al. 2016). This gives students (and teachers) timely updates on how the students are performing and learning in the game, and provides feedback on how to improve and clear up any misconceptions via in-game learning supports (Shute and Emihovich 2018).

Recently, researchers have begun to examine how stealth assessment within games can measure learning in different contexts through analyzing various game behaviours stored in log data (e.g., de Klerk, Veldkamp, and Eggen 2015; Kim, Almond, and Shute 2016). For example, Shute and Ventura (2013) embedded stealth assessment within a digital physics game to measure and enhance students' understanding of qualitative physics. The results showed that (a) the stealth assessment was valid (correlating with an external measure of physics understanding), (b) students playing the game improved their physics understanding after gameplay, and (c) they generally enjoyed the experience, rating it on average a 4 on a 1–5 scale from 1/strongly dislike to 5/strongly like, with males and females enjoying it equally.

Over the past decade, Shute and colleagues have examined stealth assessments in various digital games, measuring a range of knowledge and skills, such as systems thinking in *Taiga Park* (Shute, Masduki, and Donmez 2010), creative problem solving in *Oblivion* (Shute et al. 2009), causal reasoning in the *World of Goo* (Shute and Kim 2011), problem solving in *Plants vs Zombies 2* (Shute et al. 2016), and Newtonian physics, creativity, and persistence in *Physics Playground* (Kim et al. 2016; Shute and Ventura 2013; Shute and Wang 2016).

Similar to the stealth assessment work done by Shute's team, Rowe et al. (2017) used educational data mining techniques to examine how students' in-game behaviours were related to implicit understanding of Newtonian physics in two physics games. Rowe and colleagues used a screen capturing technique, a game data collection architecture, and a human coding technique to collect and then analyze game data. The results showed that students' learning gains in each of the two games could be predicted by particular game behaviour patterns. Also, they found that in-game measures of implicit learning developed through data mining techniques, correlated with external learning outcome measures.

DiCerbo's (2014) study using a commercial children's game called *Poptropica* also explored how game data could be captured and used for making inferences about persistence. This game asked players to visit themed 'islands' and fulfill various quests which were considered hard tasks for the players ranged from 6 to 14 years old. Researcher created game indicators of persistence were based on previous research. The results showed that the game-based measure of persistence was reliable with an alpha coefficient of .87.

Min and colleagues (2015) also employed a collection of machine-learning methods to measure students' computational thinking skills within an immersive game-based learning environment called ENGAGE. The results showed that their deep learning stealth assessment approach could better predict students' posttest scores compared with standard classification techniques. And, comparing the pretest and posttest scores, students' learning gains were significant $t(48) = 6.22$, $p < .001$, with a large effect size ($d = .89$).

In summary, the evidence for digital game-based stealth assessment suggests that it is valid, reliable, and can be used to support learning. Section 3 describes how it works.

3. Stealth assessment: how does it work?

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 probabilistic models like item-response theory, or IRT), which show the evolving probabilities of students' estimated mastery levels on targeted competencies. Stealth assessment follows the Evidence-Centred Design (ECD) framework (Mislevy, Steinberg, and Almond 2003) that provides a way to reason about assessment design and student performance. ECD consists of three key models: competency model (CM), evidence model (EM), and task model (TM).

The CM defines what is intended to be assessed (e.g., knowledge, skill, or other attributes) by operationalizing the target competency into its constituent facets. Students are then classified into various levels (e.g., low, medium, and high) based on their actions within the game/environment relative to CM variables. The EM delineates what students do (i.e., the observables) comprising the indicators of the competencies of interest. The EM includes two parts: (1) *scoring rules*, which are used to score the observables via weighted evidence, and (2) *statistical models*, which set the values of the evidence, accumulate all relevant evidence, then statistically link the observables to the unobservables (i.e., 'feeding the CM'). Finally, the TM describes the task types and characteristics (e.g., task difficulty and format) that can elicit evidence from student performance about the competency of interest (i.e., 'feeding the EM') (see Shute 2011). These three models work together dynamically to assess students' competency levels – at various times and grainsizes.

In addition to employing ECD-based models, another key feature of stealth assessment is that it can provide real-time feedback and other types of learning supports directly within the game (Shute 2008; 2011; Shute and Kim 2013). That is, based on the student's performance and stealth assessment results, the game or digital learning environment can adapt to the student's level of competency. The process of making valid assessments and adapting the ways of providing learning support is essential for the growth of students' competencies (Shute and Wang 2016).

Shute, Ke, and Wang (2017) summarized stealth assessment as a nine-step iterative process. A recap of the nine steps and the addition of another step (step 10) with explanations for each step follows:

1. Develop the competency model (CM) of targeted knowledge, skills, or other attributes based on full literature and expert reviews. It is best to come up with at least two versions of the CM and discuss the two versions with content experts rather than one CM (see Almond et al. 2017).
2. Determine which game (or digital learning environment) the stealth assessment will be embedded into. Having access to the source code of the game (either commercial or homegrown game) is necessary for embedding stealth assessment seamlessly into it.
3. Delineate a full list of relevant gameplay actions/indicators that serve as evidence to inform the CM and its facets. This step can be done either through consultation with content experts or engaging in extensive gameplay (and/or watching YouTube videos of expert solutions) when experts are unavailable. Knowing how to accurately identify different in-game behaviours (i.e., observables) provides links between gameplay and associated competencies.
4. Create new tasks in the game, if necessary (TM). Specific tasks are designed to elicit evidence (i.e., in-game behaviour) of desirable competencies. An example is *Physics*

Playground, a homegrown game (see Shute and Ventura 2013). They started by developing the game, and in current work by Shute and her team, they have developed a range of new task types (e.g., sketching and manipulation levels) and new tasks per task type to measure students' understanding of Newtonian laws of motion, torque and conservation of momentum, and energy and dissipative forces. One of the new task types (manipulation) require students to manipulate three physics parameters (i.e., gravity, mass, and air resistance) via sliders and add external forces using blowers to solve game levels.

5 Create a Q-matrix, which is basically a spreadsheet in which tasks (game levels) are on the rows and the competencies on the columns. The cells contain 0/1 (absent/present) data showing which task is connected to which competency or competencies. This also helps to establish a balanced set of tasks for all the competencies in the CM. Including task difficulty and discrimination estimates in the matrix make it an augmented Q-matrix (Shute, Ke, Almond, Sun, Rahimi, and Lu 2018).

6 Determine how to score indicators using classification into discrete categories (e.g., yes/no, poor, OK, good, very good – relative to quality of the actions). This becomes the automated 'scoring rules' part of the evidence model (EM). The scoring process generally requires a modification for the game codes used to capture player data based on scoring rules. Thus, having a programmer who also knows Bayes networks (BNs) is a plus (see Zhao, Shute, and Wang 2015).

7 Establish statistical relationships between each indicator and associated levels of the CM variables. This is the 'statistical model' part of the EM. Subject-matter experts can provide valuable a priori input.

8 Pilot test BNs and modify parameters. The pilot data can provide valuable empirical data for updating the model. That is, the pilot data can be scored with the expert-provided BNs. If the correlation between the expected posteriori scores and the posttest is low, then calibration needs to be made on the BNs (see Kim, Almond, and Shute 2016).

9 Validate the stealth assessment with external measures (e.g., employ a pretest/posttest of content knowledge/skills to validate the in-game stealth assessment of the content and competencies). Comparing the results of the stealth assessment and pretest/posttest data allows for detecting the flaws of the BNs and gives the direction for model improvement (see Shute and Ventura 2013).

10 Use the current information about a player's competency states to provide adaptive learning support (e.g., targeted formative feedback, progressively harder levels relative to the player's abilities, and so on). This represents current research in Shute's lab and the goal is to incorporate learning supports (cognitive and affective) that do not disrupt, but actually facilitate flow.

Each step can be revisited and revised based on experts' feedback and playtest results. The process involves a team of learning scientists, game developers, instructional designers, measurement experts, and content experts. An accurate stealth assessment can be designed, developed, and implemented via these ten steps to measure and support a variety of content area knowledge and skills. Section 4 examines some of the opportunities and challenges associated with stealth assessment.

4. Stealth assessment: challenges and future research

As mentioned, the largest benefits offered by stealth assessment embedded in well-designed games (or other types of DLEs) are that these are engaging assessments that can reduce test

anxiety and bias while concurrently fostering the acquisition of important knowledge and skills (de Klerk and Kato 2017; Kato and de Klerk 2017; Shute 2011). But for this approach to assessment to become mainstream – as ubiquitous, unobtrusive, engaging, and valid – there are a number of large hurdles to overcome. Following are some of the more pressing issues that need more research.

The first hurdle relates to variability in the quality of assessments within existing games and DLEs. That is, because schools are under local control, students in a given state could engage in hundreds or thousands of DLEs during their educational tenure. Teachers, publishers, researchers, and others will be developing DLEs. 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 DLE, tasks, and inferences on knowledge and skill acquisition that can justifiably be made from successfully completing activities in the learning 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).

The second hurdle involves accurately capturing and making sense of students' learning progressions. That is, while DLEs can provide a greater variety of learning situations than traditional face-to-face classroom instruction and learning, 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 and 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.

Finally, the third 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.

Despite the foregoing hurdles, constructing the envisioned ubiquitous and unobtrusive 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 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 on the basis of acquired competencies (see Sturgis 2014).

It's time to derive and deploy new methods, like stealth assessment, to measure and support content, as well as important yet hard-to-measure competencies like creativity, problem solving, persistence, 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. It could be that twenty-first-century assessment will be so well integrated into students' day-to-day lives that they are unaware of its presence. This contrasts with current testing contexts. However, while the benefits of using a seamless-and-ubiquitous model are clear, applying this idea is tricky. For instance, one risk associated with deploying stealth assessment is that students may come to feel like they are constantly being evaluated which could negatively affect their learning and possibly add stress to their lives. Another risk of continuous assessment could result in teaching and learning turning into 'gaming the system' depending on how it is implemented and communicated. But the aforementioned hurdles and risks, 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.

Towards realizing some of the ideas presented herein regarding stealth assessment, ongoing work is being conducted to develop iterative design processes to achieve a smooth integration of the learning game and its assessment/support mechanisms, which can yield optimal learning results (Ke and Shute 2015). For instance, several funded projects are currently underway, focusing on the enhancement of *Physics Playgroun*d, where both cognitive and non-cognitive supports are being embedded in the game to promote formal physics understanding (e.g., game tutorials, animations, worked examples, interactive definitions, formulas, Hewitt videos, and glossary). In addition to the design and development of learning supports, we are currently developing an adaptive stealth assessment-based level selection algorithm. All of the current projects are using stealth assessment technology to parse data obtained from game log files and affective detectors. For more details and game demonstrations, please visit Shute's lab webpage: https://pluto.coe.fsu.edu/ppteam/.

References and further reading

Almond, R.G., Tingir, S., Lu, X., Sun, C., and Rahimi, S. (presented 2017, August). 'A Validation Tool for Conditional Probability Tables (CPT) for Physics Playground', paper presented at the Bayesian Modeling Application Workshop 2017, Symposium conducted at the meeting of Association for Uncertainty in Artificial Intelligence, Sydney, Australia. John-Mark Agosta and Tomas Singlair (Chair). Available from http://bmaw2017.azurewebsites.net/

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., and Rumble, M. (2010) 'Defining 21st Century Skills. Assessment and Teaching of 21st Century Skills', draft white paper, The University of Melbourne.

de Klerk, S. and Kato, P. (2017) 'The Future Value of Serious Games for Assessment: Where Do We Go Now?', *Journal of Applied Testing Technology* 18(S1): 32–37.

de Klerk, S., Veldkamp, B.P., and Eggen, T.J.H.M. (2015) 'Psychometric Analysis of the Performance Data of Simulation-based Assessment: A Systematic Review and a Bayesian Network Example', *Computers and Education* 85: 23–34. doi:10.1016/j.compedu.2014.12.020.

DiCerbo, K. (2014) 'Game-based Assessment of Persistence', *Journal of Educational Technology and Society* 17(1): 17–28.

DiCerbo, K.E., Shute, V., and Kim, Y.J. (2017) 'The Future of Assessment in Technology Rich Environments: Psychometric Considerations of Ongoing Assessment', in Spector, J.M., Lockee, B., and Childress, M. eds. *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy*, New York, NY: Springer.

IfenthalerD., EseryelD., and GeX. (2012) 'Assessment for Game-based Learning', in Ifenthaler, D., Eseryel, D., and Ge, X. eds. *Assessment in Game-based Learning*, New York, NY: Springer.

Kato, P. and de Klerk, S. (2017) 'Serious Games for Assessment: Welcome to the Jungle', *Journal of Applied Testing Technology* 18(S1): 1–6.

Ke, F., and Shute, V.J. (2015) 'Design of Game-based Stealth Assessment and Learning Support', in Loh, C., Sheng, Y., and Ifenthaler, D. eds. *Serious Games Analytics*, New York, NY: Springer, pp. 301–318.

Kim, Y.J., Almond, R.G., and Shute, V.J. (2016) 'Applying Evidence-centered Design for the Development of Game-based Assessments in Physics Playground', *International Journal of Testing* 16(2): 142–163.

Mayer, R. (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.

Min, W., Frankosky, M.H., Mott, B.W., Rowe, J.P., Wiebe, E., Boyer, K.E., and Lester, J.C. (2015) 'DeepStealth: Leveraging Deep Learning Models for Stealth Assessment in Game-based Learning Environments', in *International Conference on Artificial Intelligence in Education*, Heidelberg: Springer, pp. 277–286.

Mislevy, R.J., Steinberg, L.S., and 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

National Council of Teachers of Mathematics (NCTM) (2000) *Principles and Standards for School Mathematics*, Reston VA: National Council of Teachers of Mathematics.

Partnership for 21st Century Skills (P21) (2015) *Framework for 21st Century Learning*. Retrieved from http://www.p21.org/storage/documents/docs/P21_Framework_Definitions_New_Logo_2015.pdf

Rowe, E., Asbell-Clarke, J., Baker, R., Eagle, M., Hicks, A., Barnes, T., Brown, R., and Edwards, T. (2017) 'Assessing Implicit Science Learning in Digital Games', *Computers in Human Behavior* 76: 617–630.

Shavelson, R.J. and Kurpius A. (2012) 'Reflections on Learning Progressions', in Alonzo, A.C. and Gotwals, A.W. eds. *Learning Progressions in Science*, Rotterdam: SensePublishers, pp. 13–26.

Shute, V.J. (2008) 'Focus on Formative Feedback', *Review of Educational Research* 78(1): 153–189.

Shute, V.J. (2009) 'Simply Assessment', *International Journal of Learning, and Media* 1(2): 1–11. doi:10.1162/ijlm.2009.0014.

Shute, V.J. (2011) 'Stealth Assessment in Computer-based Games to Support Learning', in Tobias, S. and Fletcher, J.D. eds. *Computer Games and Instruction*, Charlotte, NC: Information Age Publishers, pp. 503–524.

Shute, V.J. (2015) 'Stealth Assessment', in Spector, J.M. ed. *Encyclopedia of Educational Technology*, Thousand Oaks, CA: Sage Publications, pp. 675–678.

Shute, V.J. and Emihovich, B. (2018) 'Assessing Problem-solving Skills in Immersive Environments', in Voogt, J., Knezek, G., Christensen, R., and Lai, K.W. eds. *International Handbook of Information Technology in Primary and Secondary Education*, Cham, Switzerland: Springer, pp. 635–646.

Shute, V.J., Ke, F., and Wang, L. (2017) 'Assessment and Adaptation in Games', in Wouters, P. and van Oostendorp, H. eds. *Instructional Techniques to Facilitate Learning and Motivation of Serious Games*, New York, NY: Springer, pp. 59–78.

Shute, V.J., Ke, F., Almond, R., Sun, C., Rahimi, S., and Lu, X. (presented 2018, April) 'Promoting Formal Knowledge and Skills Acquisition in Physics Playground', paper presented at American Educational Research Association, AERA, NYC, NY.

Shute, V.J. and Kim, Y.J. (2011) 'Does Playing the World of Goo Facilitate Learning?', in Dai, D.Y. ed. *Design Research on Learning and Thinking in Educational Settings: Enhancing Intellectual Growth and Functioning*, New York, NY: Routledge Books, pp. 359–387.

Shute, V.J. and Kim, Y.J. (2013) 'Formative and Stealth Assessment', in Spector, J.M., Merrill, M.D., Elen, J., and Bishop, M.J. eds. *Handbook of Research on Educational Communications and Technology* (4th edn), New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group, pp. 311–323.

Shute, V.J., Leighton, J.P., Jang, E.E., and Chu, M.-W. (2016a) 'Advances in the Science of Assessment', *Educational Assessment* 21(1): 34–59.

Shute, V.J., Masduki, I., and 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. and Rahimi, S. (2017) 'Review of Computer-based Assessment for Learning in Elementary and Secondary Education', *Journal of Computer Assisted Learning* 33: 1–19.

Shute, V.J., Ventura, M., Bauer, M.I., and Zapata-Rivera, D. (2009) 'Melding the Power of Serious Games and Embedded Assessment to Monitor and Foster Learning: Flow and Grow', in Ritterfeld, U., Cody, M., and Vorderer, P. eds. *Serious Games: Mechanisms and Effects*, Mahwah, NJ: Routledge, Taylor & Francis, pp. 295–321.

Shute, V.J. and Ventura, M. (2013) *Stealth Assessment: Measuring and Supporting Learning in Video Games*, Cambridge, MA: The MIT Press.

Shute, V.J. and Wang, L. (2016) 'Assessing and Supporting Hard-to-measure Constructs', in Rupp, A.A. and Leighton, J.P. eds. *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Application*, Hoboken, NJ: John Wiley and Sons, pp. 535–562.

Shute, V.J., Wang, L., Greiff, S., Zhao, W., and Moore, G. (2016) 'Measuring Problem Solving Skills Via Stealth Assessment in an Engaging Video Game', *Computers in Human Behavior* 63: 106–117.

Snow, E., Allen, L., Jacovina, M., and McNamara, D. (2015) 'Does Agency Matter? Exploring the Impact of Controlled Behaviors Within a Game-based Environment', *Computers and Education* 82: 378–392.

Sturgis, C. (2014) *Progress and Proficiency: Redesigning Grading for Competency Education*, Vienna, VA: International Association for K-12 Online Learning (iNACOL). Retrieved 15 May 2018 from: https://www.inacol.org/resource/progress-and-proficiency-redesigning-grading-for-competency-education/

Taub, M., Mudrick, N., Azevedo, R., Millar, G., Rowe, J., and Lester, J. (2017) 'Using Multi-channel Data With Multi-level Modeling to Assess In-game Performance During Gameplay With CRYSTAL ISLAND', *Computers in Human Behavior* 76: 641–655.

Tobias, S., Fletcher, D. J., and Wind, A. (2014) 'Game Based Learning', in Spector, J.M., Merrill, M.D., Elen, J., and Bishop, M. eds. *Handbook of Research on Educational Communication and Technology* (4th edn), New York, NY: Springer, pp. 485–503.

Van Eck, R.N., Shute, V.J. and Rieber, L.P. (2017) 'Leveling Up: Game Design Research and Practice for Instructional Designers', in Reiser, R. and Dempsey, J. eds. *Trends and Issues in Instructional Design and Technology* (4th ed.), New York, NY: Pearson, pp. 227–285.

Zhao, W., Shute, V.J., and Wang, L. (2015) 'Stealth Assessment of Problem-solving Skills from Gameplay', in *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL*, pp. 2226–2236. http://www.iitsecdocs.com/volumes/2015