Personalized Learning in Educational Games Using Stealth Assessment

Seyedahmad Rahimi & Valerie Shute
University of Florida
Florida State University

Abstract

This chapter addresses a widespread issue associated with the "one-size-fits-all" approach in educational systems, emphasizing the need for personalized learning. Our proposed solution to this issue is an innovative assessment method called *stealth assessment*, utilizing Evidence Centered Design (ECD) to create valid, reliable, and fair assessments, then applying results of ongoing assessments to enhance individual learning. In this chapter, we describe what stealth assessment is and show how stealth assessment can be used for personalized learning.

Keywords: Personalized Learning, Stealth Assessment, Educational Games, ECD

Introduction

It is usually and rightly esteemed an excellent thing in a teacher that he should be careful to mark diversity of gifts in those whose education he has undertaken, and to know in what direction nature inclines each one most. For in this respect there is an unbelievable variety, and types of mind are not less numerous than types of body. (Quintilian, ca. 90 A.D.)

As the quote above shows, it has long been known that differences among individuals influence learning. However, educational systems in the U.S. and globally have been applying a "one-size-fits-all" style of teaching and learning for centuries (Shute, 2007). Personalized Learning (PL) strives to address this issue by monitoring relevant learner characteristics and adapting the learning environment and experiences accordingly (Shute & Zapata-Rivera, 2012). A successful PL environment should accurately define and assess particular learner characteristics (i.e., the *what* to adapt), then use that information as the basis for real-time personalization (i.e., the *how* to adapt). So, PL environments need learners' data to work. The data can be gathered before and/or during the learning experience and can reflect various aspects of the learner, such as prior knowledge and skills

related to the targeted domain, current affective state, background, and even dispositional variables (e.g., personality traits and motivations). These personal data (what to adapt) are monitored during interactions with the learning environment which requires a digital infrastructure for identifying and accumulating the pertinent data. Consequently, various machine-learning algorithms can determine the how to adapt part of PL. For instance, tasks of particular difficulty and relevant learning supports can be matched relative to learners' current competency level (e.g., easier tasks with targeted supports for those with a low level of competency). Such personalization can prevent cognitive overload and promote learning (Corbalan et al., 2006).

There are several terms employed by researchers in the field to address the needs of individual learners. As van Merriënboer (2023) indicated in his farewell remarks, terms such as *personalized learning* (i.e., personalization based on a complete learner profile of various characteristics), *adaptive learning* (i.e., adaptation of content and instructional strategies using data analytics in real-time), and *differentiated instruction* (i.e., instruction based on assessment results to determine a fixed learning path for different individuals) have been used interchangeably. van Merriënboer (2023) contends that this variation in terminology underscores a lack of fundamental theoretical progress in effectively addressing learners' needs. While acknowledging this concern, we perceive these different categories and names as representing fidelity levels of learning personalization, ranging from low (as in the case of differentiated instruction) to high (as exemplified by personalized learning using a complete learner profile).

Similarly, Plass and Pawar (2020) introduced a taxonomy of adaptivity, adaptive learning, and personalization. Specifically, this taxonomy indicates that a learning environment can be adapted at various *macro levels* (i.e., general categories of the overall learning context in which learners operate such as progression through a course or prerequisite units before taking a course) and *micro levels* (i.e., ongoing learning task becomes adaptive based on learner needs, interests, and real-time assessment of learner knowledge and skills). Ideally, use of any level of learning personalization should lead to greater learning compared to a one-size-fits-all learning approach which ignores learners' differences, needs, and particular assets.

In this chapter, we focus on PL in the context of game-based learning environments. Adoption of PL in educational games has faced some challenges due to the difficulties in designing and implementing games (Zarraonandía et al., 2016) – especially games with accurate measurements and support of learning. Stealth assessment (Shute, 2011; Shute & Rahimi, 2022) is formative in nature, with the main goal to blur the boundaries between learning and assessment (i.e., assessment

for learning; Shute & Rahimi, 2017), thus providing a promising route for PL in educational games. That is, stealth assessment quietly and continuously collects and analyzes learners' interaction data and makes inferences about learners' knowledge, skills, and other attributes in real time. Then, using appropriate algorithms, personalization occurs in various forms (e.g., by adjusting game difficulty or by delivering an appropriate cognitive or affective learning support).

As mentioned above, stealth assessment is intended to be formative (i.e., to support learning; Black & Wiliam, 2012; Shute & Rahimi, 2017). These assessments are usually low-stakes and are performed to help learners achieve their learning goals or to help instructors make necessary changes in their instruction (e.g., repeat a challenging topic). Summative assessments, or high-stakes tests, are usually conducted at the end of an instructional unit and used for purposes such as selection of qualified candidates, promotion, or accountability. The added value of formative assessments is that they support learning in addition to measuring it. Stealth assessment, rooted in the psychometrics field, aims to support learning without sacrificing the psychometric qualities of validity, reliability, and fairness. According to Messick (1994a), "…validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made" (p. 13). *Validity* refers to the extent to which an assessment is assessing what it claims to assess (Messick, 1994; Shute, 2009). *Reliability* refers to the consistency of an assessment (Shute, 2009). *Fairness* refers to the extent to which an assessment is equitable and unbiased for various subgroups (DiCerbo et al., 2016; Dorans & Cook, 2016; Mislevy et al., 2013). Other chapters in this volume also discuss the topic of assessment in different contexts (e.g., Aslan & Tenison, 2025/this volume; Ober et al., 2025/this volume).

This chapter will briefly overview (1) the history of stealth assessment; (2) its base psychometric and instructional design frameworks; (3) different learner characteristics that influence game-based adaptation; and (4) relevant learning, engagement, and motivational theories that guide the design choices to achieve productive personalized learning. Throughout this chapter, we will attempt to link our stealth assessment to an extensible PL design framework by Bernacki and Walkington (under review), inspired by previous work on personalized learning (e.g., Aleven et al., 2017; Plass & Pawar, 2020). The extensible framework for designing PL encompasses a guiding theory for shaping the PL environment. It also consists of an infrastructure layer enabling researchers and educators to implement that theory into a customizable model. This model adjusts

the learning experience according to data gathered from the learner or their interaction with the task (see Figure 1).

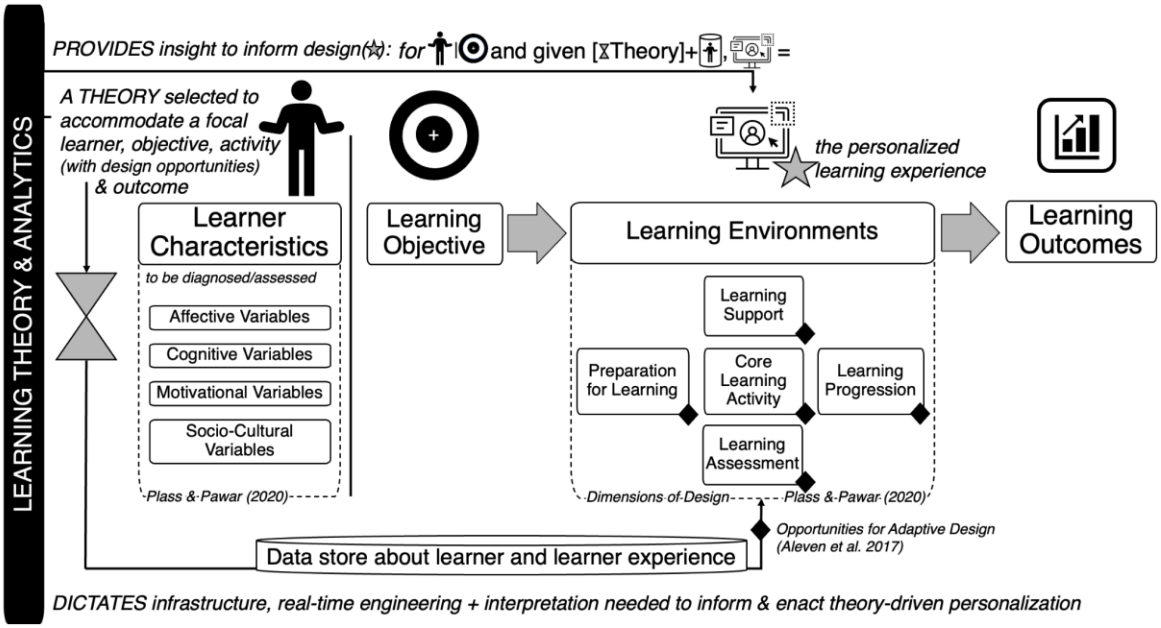
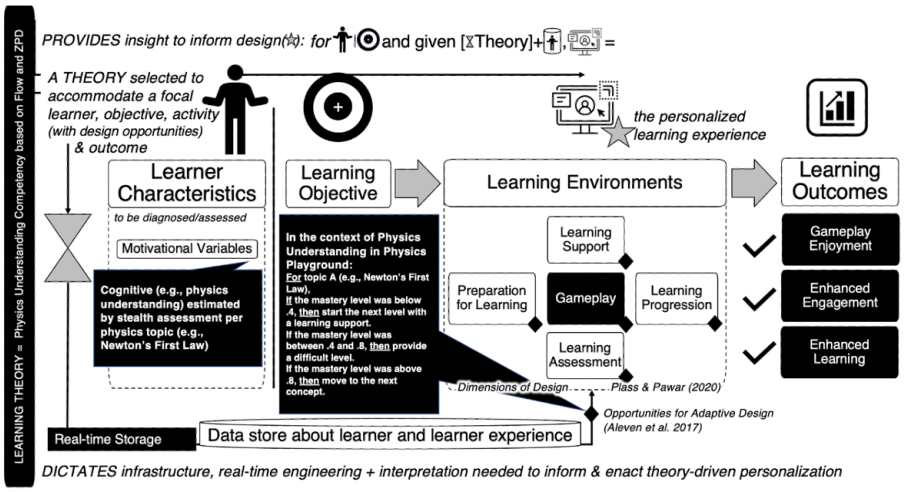


Figure 1. The Extensible PL Design Framework from Bernacki and Walkington (under review)

Notice that there are three components in this extensible PL design framework for theory-driven, analytical personalized learning: (Component 1) establishing a theory of change (i.e., conducting an extensive literature review to identify one or a small set of target variables that exhibit potential to impact the learning process ultimately contributing to the desired outcome); (Component 2) establishing an infrastructure to enact PL (e.g., measurement techniques, data store, and adaptive components); and (Component 3) applying learning analytics and artificial intelligence (AI) to the data store to inform adaptivity. Below (see Figure 2), we adapted the PL design framework to the context of a stealth assessment we designed and will discuss throughout this chapter.



This PL framework is in the context of a stealth assessment of physics understanding embedded into a digital game called Physics Playground. We used a competency model of physics understanding and based the personalized learning theory on flow and ZPD theories. For the learning objectives, we included the thresholds we selected for learning proficiency. The learning activity is gameplay and use of learning supports. Finally, the desired learning outcomes are gameplay enjoyment, enhanced engagement, and enhanced learning of physics. The stealth assessment estimates get accumulated in the logfiles in real-time and will be used for personalized learning.

Figure 2. PL design framework in the context of stealth assessment of Physics Understanding.

Origins, Components, and Accomplishments of Stealth Assessment

The term stealth assessment was coined about 20 years ago by Shute (to read about the full history of stealth assessment see Shute, 2023), but the ideas underlying the term originated from her earlier work on Intelligent Tutoring Systems (ITS; Shute & Psotka, 1996) back in the mid-1980s. ITSs at that time were computer-based learning environments that could measure students' knowledge and skills, then respond "intelligently" to enhance students' learning. Shute designed, developed, and evaluated several ITSs to enhance students' learning of, for example, (a) microeconomics in *Smithtown*, and (b) statistics in *Stat Lady* (Shute, 2023). Through the creation of those ITSs, Shute focused on identifying factors that improved learning (e.g., certain aptitude-treatment interactions). However, the elements responsible for developing content, measuring performance, drawing

inferences about learners' competency levels, and offering instructional supports adaptively were topics that emerged later in Shute's research journey as technology advanced.

By the early 2000s, and inspired by a psychometrics-driven framework called Evidence-Centered Design (ECD; Mislevy, 2013; Mislevy et al., 2004), Shute was able to integrate the individual components she had been working on previously. ECD provided the framework for the development of valid and reliable assessments – a crucial next phase in the evolution of stealth assessment. The core idea of ECD is to center assessment design around the evidence required to make valid and reliable inferences about a learner's knowledge, skills, and other attributes. ECD establishes a direct connection between claims of learner competency (shown as stars in Figure 3) and the data derived from their performance (Mislevy & Haertel, 2006). ECD includes four core models: Competency Model (CM), Evidence Model (EM), Task Model (TM), and Assembly Model (AM)—see Figure 3.

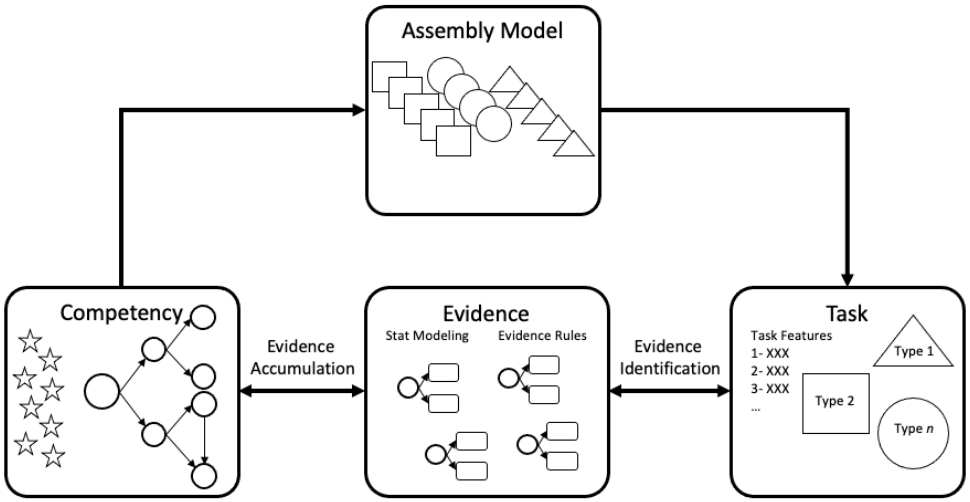


Figure 3. The four core ECD models.

The Competency Model (CM) defines what to assess, such as knowledge, skills, and other attributes. It provides a clear definition of the latent (*unobservable*) variables that characterize the targeted competencies, enabling the inference of learners' competency levels on these specific variables (Almond & Mislevy, 1999). The CM informs the student model (i.e., a profile detailing each learner's current knowledge and skill states). To define the CM, one needs to conduct a deep literature review and consult with domain experts. During the CM development process, assessment designers identify the main, and associated sub-facets that operationalize the competency of interest. Then, the relationships and strengths among those variables can be defined. See Rahimi et al. (under review) for a detailed explanation about how to design a CM.

The Evidence Model (EM) is responsible for defining specific behaviors, indicators, or *observables* that reveal the targeted competencies. The learner behaviors and their corresponding scoring mechanisms (or rubrics) are articulated through *evidence rules*. Meanwhile, *statistical models* establish connections between these behaviors and the CM variables, thereby forming the statistical model. This model can incorporate both simple dichotomous models (e.g., correct/incorrect) and graded models (e.g., low, medium, high), as seen in Bayesian Networks (Shute & Ventura, 2013). The EM encompasses two vital processes: evidence identification (EI), which identifies observables using evidence rules from log data, and evidence accumulation (EA), which aggregates these observables through statistical models and updates the student model (Almond et al., 2020).

The Task Model (TM) focuses on outlining the characteristics of assessment tasks, including elements like difficulty level and format. Its goal is to aid the construction of learning tasks that are designed to elicit behaviors that are required as evidence. Essentially, the TM plays a crucial role in producing assessment tasks (which can be presented as game levels), that are carefully crafted to align with the targeted competency variable(s). The TM encompasses a diverse collection of task types, forming the basis for the development of specific assessment tasks (Almond et al., 2014). Task types also can be seen as templates where task authors (e.g., game developers) can refer to and create as many instances as they want based on those templates (Rahimi, Almond, et al., 2023a). Each template includes various features that can help elicit the type of evidence needed for the competencies of interest.

The Assembly Model (AM) describes the final collection of tasks (e.g., game levels or test items) and how they are sequenced or delivered to the learners (e.g., in a linear or an adaptive manner). The AM ensures that sufficient evidence is collected for a reliable and valid assessment. In a game-based assessment these are often levels or challenges in the game. The reliability of a stealth assessment is determined by the number and evidentiary strength of the challenges. That is, if we have too few observations for a particular competency variable, we can't make a claim that our assessment is consistent/reliable. The best way to improve the reliability of an assessment is to manipulate the number and types of challenges offered. The validity of the assessment is established by producing a collection of challenges that span both the depth and breadth of the competencies being measured. If an assessment does not cover all the competencies (and their sub-facets), it will not be accurate/valid.

In essence, ECD spells out the way to create various models, and operates as a system where the CM, EM, TM, and AM work together, forming the backbone of a stealth assessment's

functionality. This framework ensures that assessment tasks are not only valid but also finely tuned to capture the specific competencies they aim to measure.

The main aspects of a well-designed stealth assessment include: (1) the use or creation of a technology-rich environment such as a digital game or other immersive digital environment (note that stealth assessment is not bound to digital games, but games are highly engaging and can bring out the best performance of the learners); (2) the application of ECD in the design, development, and implementation of the core models described above, (3) the embedding of the stealth assessment into the code of a technology-rich environment; and (4) the creation of capabilities for the system to provide formative assessment/feedback, and/or adaptivity—of game levels, supports, and so on.

In the past decade, we have designed, developed, and validated various stealth assessments of physics understanding (Shute, Rahimi, et al., 2020), creativity (Shute & Rahimi, 2020), calculus (Smith et al., 2019), persistence (Rahimi, Shute, & Zhang, 2021; Ventura & Shute, 2013), and problem solving (Shute et al., 2016). Recently, we conducted a systematic review to see how this method has been used by others (Rahimi, Shute, et al., 2023). We found 93 studies that used stealth assessment to assess various competencies in various contexts. We were able to categorize the competencies into basic knowledge (e.g., physics, math, biology), 21st century skills (e.g., creativity, critical thinking, problem-solving), and social and emotional learning (e.g., risk taking, self-explanation). Moreover, scholars from various fields of study such as computer science, educational technology and learning sciences, and health have used stealth assessment, indicating the diverse uptake of this method. The findings of the systematic review indicate that stealth assessment has been widely adopted and adapted by researchers in multiple contexts to assess and, in some cases, support learning in a personalized way (to read about the steps of stealth assessment see Shute et al., 2021).

Reflecting on the alignment between stealth assessment and the extensible PL design Framework by Bernacki and Walkington (under review), we argue that the CM aligns with Component 1 of the extensible PL design framework; while the EM, TM, and AM are aligned with Component 2 (establishing infrastructure to enact PL). Also, when stealth assessment is implemented, the technical underpinnings (i.e., software architecture) that make stealth assessment functionally possible (Rahimi, Almond, et al., 2023b) are also aligned with Component 2. Moreover, the statistical modeling (EM), and more specifically, the evidence accumulation process in stealth assessment is aligned with Component 3 (applying learning analytics for PL).

We see the extensible PL design Framework by Bernacki and Walkington (under review) and stealth assessment as similar frameworks that align nicely. This paper aims to show the linkages between the two frameworks and provide an example that can be followed, specifically relative to an educational game. One might use either framework to design and develop an adaptive learning environment which promotes personalized learning.

The Theories Underlying and Driving Stealth Assessment to Support Learning

There are several main theories that underlie stealth assessment design. First, there is flow theory (Csikszentmihalyi, 1997) which indicates that if the difficulty of the challenge at hand is matched with one's current knowledge or skill level, that person will likely experience engagement. In a state of flow, people typically lose track of time, perform effortlessly, and even become creative. If challenges are too easy, learners can easily become bored; and if the challenges are too difficult, the learners may get frustrated.

Various activities can induce flow such as reading a really good book or playing a well-designed game. Well-designed games usually include incremental challenges that are matched with the players' skill level (Shute & Ke, 2012) which leads to flow inducement. Stealth assessment aims to match students' knowledge/skill level and selected challenges to facilitate flow (i.e., challenges right at the outer edges of do-ability for a person). Research has shown that engagement within educational games can enhance learning (Dede, 2009; Hookham & Nesbitt, 2019; Sabourin & Lester, 2013; Shute, Rahimi, et al., 2020); therefore, if a game-based learning environment can induce flow (i.e., promote engagement) for different learners, learning can be enhanced.

The second theory underlying stealth assessment is based on is Vygotsky's Zone of Proximal Development (ZPD; Vygotsky, 1978). Similar to flow theory, ZPD refers to the range of tasks that a learner can perform with guidance or assistance from someone more knowledgeable than themselves. In other words, ZPD is the gap between what a learner can do independently and what they can achieve with the support of a skilled mentor or instructor. The ZPD is seen as the optimal area for learning, where learners are challenged to reach just beyond their current level of competence, thereby facilitating growth and skill development. Therefore, the best instruction (and consequently learning) happens where learners can perform some activity with just some small amount of guidance. In stealth assessment, it's important to collect accurate and real-time information about a learner, which is then used as the basis for delivering timely and targeted formative feedback, as well as presenting a new task or quest that is right at the cusp of the student's skill level. Again, this is in line with both flow theory and Vygotsky's ZPD. Stealth assessment aims

to figure out where each learner is in terms of their ZPD and deliver the appropriate learning support/instruction for the learner.

Stealth assessment incorporates additional learning theories. For example, situated learning theory (Lave & Wenger, 1991) as a social constructivist learning theory, indicates that learning is inherently tied to the context in which it occurs. According to this theory, learning is most effective when it takes place within authentic, real-world contexts, where knowledge and skills are acquired through active participation in meaningful activities. Situated learning emphasizes the importance of social interactions, collaboration, and apprenticeship-like experiences in the learning process. Digital games provide a close-to-real-life context for learners in which they can deepen their learning. With authentic problem solving, and immediate and formative feedback, digital games can afford a realistic framework for experimentation and situated learning, and therefore act as rich environments for active learning.

The Underlying Instructional Design Approaches in Stealth Assessment

As discussed earlier, ECD is an assessment design framework which can be coupled with some instructional design principles. That is, when designing a stealth assessment, designers—both assessment and game designers—should ensure an alignment between what is being assessed or learned (the competencies being targeted) with the content and mechanics of game tasks. Specifically, at the CM development stage, assessment designers must write claims about student learning. Claims are statements that can be made about what a learner can do in the future. One form of writing claims could be in the form of learning objectives (see Rahimi et al., under review). For instance, a learner who is high on creativity will be able to produce many ideas about solving a given problem.

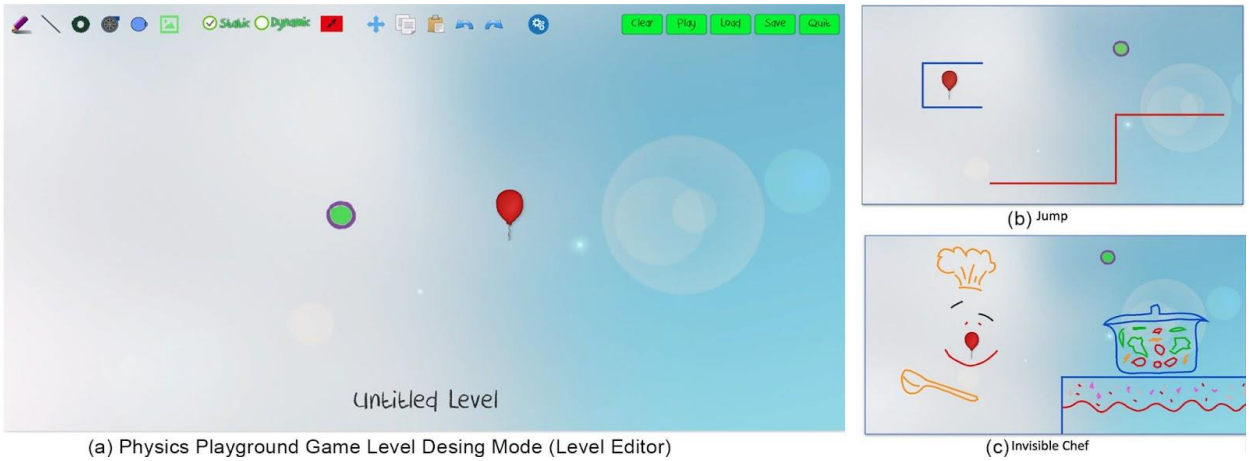
Other concepts shared between psychometrics and instructional design that are relevant to stealth assessment include coverage and sequencing. In psychometrics, assessment designers should make sure to cover all aspects of the competency of interest by including enough test items (in the case of stealth assessment, game levels); similar to instructional design approaches where an instructional designer should include appropriate content for all the objectives. Moreover, similar to sequencing strategies in an instructional design approach, at the AM stage, stealth assessment designers should think about various ways they need to sequence game levels. In relation to assessment and personalized learning, stealth assessment designers need to craft appropriate algorithms for adaptivity. For instance, one can select a data-driven cutoff for learners' mastery level

per competency in a game. If a learner, for example, reached the threshold set by the assessment designers, they can move to game levels covering the next competency.

Physics Playground as a Case Study for Stealth Assessment

At the EM, TM, and AM development stages, assessment designers should think about various elements of their well-designed learning game (Plass et al., 2015; Shute & Ke, 2012). These learning game design elements include game mechanics, visual aesthetic design, narrative design, incentive systems (Rahimi, Shute, Kuba et al., 2021), and learning supports (Bainbridge et al., 2022; Rahimi et al., 2022; Shute, Smith et al., 2020). One of the biggest challenges that game-based learning researchers have faced is the smooth integration of high-quality learning supports (content) in games without sacrificing the fun of gameplay (Shute, Rahimi, et al., 2020). To address this challenge, we used Mayer's (2020) multimedia principles and Merrill's (2002) First Principles of Instruction to design the learning supports in an educational game called *Physics Playground* (Shute et al., 2019).

Physics Playground is a 2-dimensional game created for 7th and 8th graders targeting Newtonian physics understanding with more than 100 game levels. The goal of the game is to hit a red balloon with a green ball achieved by drawing simple machines (i.e., ramp, springboard, lever, and pendulum) and objects (e.g., weights) on the screen. *Physics Playground* also includes a level editor whereby non-technical users (e.g., students and teachers) can create their own levels by drawing objects (e.g., lines, shapes, etc.) on the screen. Figure 4 shows the level editor with two game level examples.



(a) Physics Playground Game Level Desing Mode (Level Editor)

Figure 4. *Physics Playground*'s level editor and two example game levels.

Stealth Assessment of Physics Understanding. To accurately assess learners' physics understanding, we consulted with two physics experts and identified the appropriate competency model of physics understanding (Figure 5) encompassing three levels: (1) physics understanding which represents an aggregated, overall estimated of the targeted physics content; (2) middle-level competency variables which include facets of force and motion, linear momentum, energy, and torque (we did not have direct game levels for these sub facets); and (3) low-level competency variables which include sub-facets of Newton's Laws, properties of momentum and conservation of momentum, energy can transfer and dissipate, properties of torque, and static equilibrium.

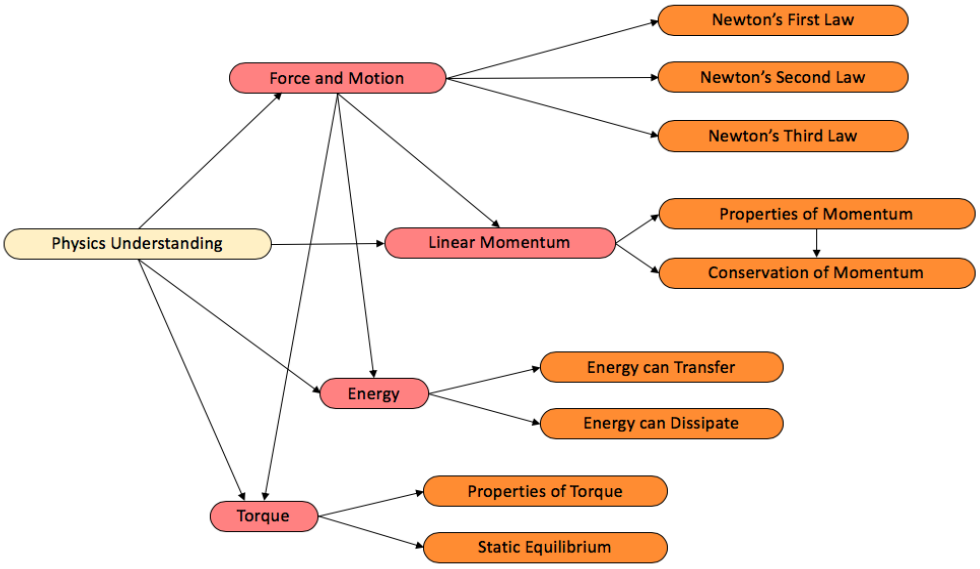


Figure 5. Physics Understanding competency models.

We included two main components in our Task Model: (1) Presentation material: this component defined what the learners could see for a given game level, and also the possible game mechanics present; and (2) Work product: This related to what the learners could do in the game level (e.g., draw lines, create simple machines). Then using an iterative process, we specified the Assembly Model, and after consulting with our physics experts, we created about 10-15 game levels per low-level competency model variable. In the Evidence Model, we came up with a list of observables (indicators) that would provide evidence for the low-level competency variables. We then came up with two important aspects of the EM: (1) *rules of evidence* which are the rubrics related to scoring the observables; and (2) *statistical modeling* for which we used Bayesian Networks to accumulate the scored evidence and estimate students' physics understanding—both on the mid-level variables and overall physics understanding. These estimates then were used for personalization

of learning. For instance, the stealth assessment estimates were the basis for delivering learning supports to the students in the adaptive version of *Physics Playground*.

Learning Supports in *Physics Playground*. *Physics Playground* includes cognitive supports, such as short videos/animations explaining the physics behind each game level using the same look and feel of the game. Our iterative, designed-based research on learning supports in games showed that designing such supports using two instructional design models can lead to the best learning results (Kuba et al., 2021; Shute, Smith, et al., 2020).

First, we designed all learning supports so they had the same look and feel of the game following Mayer's principles of multimedia (Mayer, 2020). Each game level was connected to one physics competency. Therefore, we created very short (just less than one minute) videos for each of the underlying physics concepts of the game level at hand. Each video was designed to avoid extraneous cognitive load and be easy to follow. We also incorporated Mayer's multimedia principals such as signaling, coherence, temporal contiguity, and use of human voice in their design.

Second, we used Merrill's First Principles of Instruction (2002) to guide the learning support development efforts. Investigating many learning theories and instructional models, Merrill (2002) identified five principles of instruction for effective, efficient, and engaging learning: (1) *Problem-centered*: learners learn better when they engage in solving real-world problems; (2) *Activation*: learners learn better when they get to activate relevant prior knowledge or previous experiences; (3) *Demonstration*: learners learn better when they observe a demonstration of what is to be learned rather than merely being told what is to be learned; (4) *Application*: learners learn better when they apply the new knowledge or skill to solve problems; and (5) *Integration*: learners learn better when they integrate the new knowledge or skill into their everyday life.

We followed these principles in designing the learning supports in *Physics Playground* (Kuba et al., 2021). Specifically, we designed the learning supports around a problem (i.e., game level) similar to what learners would need to solve (i.e., the problem-centered principle). Moreover, we designed these learning support videos using already-familiar elements and game mechanics (i.e., the activation principle). Also, we demonstrated a failed attempt followed by a successful attempt in these videos (i.e., the demonstration principle). Finally, since it is important to direct learners to the content in educational games, we designed an incentive system that was used to only nudge the learners toward these learning supports (Rahimi, Shute, Kuba, et al., 2021). This incentive system would give game points to the learners only for the first time they attempted to see the learning

supports. Our findings suggest that even when these learning supports were not incentivized, learners accessed them showing the value of the supports (Rahimi, Shute, Kuba, et al., 2021). Additionally, we used the stealth assessment estimates of competency levels in *Physics Playground* to deliver learning supports to the student who was at just the right level. If a person's estimated level of mastery was below a certain point, the game showed the learner a support associated with the next game level before the learner started to play the level.

In summary, both in the design phases of assessment and learning supports in educational games, the use of good instructional design models, theories, and principles can lead to successful personalization. There are, however, other learner characteristics that can be assessed and used for adaptation—e.g., cognitive, affective, and interest—discussed next.

Learner Characteristics for Personalized Learning. To date, we have considered two main kinds of learner characteristics on which to base adaptivity: (a) cognitive—e.g., the current estimated level of learners' mastery during gameplay (Shute, Rahimi, et al., 2020), and (b) affective—e.g., frustration and boredom (Bainbridge, Smith, et al., 2022). After going through several short introductory levels and game tutorials, the stealth assessment system in the game (e.g., *Physics Playground*) starts to gather evidence on student mastery of various physics concepts. After sufficient evidence is gathered about one's competency level on a given topic (e.g., Newton's Second Law of Force and Motion), the stealth assessment adapts the level of difficulty of the next game level to match learner's level of competency.

Stealth assessment estimates can also be used in deciding on the delivery of learning supports. For instance, in the adaptive version of *Physics Playground*, we used a proficiency range of .4 to .8 as an indicator that a learner was at an acceptable proficiency level where they did not need a learning support. If the learner fell below .4 on a given competency (e.g., Newton's First Law), they would start the next level with a relevant learning support. An estimate greater than .8 on a topic indicated that the learner mastered that topic and can move on to the next one. The possible range computed by the stealth assessment system was any number between 0 and 1. We chose .4 to .8 based on our experts' opinion.

Similar to estimating cognitive skill levels, stealth assessments can be designed to detect evidence of students' affective states of frustration and boredom. Those detectors then can be used to trigger appropriate interventions to regulate and enhance one's affective state. For instance, using data mining techniques, we developed a quit prediction model (Karumbaiah et al., 2018; Slater et al.,

2020). This model included variables from log data that were predictors of quitting behavior. Once the probability of quitting reached a certain threshold, the model triggers an affective support for the learners (e.g., a motivational message, a fun physics-related video, or a short breathing game). The goal of such interventions is to help learners feel less frustrated and come back to gameplay and engage in solving problems and learning.

Apart from cognitive and affective characteristics being detected and treated (personalized) by the stealth assessment machinery, we have designed opportunities for learners to create and customize game levels for themselves, based on their interests (both situational and general interests are relevant; Krapp, 2005; Renninger, 2000). Many popular games have a level editor or a "create" mode where players can make their own game levels (e.g., *Little Big Planet*, *Minecraft*, *Portal 2*). These sandbox environments allow players to act based on their interest and be creative through use of multiple tools and endless possibilities.

New levels can be saved and later uploaded to the game via a website where all the levels get stored and added to the game without the need for any programming. Starting with an empty stage, students could place the ball and the balloon anywhere on the screen and draw any number of obstacles between them (see Figure 6 for four example levels that students created). The level editor can also provide a lot of opportunities for learners to practice being creative (Rahimi & Shute, 2021).

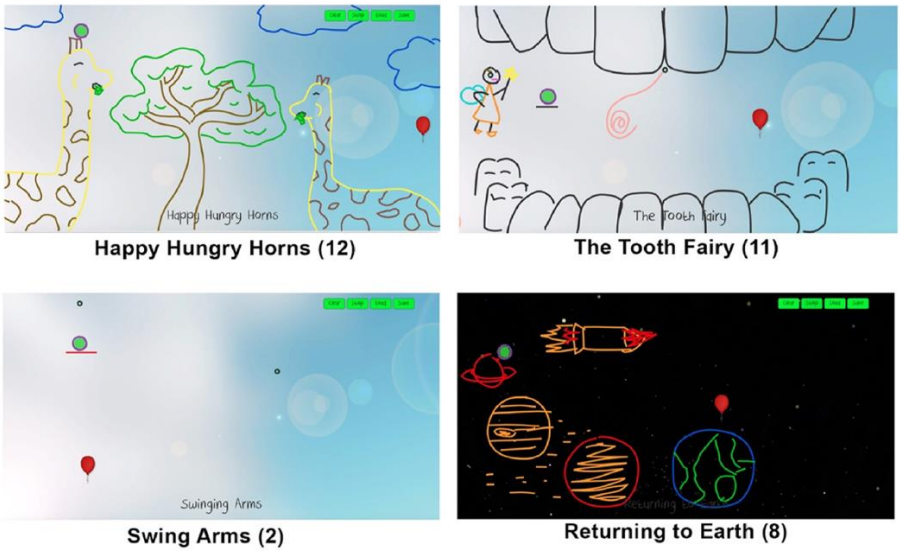


Figure 6. *F*our example levels created by learners.

The second game aspect we included in *Physics Playground* to let learners personalize their gameplay pertains an in-game store (Figure 7). Using in-game money that the learners' earned in the game, they can unlock the in-game store and spend their money to change the background music,

the face/type of ball, and alter the background image for subsequent game levels. This in-game store has also been used as an affective support in *Physics Playground* (Bainbridge, Smith, et al., 2022) since the store can serve as a way to redirect learners' attention from playing the game to a different, less taxing task (i.e., selecting different background music).

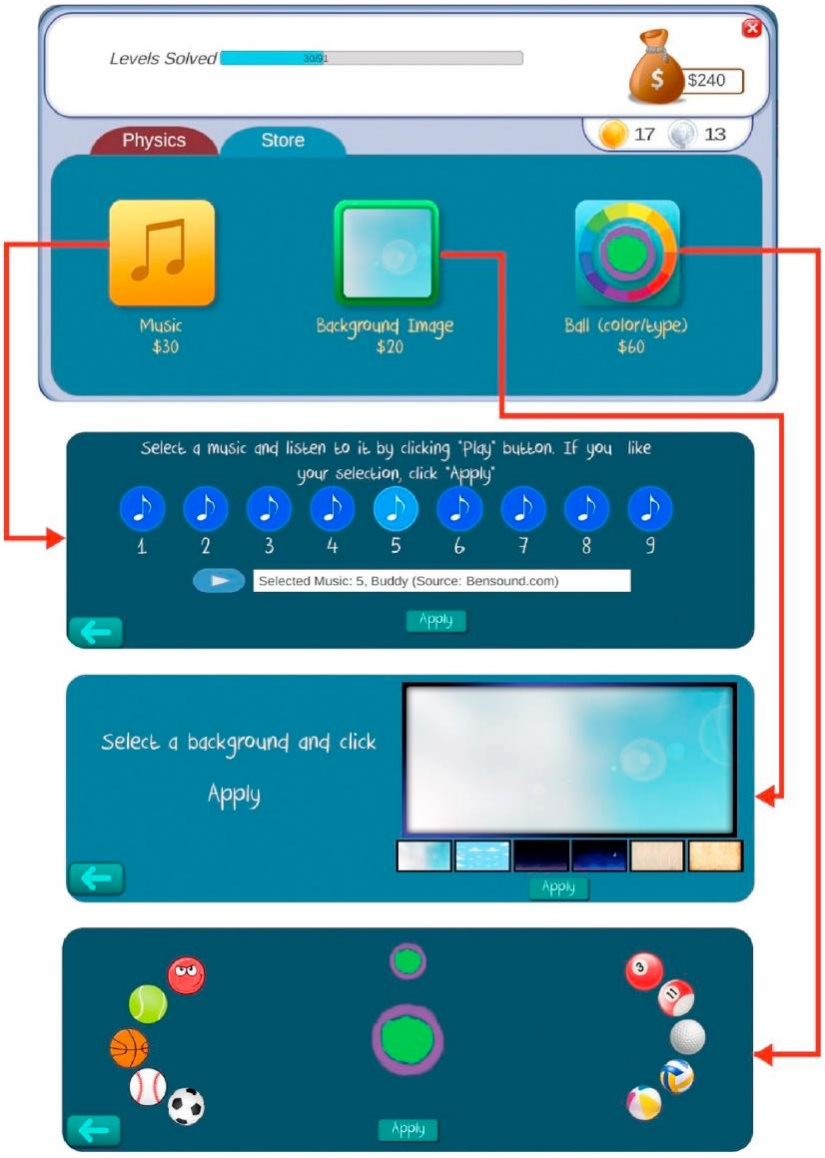


Figure 7. *Physics Playground*'s in-game store for customization.

Conclusion

This chapter examines our work on personalized learning, with a particular focus on stealth assessment in educational games. We discussed the roots of stealth assessment, its theoretical foundations, instructional design principles, and the integration of learner characteristics for

effective personalization of learning. We also presented some examples of personalized learning and adaptivity in *Physics Playground*. Stealth assessment, particularly when ECD is used, is a promising technique to promote personalized learning. As we discussed throughout this paper, stealth assessment and the extensible PL design framework align well. This alignment indicates that stealth assessment is not just about assessing learners' various knowledge and skills. Rather, this assessment technique is complete when the learning environment adapts to learners' competency levels, needs, and interests.

Moving forward in this arena, there are certain challenges that we need to think about and address (Shute, Leighton, et al., 2016; Shute & Rahimi, 2017). For instance, there are ethical considerations surrounding data privacy which demand attention. Another challenge lies in the integration of personalized learning approaches, especially game-based ones such as stealth assessment, with traditional educational structures. This challenge requires the demonstration of clear, measurable learning outcomes that align with established standards. Despite these challenges, the future of personalized learning is promising. Leveraging advances in Artificial Intelligence to personalized learning can enhance accuracy, providing a more nuanced understanding of learner characteristics and subsequent help to personalize learning experiences more successfully than before. These advancements offer the potential for more precise personalization, addressing the challenges discussed and contributing to more inclusive, adaptable, and effective personalized learning environments.

References

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of Research on Learning and Instruction* (pp. 522–560). Routledge.

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. https://doi.org/10.1080/15366367.2014.910060

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223–238.

Almond, R., Shute, V., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in game-based assessments. In *Innovative Psychometric Modeling and Methods* (pp. 163–192). Information Age Publishing.

Bainbridge, K., Shute, V. J., Rahimi, S., Liu, Z., Slater, S., Baker, R. S., & D'Mello, S. (2022). Does embedding learning supports enhance transfer during game-based learning? A case study with Physics Playground. *Learning and Instruction*, *77*, 1–11. https://doi.org/10.1016/j.learninstruc.2021.101547

Bainbridge, K., Smith, G. L., Shute, V. J., & D'Mello, S. (2022). Designing and testing affective supports in an educational game. *International Journal of Game-Based Learning*, *12*(1), 1–32. https://doi.org/10.4018/IJGBL.304434

Black, P., & Wiliam, D. (2012). Assessment for Learning in the Classroom. In *Assessment and Learning* (pp. 11–32). SAGE Publications Ltd. https://doi.org/10.4135/9781446250808.n2

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science*, *34*(5), 399–422. https://doi.org/10.1007/s11251-005-5774-2

Csikszentmihalyi, M. (1997). *Creativity: Flow and the psychology of discovery and invention*. Basic Books.

Dede, C. (2009). Immersive Interfaces for Engagement and Learning. *Science*, *323*(5910), 66–69. https://doi.org/10.1126/science.1167311

DiCerbo, K. E., Shute, V., & Kim, Y. J. (2016). The future of assessment in technology-rich environments: Psychometric considerations. In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, Design, and Technology* (pp. 1–21). Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_66-1

Dorans, N. J., & Cook, L. L. (2016). *Fairness in Educational Assessment and Measurement*. Routledge.

Hookham, G., & Nesbitt, K. (2019). A Systematic Review of the Definition and Measurement of Engagement in Serious Games. *Proceedings of the Australasian Computer Science Week Multiconference*, 1–10. https://doi.org/10.1145/3290688.3290747

Karumbaiah, S., Baker, R. S., & Shute, V. (2018). Predicting quitting in students playing a learning game. *11th International Conference on Educational Data Mining*, 1–10.

Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. *Learning and Instruction*, *15*(5), 381–395.

Kuba, R., Rahimi, S., Smith, G., Shute, V., & 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. https://doi.org/10.1007/s11423-021-09994-3

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

Mayer, R. (2020). *Multimedia Learning*. Cambridge University Press.

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, *50*(3), 43–59.

Messick, S. (1994a). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Messick, S. (1994b). *Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performancesas Scientific Inquiry into Score Meaning*. 33.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine*, *178*, 107–114.

Mislevy, R. J., Almond, R. G., & Lukas, J. (2004). *A brief introduction to evidence-centered design* (CSE Technical Report 632). The National Center for Research on Evaluation, Standards, Student Testing (CRESST). http://www.cresst.org/reports/r632.pdf

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., & others. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation*, *19*(2–3), 121–140.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of Game-Based Learning. *Educational Psychologist*, *50*(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Rahimi, S., Almond, R. G., & Shute, V. J. (2023a). Getting the first and second decimals right: Psychometrics of stealth assessment. In M. P. McCreery & S. K. Krach (Eds.), *Games as Stealth Assessments* (pp. 1–40). IGI Global.

Rahimi, S., Almond, R. G., & Shute, V. J. (2023b). Stealth assessments' technical architecture. In M. P. McCreery & S. K. Krach (Eds.), *Games as Stealth Assessments* (pp. 61–80). IGI Global. https://doi.org/10.4018/979-8-3693-0568-3.ch003

Rahimi, S., Almond, R., Ramírez-Salgadoa, A., Wusylkoa, C., Weisberg, L., Song, Y., Lu, J., Myersa, T., Wanga, B., Wang, X., Francois, M., Moses, J., & Wright, E. (under review). Competency

model development: The backbone of successful stealth assessments. *Submitted to Journal of Computer Assisted Learning*.

Rahimi, S., & Shute, V. J. (2021). First inspire, then instruct to improve students' creativity. *Computers & Education*, *174*, 104312. https://doi.org/10.1016/j.compedu.2021.104312

Rahimi, S., Shute, V. J., Fulwider, C., Bainbridge, K., Kuba, R., Yang, X., Smith, G., Baker, R. S., & D'Mello, S. K. (2022). Timing of learning supports in educational games can impact students' outcomes. *Computers & Education*, *190*(C), 104600. https://doi.org/10.1016/j.compedu.2022.104600

Rahimi, S., Shute, V. J., Kuba, R., Dai, C.-P., Yang, X., Smith, G., & Alonso Fernández, C. (2021). The use and effects of incentive systems on learning and performance in educational games. *Computers & Education*, *165*, 1–17. https://doi.org/10.1016/j.compedu.2021.104135

Rahimi, S., Shute, V., Khodabandelou, R., Kuba, R., Babaee, M., & Esmaeiligoujar, S. (2023). Stealth assessment: A systematic review of the literature. In P. Blikstein, J. Van Aalst, R. Kizito, & K. Brennan (Eds.), *Proceedings of the 17th International Conference of the Learning Sciences—ICLS 2023* (pp. 1977–1978). https://doi.org/10.22318/icls2023.395429

Rahimi, S., Shute, V., & Zhang, Q. (2021). The effects of game and student characteristics on persistence in educational games: A hierarchical linear modeling approach. *International Journal of Technology in Education and Science*, *5*(2), 141–165. https://doi.org/10.46328/ijtes.118

Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In *Intrinsic and extrinsic motivation* (pp. 373–404). Elsevier.

Sabourin, J. L., & Lester, J. C. (2013). Affect and engagement in Game-BasedLearning environments. *IEEE Transactions on Affective Computing*, *5*(1), 45–56.

Shute, V., Almond, R., & Rahimi, S. (2019). *Physics Playground* (1.3) [Computer software]. https://pluto.coe.fsu.edu/ppteam/pp-links/

Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 139–187). Lawrence Erlbaum Associates, Taylor & Francis Group.

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

Shute, V. J. (2023). History of stealth assessment and a peek into its future. In M. P. McCreery & S. K. Krach (Eds.), *Games as Stealth Assessments* (p. 31 pages). DIO Press.

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives* (pp. 43–58). Springer New York. https://doi.org/10.1007/978-1-4614-3546-4_4

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the Science of Assessment. *Educational Assessment*, *21*(1), 34–59. https://doi.org/10.1080/10627197.2015.1127752

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education: Computer-based assessment for learning. *Journal of Computer Assisted Learning*, *33*(1), 1–19. https://doi.org/10.1111/jcal.12172

Shute, V. J., & Rahimi, S. (2020). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C., Kuba, 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*, *37*(1). https://doi.org/10.1111/jcal.12473

Shute, V. J., Smith, G., Kuba, R., Dai, C.-P., Rahimi, S., Liu, Z., & Almond, R. (2020). The design, development, and testing of learning supports for the Physics Playground game. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-020-00196-1

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

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. https://doi.org/10.1016/j.chb.2016.05.047

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach (Ed.), *Adaptive technologies for training and education* (pp. 7–27). Cambridge University Press.

Shute, V., Lu, X., & Rahimi, S. (2021). Stealth assessment. In J. M. Spector (Ed.), *The Routledge Encyclopedia of Education* (p. 9). Taylor & Francis group.

Shute, V., & 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.

Slater, S., Baker, R. S., & Wang, Y. (2020). Iterative feature engineering through text replays of model errors. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020).*

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

van Merriënboer, J. (2023). *Farewell lecture Prof. Dr. Jeroen van Merriënboer*. https://www.youtube.com/live/M3STl3izLJs?si=OkzE6TSdfNLK5P4c

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572. https://doi.org/10.1016/j.chb.2013.06.033

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. Harvard University Press.

Zarraonandía, T., Díaz, P., & Aedo, I. (2016). Modeling games for adaptive and personalized learning. In B. Gros, Kinshuk, & M. Maina (Eds.), *The Future of Ubiquitous Learning: Learning Designs for Emerging Pedagogies* (pp. 217–239). Springer. https://doi.org/10.1007/978-3-662-47724-3_12