Education Tech Research Dev (2024) 72:2417–2441
https://doi.org/10.1007/s11423-023-10232-1

DEVELOPMENT ARTICLE

Stealth assessment: a theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments

Seyedahmad Rahimi[1] · **Valerie J. Shute**[2]

Accepted: 29 April 2023 / Published online: 17 May 2023
© Association for Educational Communications and Technology 2023

Abstract
Research fields related to learning (e.g., educational technology and learning sciences) have historically focused on *what* questions using traditional methods (e.g., comparing different learning tools and methods). New methodologies that are grounded in learning, engagement, and motivational theories are needed to additionally address the *how* questions. Methodologies that use learners' process data shed more light on how learners learn and if the learning tools are effective, compared to methodologies that use just outcome data. In this paper, we discuss stealth assessment—an evidence-based methodology that can be used in technology-rich environments (e.g., games) to assess and support hard-to-measure constructs (e.g., creativity) as well as knowledge acquisition (e.g., physics). We also discuss evidence-centered design (ECD), and present specific steps to design, embed in a digital learning environment, and evaluate a stealth assessment. Additionally, we provide two examples of stealth assessment studies in the context of an educational game called *Physics Playground:* Study 1 illustrates a stealth assessment of creativity and Study 2 describes a stealth assessment of physics understanding and how we used it to make the game adaptive. The purpose of this paper is to provide sufficient detail about stealth assessment to help researchers in the field of educational technology and related fields to adopt this method to assess, foster, and investigate learning processes in various technology-rich environments.

Keywords Stealth assessment · Methodology · Learning environments · Educational technology · Creativity · Physics

✉ Seyedahmad Rahimi
srahimi@ufl.edu

Valerie J. Shute
vshute@admin.fsu.edu

[1] School of Teaching and Learning, College of Education, University of Florida, 0501D, 2821 Norman Hall, PO Box 117048, Gainesville, FL 32611, USA

[2] Educational Psychology & Learning Systems, Florida State University, Tallahassee, USA

Introduction

For decades, researchers in the field of educational technology have been conducting pretest–posttest studies comparing a learning technology with a control (or business-as-usual) condition. But the results often fail to shed light on the reasons how and why the treatment condition affected students' learning (e.g., Reeves & Lin, 2020; Reeves & Oh, 2017). In other words, research fields related to learning (e.g., educational technology and learning sciences) have tended to focus on effectiveness studies (usually by examining the effects of a treatment based on pretest to posttest changes) via traditional methods. New methodologies that are grounded in learning, engagement, and motivational theories are needed that can additionally answer the *how* and *why* questions. Conducting effectiveness studies using traditional methods, researchers have been treating the learning interventions as black boxes. With advances in learning technologies, learning sciences, and psychometrics (Shute et al., 2016a, 2016b), researchers can and should go beyond simply evaluating students' learning before and after they engage with some intervention. Using new methods and technologies for collecting and analyzing students' interaction data and making inferences of learning in real time will allow researchers to test their educational interventions as *glass boxes*.

Apart from the black vs. glass box issue above, traditional comparison methodologies are not able to capture how learners acquire, express, and improve such hard-to-measure competencies like creativity, problem solving, persistence, collaboration, and systems thinking (Shute & Wang, 2016). For instance, asking learners about their problem-solving skills is not as reliable and valid as seeing learners going through the problem-solving stages when they attempt to solve a game level. Advances in new technologies and methods can help us identify and interpret what students know and can do as they interact with technology-rich learning environments (e.g., digital games).

Glass-box approaches can provide researchers with a wealth of information regarding a person's learning trajectory which can be used, in real-time, to assess learners' competencies and adapt the learning environment to fit learners' needs (e.g., adapt the difficulty of challenges, or provide targeted feedback). Such adaptivity is closely linked to learning, engagement, and motivation theories. Learning and engagement theories—such as the zone of proximal development (Vygotsky, 1978) and flow (Csikszentmihalyi, 1990)—suggest that challenges in a learning environment should match learners' knowledge and ability. Moreover, motivation theories and models such as self-determination theory (SDT; Deci & Ryan, 2012) and the ARCS model (Keller, 1987) indicate that learners' mastery of and confidence in what they know and can do increases their motivation and effort they are willing to put in the learning experience. Learning environments that maximize students' learning via ongoing formative assessments and delivery of appropriate learning supports, are environments that can promote learners' motivation. With the advances in technology, as well as in the learning and assessment sciences (Shute & Rahimi, 2017; Shute et al., 2016a, 2016b), educators can develop learning environments that accurately assess and support students' knowledge, skills, dispositions, and other attributes. As mentioned above, such learning environments use the real-time assessment estimates of students' competency levels to adapt their challenges to students' ability levels or to provide tailored supports to maximize student learning. The real-time assessment of learning can also be used for research purposes (e.g., investigating learners' processes of learning).

In this paper, we (1) discuss an innovative methodology called stealth assessment (Shute, 2011), (2) elaborate the underlying design framework of stealth assessment (i.e.,

evidence-centered design; ECD; Mislevy et al., 2003) with its psychometrics properties (i.e., validity, reliability, and fairness), (3) walk the readers through the steps for designing and developing a stealth assessment, and (4) illustrate two examples of stealth assessment. The purpose of this paper is to provide sufficient detail about stealth assessment to help researchers in the field of educational technology adopt this method to assess, foster, and investigate learning in various technology-rich environments. We begin by a general description of what stealth assessment is.

Background

Stealth assessment

Stealth assessment (Shute, 2011, 2023) uses technology-rich environments (e.g., digital games, virtual reality, or other digital simulations) as vehicles for assessing and fostering learner' various competencies (we call those competencies *unobservables* as they can be latent traits). Technology-rich learning environments, especially digital games, have many affordances for learners to show what they know and can do, and for researchers to collect data in real-time (Shute, 2023). The primary goal of stealth assessment is to seamlessly blur the boundaries among gameplay (if the assessment is embedded in a digital game), learning, and assessment in an unobtrusive way. To that end, stealth assessment machinery continuously collects learner data and estimates their cognitive and non-cognitive competencies during the gameplay.

The main job of stealth assessment is to serve as a type of formative assessment (assessment for learning). That is, stealth assessment collects data in real-time and estimates students' learning as they learn, thus researchers can make inferences about how people learn in the process. The future of stealth assessment should include data sources other than interaction data (i.e., multimodal data sources) to help researchers/educators make accurate estimations about how people learn in environments equipped with stealth assessment.

As shown in Fig. 1, with stealth assessment machinery in place, learners become engaged in playing or interacting with the digital learning environment (e.g., a digital game). The system continuously captures interaction data in logfiles. Then, the stealth assessment machinery identifies data that show evidence for the learner's competencies (we call those relevant data *observables* as we can see a learner perform them while playing a digital game). A statistical modeling approach then accumulates the evidence (observables) in real time into a learner model. The evidence accumulation process can be accomplished by using a simple count of observable frequencies, or by more sophisticated statistical modeling, e.g., Bayesian network modeling or Item Response Theory (IRT).

The more a learner interacts with the digital learning environment the more accurate the stealth assessment estimates of the learner's competencies will become. Stealth assessment estimates can be used by the digital learning environment for various purposes. For instance, in a digital game context, the difficulty of the challenges can be adjusted to match the learner's current ability level to facilitate the flow state (Csikszentmihalyi, 1990) and place the learner's experience within their zone of proximal development (Vygotsky, 1978). Adaptive learning systems such as Intelligent Tutoring Systems (ITSs) have used a similar approach using student models and task adaptation (e.g., Shute, 1995; Brusilovsky, 2002; Phobun & Vicheanpanya, 2010). Alternatively, tailored cognitive and/or affective supports can be delivered based on the stealth assessment estimates. Stealth assessment, as the name implies, is unobtrusive and ongoing. The intention is to not disrupt engagement

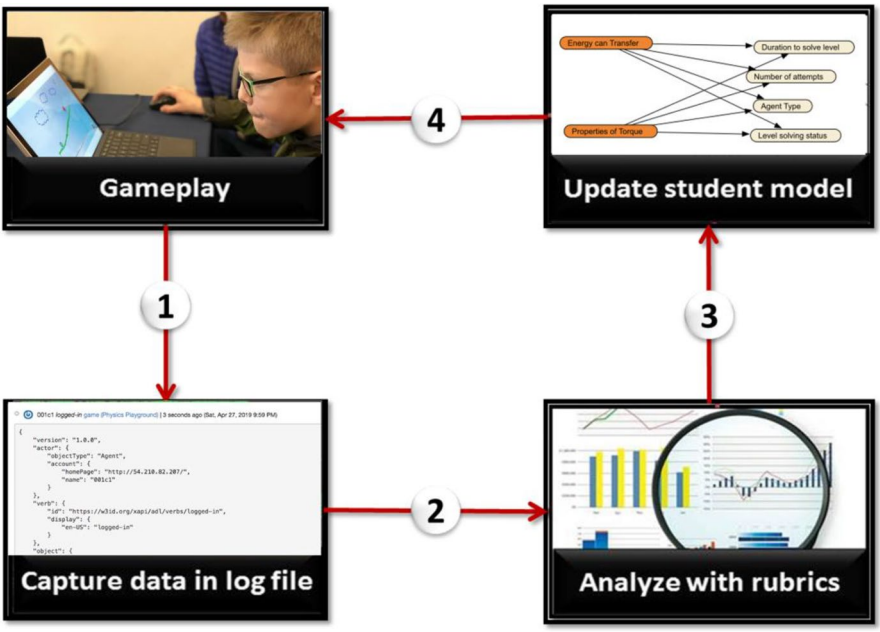

Fig. 1 Stealth assessment process

with the learning experience and to effectively blur the distinction between learning and assessment.

In a systematic review of studies that used stealth assessment during the past decade, we have identified about 100 studies (see Rahimi et al., 2023a for more details). These studies (see Table 1), with participants from third grade to adults, were placed into low, medium, and high-fidelity categories in terms of following the steps of stealth assessment (shown in Table 2). The competencies that were assessed among those 100 studies included: (a) *hard-to-measure competencies* such as creativity, persistence, problem solving, computational thinking, risk taking, safety and emergency readiness; and (b) *knowledge and skills acquisition* such as mathematics, physics, genetics, geometry, reading and writing, and ratio and proportional reasoning (Table 1). This list of competencies that can be assessed using stealth assessment shows the power and generalizability of this evidence-based method.

In addition to testing different competencies in the reviewed list of stealth assessment studies, various fields of study have adopted stealth assessment. Some of these include health and medical, computer science, AI in education, educational technology, learning sciences, and bioengineering. The diversity of fields leveraging stealth assessment is promising. For instance, some of the reviewed studies have attempted to modify the way stealth assessment is created and embedded in digital learning environments (e.g., automatizing some of the processes of ECD using machine learning). Additionally, although stealth assessment can be used for summative purposes, it was not originally intended for that. As shown in Table 1, several studies we reviewed used stealth assessment for summative purposes. Our view is that the real power of stealth assessment, however, is its formative function—measuring and *enhancing learning*.

One important feature of stealth assessment includes the design of its various models—particularly the competency and evidence models. We have historically (e.g., Shute et al.,

Table 1 Selected sample of the 100 studies that used stealth assessment in the past 10 years

First Author (Year)	Field	Game/LE	Competency	Edu. Level	N	Use
DeRosier, M. E (2012)	Health	Zoo U	Social Skills	3rd & 4th g	187	V
Shute, V. J (2013)	ET/LS	PP	Physics	8th & 9th g	154	V
Ventura, M (2013)	ET/LS	PP	Persistence	8th & 9th g	154	S
DiCerbo, K. E (2014)	ET/LS	Poptropica	Persistence	6 to 14 y	892	V
Halverson, R (2014)	ET/LS	Progenitor X	Stem-cell science	M	110	V
Min, W (2015)	CS/AI	ENGAGE	CT	M	182	V
Capuano, N (2015)	IE, EE, AM	VR	Safety	P & SE	45	D
Snow, E., L (2015)	ET/LS	iSTART-2	Students' Agency	College	70	V, R
Kiili & Ketamo (2018)	ET/LS	Semideus	Fraction	6th g	51	V
Shute, V. J (2016)	ET/LS	Use Your Brainz	Problem solving	7th g	47	V
Chin, D. B (2016)	ET/LS	Stroylet	Data literacy	10th g	93	V
Snow, E., L (2016)	ET/LS	iSTART-2	Self-explanation ability	H	40	V
DiCerbo, K. E (2017)	ET/LS	Insight Learning	Geometry	3rd g	131	D
Antoniou, P. E (2017)	Health	Serious talk	Tech acceptance	Adults	21	S
Akram, B (2018)	CS/AI	ENGAGE	Problem solving	M	244	V
DeRosier, M. E (2018)	Health	Zoo U	Social emotional skills	3rd & 4th g	270	V, R
Mayer, I (2018)	AS	TEAMUP	Team Quality	Adults	424	V, R
Georgiadis, K (2019)	ET/LS	abcdeSIM	Medical Caring	Adults	267	V
Smith, G (2019)	ET/LS	Variant: Limits	Calculus	College	148	V
de-Juan-Ripoll, C (2020)	BE	Spheres & Shield	Risk Taking	Adults	38	V
Shute, V. J (2020)	ET/LS	PP	Physics	9th to 11th g	263	V
Yang, D (2020)	ET/LS	Wke-Book	Reading	3rd to 5th g	573	V
Chen, F (2020)	ET/LS	Raging Skie	Weather Phenomena	5th g	460	V
Henderson, N (2020)	CS/AI	Geniventure	Genetics	M & H	462	S
Shute, V. J (2021)	ET/LS	PP	Creativity	9th & 8th g	167	V
Verma, V (2021)	ET/LS	Chemo-o-crypt	Chemistry	College	107	F, A
Gupta, A (2021)	CS/AI	Crystal Island	Microbiology	M	119	S

ET educational technology, *LS* learning sciences, *CS* computer sciences, *AI* artificial intelligence, *IE* information engineering, *EE* electric engineering, *AM* applied mathematics, *AS* applied sciences, *BE* bioengineering, *PP* physics playground, *LE* learning environment, *VR* virtual reality, *PS* problem solving, *CA* cyberbullying awareness, *CT* computational thinking, *E* elementary school, *P* primary school, *SE* secondary school, *M* middle school, *H* high school, *V* validation, *S* summative, *D* design, *R* research, *F* formative, *A* adaptivity

2009) used evidence-centered design (ECD; Mislevy et al., 2003) to design and develop our models used in stealth assessment, and then used the stealth assessment estimates from those models to adapt the environment or provide relevant feedback based on current information about a learner. Next, we discuss ECD and its models.

Table 2 Stealth assessment's steps

Step	Qual/Quant	Description	ECD model
1	Qual	Develop competency model of targeted knowledge, skills, or other attributes based on full literature and expert reviews	CM
2	Qual	Determine which game (or learning environment) the stealth assessment will be embedded into	
3	Qual	Delineate a full list of relevant gameplay actions/indicators that serve as evidence to inform the CM and its facets	EM
4	Qual	Create new tasks in the game, if necessary	TM
5	Qual	Create a Q-matrix (spreadsheet) to link actions/indicators/game levels to relevant facets of target competencies	EM & AM
6	Qual	Decide on the collection of activities (both assessment tasks and learning activities), rules for navigating between them, and stopping rules for when there is enough information	AM
7	Quant	Determine how to score indicators using classification into discrete categories (e.g., yes/no, very good/good/ok/poor relative to quality of the actions). This becomes the "scoring rules" part of the evidence model	EM
8	Quant	Establish statistical relationships between each indicator and associated levels of the CM variables	EM
9	Qual & Quant	Pilot test scoring model (e.g., BN) and modify parameters	EM
10	Qual & Quant	Validate the stealth assessment with external measures	
11	Quant	Use current competency estimates to adapt game challenges, provide targeted support, or use the estimates to analyze learning patterns/trajectories to answer research questions	

CM competency model, *EM* evidence model, *TM* task model, *AM* assembly model, *Quant* quantitative, *Qual* qualitative

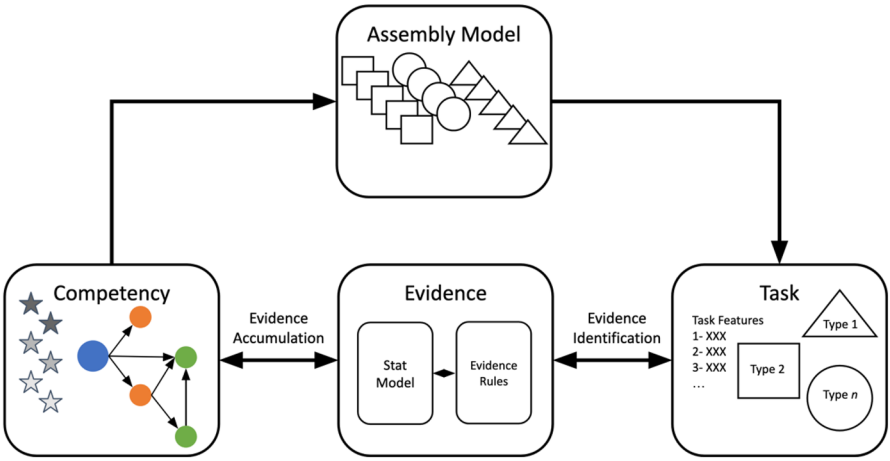

Fig. 2 ECD models

Evidence-centered design (ECD)

As mentioned, stealth assessment uses the ECD framework as the basis for assessment design. ECD includes four core models (Fig. 2). First, the *Competency Model* (CM) defines the competency of interest and its sub-facets (unobservables), their relationships to each other (e.g., prerequisite, part of, correlational; and the strength of the relationship; shown as the arrows connecting the circles in Fig. 1), and claims we want to make about what learners with various levels of competency can do (shown as the stars in Fig. 2). When defining the competency model, researchers respond to the question of *what* to assess. Second, the *Evidence Model* (EM) identifies appropriate indicators (observables) in the game (i.e., establishing the rules of evidence) that provide evidence for the CM variables via statistical linkages (i.e., the statistical model). When defining the EM, researchers answer the question of *how* to assess. Third, the *Task Model* (TM) involves the creation of various task types that can elicit the evidence needed for the evidence model. The goal of a task model (when designing a game from scratch) is to facilitate the authoring process of the tasks. One can think of a TM as a template from which a game developer can instantiate as many instances as needed. By defining the TM, researchers answer the question of *where* to assess. Finally, the *Assembly Model* (AM) allows researchers to arrange various tasks together with various difficulty levels, sufficient per competency, to be delivered to the learners. Moreover, the AM includes rules for adaptivity, personalization of the learning supports and other features of the game environment. Researchers respond to the question of *how much to assess* by defining the AM.

When these four core models are created and implemented in a digital-rich learning environment (e.g., an educational game), observations made in the context of a person interacting with assessment tasks provide evidence of current competency level permitting the system to update competency estimates in real-time.

In general, there are two ways to design a stealth assessment: (1) creating and embedding a stealth assessment into an existing game (e.g., Shute et al., 2016a, 2016b), and (2) creating a game from scratch (e.g., Shute et al., 2020). Regarding the first way—embedding stealth assessment into an existing game—this has some limitations, as it is not easy

to gain access to a game's source code necessary to integrate what is needed for a stealth assessment in the game. Also, because expansion of the game to accommodate all sub-facets of a competency model is not feasible, researchers would need to compromise the assessment of some facets of a CM and only include facets for which the existing game has good indicators. Regarding the creation of a game from scratch and concurrently designing and developing a stealth assessment in the game, while this might be more resource intensive than the first option, it provides greater opportunities to the researchers to assess and foster what they wish (Shute et al., 2017; Smith et al., 2023 for more on this).

In both cases, the same steps can be employed to design and develop a stealth assessment (see Shute et al., 2021). Each step shown in Table 2 is linked to at least one ECD model except Step 11. Also, the qualitative and quantitative nature of each step is identified. It is important to note that these steps are not intended to be followed in a linear fashion. Instead, stealth assessment design, development, and testing should be done iteratively with the help of various experts (e.g., learning scientists, game designers and developers, psychometricians, subject matter experts) working together in a team.

In this paper, we expanded Step 11 by adding the following: "… or use the estimates to analyze learning patterns/trajectories to answer research questions." The main goal of stealth assessment is to improve learning, however, since stealth assessment estimates are continuously computed in the background, researchers can leverage these data to answer various research questions related to the processes of learning. Following these steps, researchers can develop stealth assessments that are psychometrically sound—discussed next.

Psychometric properties of stealth assessment: validity, reliability, and fairness

Any assessment, including stealth assessment, is expected to be psychometrically sound (i.e., to be valid, reliable, and fair; Messick, 1994; Rahimi et al., 2023b; Shute, 2009). In short, *validity* refers to the extent to which an assessment is assessing what it claims to assess; *reliability* refers to the consistency of an assessment in different times and places; and *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).

Rahimi et al. (2023b) described various recommendations to improve the psychometric properties of a stealth assessment. For instance, consulting with a subject matter expert and conducting a thorough literature review when developing a competency model improves the construct and face validity of a stealth assessment. Moreover, a stealth assessment that assesses learners' abilities in an authentic learning environment (e.g., an educational game) tends to have a high level of ecological validity (i.e., the extent to which one can generalize what was found to real-life situations). Additionally, including a diverse group of people in terms of expertise, ethnicity, and gender on the design team can improve the fairness of a stealth assessment.

Once a stealth assessment in designed and developed, researchers should first attempt to evaluate the psychometric properties of a stealth assessment via validation studies. The most common type of validation reported in various stealth assessment studies is convergent validity (Rahimi et al., 2023a)—i.e., evaluating the correlations among stealth assessment estimates and external measures that assess same competency. If the stealth assessment shows a positive correlation with an external measure, this suggests that the stealth assessment is measuring what it claims to measure. Most of the correlational analyses reported in the literature yield values between 0.1 and 0.6 (DiCerbo et al., 2016; Rahimi

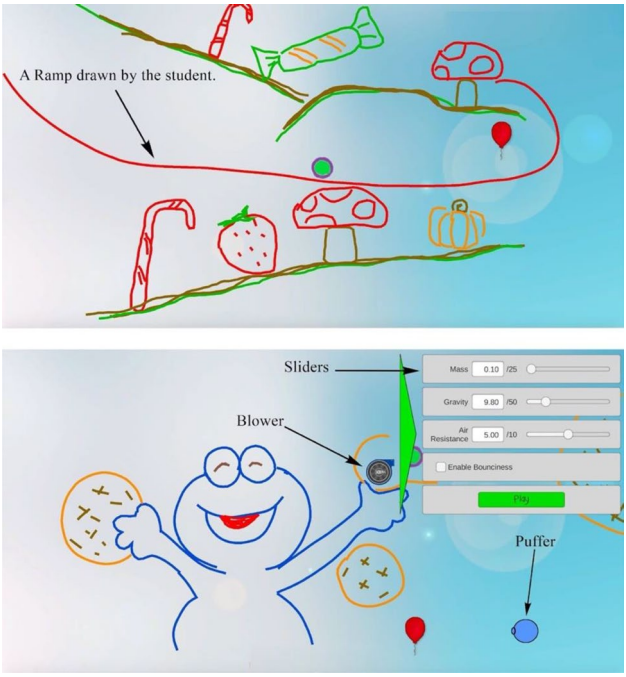

Fig. 3 Sketching level (top) and manipulation level (bottom) in *PP*

et al., 2023b). We do not see this as a weakness of stealth assessment since the conditions of traditional assessments and stealth assessment are different. However, in the case of non-significant or very small correlations, we suggest that researchers improve their stealth assessment by (1) expanding their task models to capture more evidence for the CM variables if needed, (2) identifying poor observables and replacing them with more robust ones, and (3) revising the parameters of the statistical modeling used (e.g., tweaking the conditional probability table entries after a pilot and validation study).

In the following sections, we discuss two studies that we conducted that illustrate the steps listed in Table 1. Before presenting those studies, we first introduce the game that we designed and used in both studies—*Physics Playground*.

Physics playground

A 2-dimensional computer-based game, *Physics Playground* (*PP*; Shute et al., 2019) was developed to assess and support middle-to-high-school students' understanding of Newtonian physics. This puzzle game has a clear and simple goal—i.e., hitting a red balloon using a green ball. Initially, *PP* only included one type of task, referred to as *sketching levels* (Fig. 3 top) where learners draw lines on the screen using a stylus or mouse, and use tools in the game (e.g., pins) to create simple machines (or *agents* of force and motion; e.g., lever, ramp, springboard, and pendulum) to guide the ball to the balloon. The most recent version of *PP* includes a new task type called *manipulation levels* (Fig. 3 bottom) where

the only way to solve the levels is to interact with three sliders (i.e., mass of the ball, gravity, and air resistance), or to enable the bounciness option, and/or use puffers and/or blowers (if included in a level). Drawing is disabled in this task type. Using the level editor of *PP*, non-technical members of the team allowed us to design over 100 game levels.

Study 1: stealth assessment of creativity

Background and method

The first study we discuss is a stealth assessment of creativity in *PP* (see Shute & Rahimi, 2020). Creativity is one of the most important skills for success in our complex world with all of its everchanging and unexpected problems (Glaveanu et al., 2020; Gray, 2016; Resnick, 2018). The type of creativity we are referring to here is not necessarily artistic creativity; rather, it is the creativity people need in their everyday life, or *little-c* creativity (Kaufman & Beghetto, 2009; Richards, 2010). To improve people's creativity, we first need to accurately assess it. However, creativity is a hard-to-measure construct (Shute & Wang, 2016) and traditional, self-report measures are not able to assess it accurately or fully. Therefore, we used stealth assessment to automatically assess students' creativity when they solved game levels in *PP* (Step 2 in Table 1). In this study, we asked the following research questions:

A) Is our stealth assessment estimate of creativity valid (i.e., does it correlate with other external measures of creativity)?
B) Does creativity predict in-game performance (i.e., number of levels solved, number of gold and silver coins earned)?
C) Does creativity predict enjoyment of the game?
D) Does creativity predict physics learning?

First, we conducted a literature review to define our competency model of creativity in the context of *PP* (Step 1 in Table 1). In the creativity research literature, one of the most agreed upon definitions of creativity defines it as any product (i.e., idea, solution, artwork, writing) that is both novel (original) and appropriate (applicable) (Runco & Acar, 2012; Runco & Jaeger, 2012). Moreover, the four-p model of creativity (Rhodes, 1961) suggests that there are four approaches when it comes to researching creativity: *person* (i.e., personalities, dispositions), *process* (i.e., steps one takes to come up with a creative product), *press* (i.e., the environment one interacts with or lives in that leads to a creative product), and *product* (i.e., the outcome of a creative process). Other operationalizations define creative thinking as divergent thinking which includes *fluency* (i.e., the ability to produce many relevant ideas); *originality* (i.e., the ability to produce ideas that are statistically rare); *flexibility* (i.e., the number of categories or themes used when solving a problem or the ability to come up with relevant ideas from different categories or themes); and *elaboration*[1] (i.e., the ability to implement and expand on an idea in detail and high quality). In this study, we included fluency, flexibility, and originality as the sub-facets of gameplay creativity, shown

[1] Because the facet of *elaboration* generally overlaps with the other facets (and we could not ascertain unique indicators for it), we excluded it from our model.

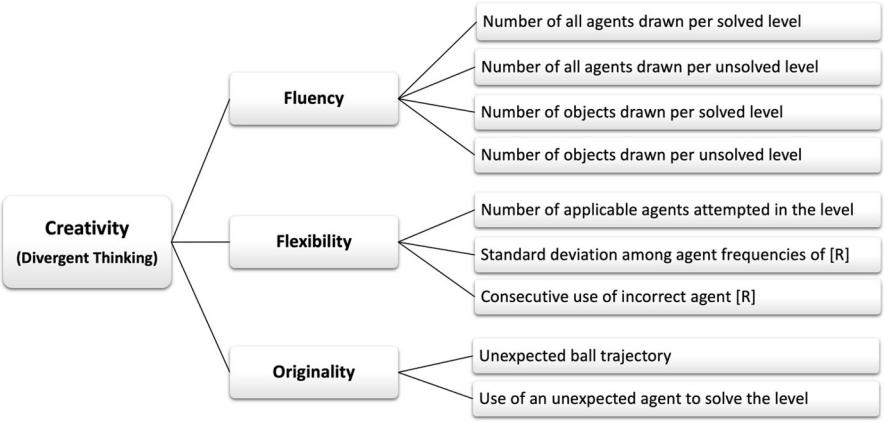

Fig. 4 CM of creativity with evidence model variables (indicators) on the right. *Note*. *R indicates reverse coding (for negative evidence)

in Fig. 4. We also identified appropriate EM variables (i.e., observables) as indicators in *PP* per sub-facet (Step 3 in Table 2).

Participants, research design

The participants of this study consisted of 167 8th and 9th graders (76 male and 91 female) from a K-12 school in Florida. Upon the completion of the study, each student received a $25 gift card. We used a one-group, pretest–posttest research design. The total gameplay time was about 4 h (across six 45-min sessions in a week). Thirty computers, in one of the school's two computer labs, were used for this study. Separators were used between the computers to make sure students did not talk to each other during gameplay. All students played the same version of *PP*.

After introducing the study, researchers administered a demographic questionnaire about students' age, gender, grade, and gameplay frequency. Afterwards, an online physics pretest, followed by some performance-based measures of creativity (i.e., identifying alternative uses for three common tools—discussed later) were administered. Upon completing the pretest battery of measures, the researchers introduced *PP* to the students. To encourage students to pay attention during gameplay, they were told that the student with the most gold coins at the end of the study would receive an extra $25 gift card.

The first session of gameplay started by having students complete the agent-tutorial videos. Students could start playing the game when they finished watching the videos. Researchers instructed students to start by playing levels in Playground 1 (i.e., mostly easy levels), and then they could move on to any levels in any playground they wanted (students were informed that the difficulty of levels across the 7 playgrounds increased incrementally). During gameplay, students were told that they could watch the tutorial videos if they were struggling in a level.

Measures[2]

Physics Test

Working with a physics expert, we created 24 multiple-choice items, counterbalanced between two equivalent forms (Form A and Form B), and used for pretest and posttest in the study. Each form included two items for each of our six main physics concepts. The tests measured students' understanding of Newtonian physics. The Cronbach's α for the physics test Form A was 0.72 and 0.73 for Form B.

Game enjoyment

We created a scale with questions about how the students enjoyed playing the game using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Two examples of game-enjoyment items include: "I enjoyed playing *Physics Playground*" and "I would play this game in my spare time" using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Stealth assessment of creativity

We developed three different stealth assessment measures of creativity—for fluency, flexibility, and originality as shown in Fig. 4. To estimate *fluency*, we identified different variables such as "Number of drawn objects per solved problem," and "Number of agents used in a problem." For *flexibility*, we collected data on variables such as, "number of correct agents attempted in the problem," and "standard deviation among frequencies of agent used" [per session; reverse coded]. Finally, for *originality*, we captured the x, y coordinates of each student's solution trajectory (i.e., the path the ball took from origin to hitting the balloon) in the log files. We also had, for each level, the expected solutions' x, y coordinates. Therefore, we could compute the difference between the student's and expected solution path—the larger the difference the more original the solution.

After establishing the relevant observables for each of the three facets of creativity, we created a Bayesian network (BN) to estimate students' creativity in *PP* (see Fig. 5). As students play and provide evidence (or counter evidence) for each facet of creativity (i.e., fluency, flexibility, and originality), the parent node gets updated. That is, at the end of each level the log files are automatically parsed, observables are identified and scored (using the scoring rules from the EM), and the scores are accumulated by the Bayes net (BN) for each student. The BN automatically calculates the low, medium, and high probability estimates for fluency, flexibility, originality, and finally for overall creativity, per student. For example, as shown in Fig. 5, the trajectory of the solution for a given level has been scored as "rare" which generated a high probability for originality, and in turn, for overall creativity. Considering other pieces of evidence coming into the BN in this example, the high,

[2] We also had students design their own game levels and we scored the creativity of their levels [see Shute and Rahimi (2021) for a full report]. We focus on the in-game stealth assessment and external measure of creativity in this paper.

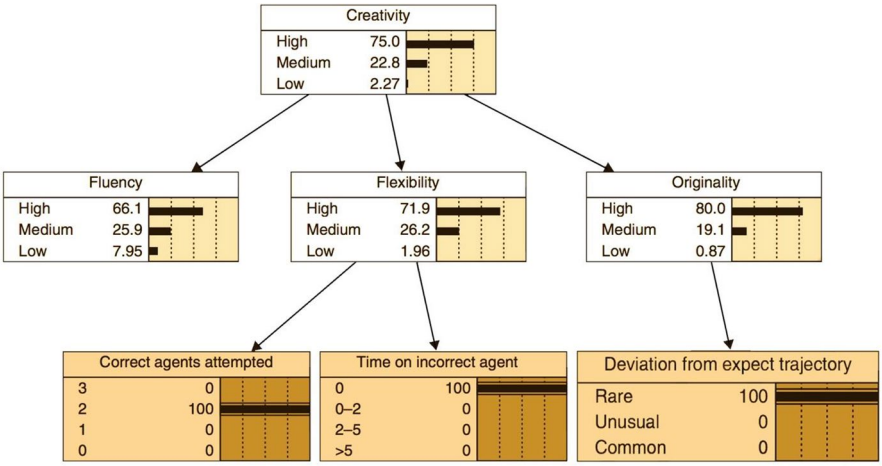

Fig. 5 Creativity BN with its child nodes and example indicators for one level in *PP*

medium, and low probabilities are calculated as $P(\text{Creativity}=\text{high}|\text{evidence})=0.75$, $P(\text{Creativity}=\text{medium}|\text{evidence})=0.23$, and $P(\text{Creativity}=\text{low}|\text{evidence})=0.02$.

These estimates become increasingly more accurate as additional data is absorbed into the BN at the end of each level. For our analyses, we computed a single value for the parent node (creativity) per student. That is, the stealth assessment estimate consisted of three probabilities (i.e., high, medium, and low). We assigned numeric values to the three states and computed the expected value. This Expected A Posteriori (EAP) value can also be expressed as, $P(\theta_{ij}=\text{High})-P(\theta_{ij}=\text{Low})$, where θ_{ij} is the value for Student i on Competency j, and $1*P(\text{High})+0*P(\text{Med})+-1*P(\text{Low})=P(\text{High})-P(\text{Low})$. This results in a scale from − 1 to 1.

External measures of creativity

To validate our stealth assessment of creativity, we used Wallach and Kogan's Creativity Test (Wallach & Kogan, 1965) consisting of three *alternative-uses* test items (e.g., "How many different uses can you list for a rubber band?") with a maximum of eight responses. Students had one minute to compile their lists with as many answers as they could. When time was up, they circled their top two most creative responses. We scored students' responses for fluency (i.e., number of responses per item; Cronbach's $\alpha=0.9$) and originality (i.e., two most creative responses; Cronbach's $\alpha=0.8$).

In-game measures of game performance

Parsing the log files, we created multiple variables indicating students' game performance: (1) *the number of gold coins*: depending on the levels' difficulty, we assigned the minimum number of objects needed to solve the level. When a student solved a level at or under the minimum number of objects, a gold coin was given to the student; (2) *the number of silver coins*: when a solution used more objects than the minimum number of objects to solve the level, a silver coin was given to the student; and (3) *the number of levels solved*: which indicates the total number of levels a student solved throughout all gameplay sessions.

Results

Research question 1 concerned validity. To establish the convergent validity of our stealth assessment of creativity, we conducted several correlational analyses. Results showed significant correlations between our stealth assessment measure of creativity and the *Alternative Uses test—fluency* ($r=0.18$, $p=0.02$); and (b) *Alternative Uses test—originality* ($r=0.18$, $p=0.02$). Therefore, the significant correlations suggest that our stealth assessment measure of creativity is valid. However, there is room for improvement.

To answer research question 2 on the ability of our stealth assessment measures to predict in-game performance, we conducted three separate multiple regression analyses controlling for the pretest score (using stealth assessment estimates for research purposes—Step 11 in Table 2). Results showed that, controlling for pretest, our stealth assessment estimate of creativity was not a significant predictor of gold coins earned ($\beta_{creativity}=0.11$, $t=1.38$, $p=0.17$; $\beta_{pretest}=0.31$, $t=3.92$, $p=0.17$; $F(2, 163)=12.57$, $p<0.001$, $R^2=0.12$). However, it was a significant predictor of silver coins earned ($\beta_{creativity}=0.27$, $t=3.29$, $p=0.001$; $\beta_{pretest}=-0.09$, $t=-1.04$, $p=0.30$; $F(2, 163)=5.44$, $p=0.005$, $R^2=0.05$), and total number of levels solved ($\beta_{creativity}=0.39$, $t=5.20$, $p<0.001$; $\beta_{pretest}=0.16$, $t=2.11$, $p=0.04$; $F(2, 163)=23.38$, $p<0.001$, $R^2=0.21$). This indicates that more creative students earned more silver coins and completed more game levels than less creative students.

For research question 3, regarding the relationship between creativity and enjoyment of the game, we computed a simple regression analysis with game-enjoyment score (i.e., the average of the two items related to game enjoyment) as the dependent variable and the stealth assessment estimate of creativity as the independent variable. Results showed that our creativity estimate significantly predicted students' game enjoyment ($\beta=0.21$, $F(1, 152)=6.73$, $p=0.01$, $R^2=0.04$) showing that more creative students enjoyed playing the game more than less creative ones.

Finally, to address research question 4, testing whether our stealth assessment estimate of creativity predicts learning physics from the game, we conducted another simple regression analysis with posttest score as the dependent variable and our in-game creativity estimate as the independent variable. Results showed that our creativity estimate significantly predicted students' posttest scores ($\beta=0.19$, $F(1, 152)=5.64$, $p=0.02$, $R^2=0.04$) showing the positive yet small relationship between creativity and learning. However, when controlling for the pretest, stealth assessment creativity estimates were not a significant predictor of the posttest scores ($\beta=-0.06$, $t=-0.81$, $p=0.42$).

Brief discussion

In Study 1, we validated our stealth assessment of creativity and used the in-game estimates to create regression models to predict game performance, enjoyment, and learning. We did not use the stealth assessment estimates in real-time to make the game adaptive or provide appropriate supports to improve students' creativity. In fact, such validation studies should be carried out before using the stealth assessment estimates for real-time changes to the game environment. Herein, we showed that the stealth assessment estimates of creativity can be used as a valuable source of process data illustrating a learner's trajectory of learning and to answer various research questions. While using stealth assessment estimates to improve learners' creativity in real-time is one of the future directions we are taking our research, we now discuss another study where we did use the stealth assessment estimates for adaptivity and triggering learning supports.

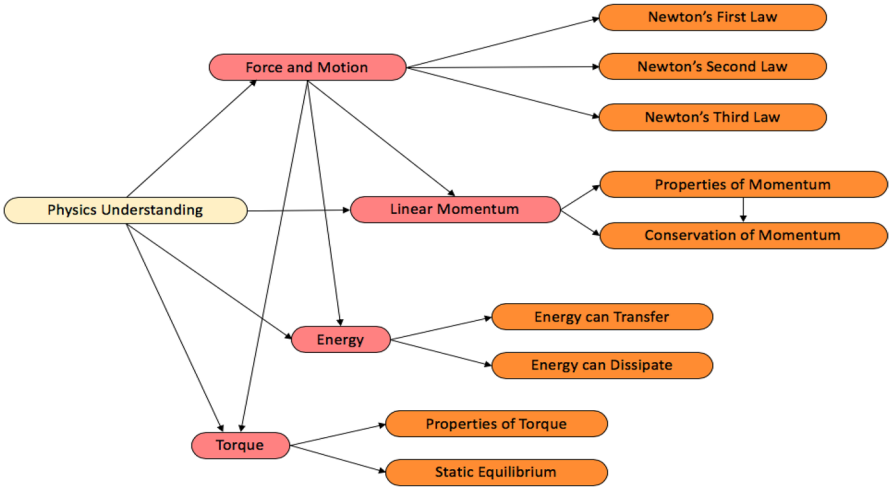

Fig. 6 Physics understanding competency model

Study 2: stealth assessment of physics understanding

Background and method

In Study 2, we pursued three main goals: (a) redesigning and validating a stealth assessment of conceptual physics understanding, (b) designing, developing, and evaluating the effects of including cognitive learning supports in multiple formats (e.g., short videos, mini games, text, etc.), and (c) using the stealth assessment estimates of students' physics understanding to create and evaluate an adaptive version of the game (i.e., showing a game level to the learner that matches their level of physics understanding; starting the next level with a learning support if the learner's competency estimate was low). To achieve these goals, we spent time going through the steps included in Table 2 in an iterative process. First, we consulted with our physics experts to develop a new competency model of physics understanding (Fig. 6; see Almond et al., 2017; Rahimi et al., 2023b for more detial on this process), in press a for more detial on this process). Developing the CM was Step 2 in Table 1.

Study 2's CM is more complex than the CM described in Study 1 as it has three levels of CM variables: physics understanding which is the high-level variable; force and motion, linear momentum, energy, and torque that are mid-level variables; and other variables (e.g., Newton's laws) that are lower-level variables. The indicators (observables) in *PP* are directly linked to the low-level variables (not included in Fig. 6). Figure 7 shows the in-game indicators we included for manipulation levels relative to Newton's Third Law of force and motion (the primary concept) and Energy can Dissipate (the secondary concept). Sketching levels and other manipulation levels that were not connected to Newton's Third Law had different indicators.

In the process of creating the CM for physics understanding we came up with two different CMs and showed them to our experts. They selected the CM shown in Fig. 6 as the best depiction of the content. During an earlier phase of Study 2, *PP* had only one level type (i.e., sketching levels) and we had about 100 sketching game levels. Step 5 in Table 2 suggests using a tool called Q-Matrix to link indicators or game levels to the CM variables.

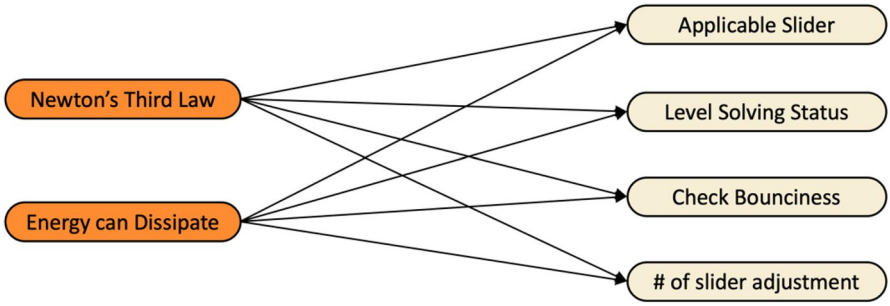

Fig. 7 Indicators for the manipulation levels with Newton's Third Law and Energy can Dissipate as primary and secondary concepts respectively

Table 3 Example of the Q-Matrix

	NFL	NSL	NTL	POM	COM	ECT	ECD	POT	Equil	GM	PU	Comp
Level 1	1	2	0	0	0	0	0	0	0	2	3	5
Level 2	0	0	0	0	0	1	0	2	0	3	5	8
…	…	…	…	…	…	…	…	…	…	…	…	…
Level n	1	2	0	0	0	0	0	0	0	2	2	4

NFL Newton's First Law; *NSL* Newton's Second Law; *NTL* Newton's Third Law; *POM* Properties of Momentum; *COM* Conservation of Momentum; *ECT* Energy Can Transfer; *ECD* Energy Can Dissipate; *POT* Properties of Torque; *Equil.* Statis Equilibrium; *GM* Game Mechanics difficulty; *PU* Physics Understanding Difficulty; *Comp.* Composite difficulty

This tool allows researchers to examine the coverage of game levels or indicators per CM variable.

Through examination of a Q-Matrix, researchers can identify any need to modify (expand or curtail) the TM. Table 3 shows a smaller version of the Q-Matrix that we used for Study 2. In this Q-matrix, a 1 indicates the primary competency and a 2 denotes the secondary competency associated with each game level (e.g., Level 1 has Newton's First Law as its primary competency and Newton's Second Law as its secondary competency). Additionally, the Q-matrix included columns indicating the game mechanics difficulty (GM, ranging from 1 to 5) and physics understanding difficulty (PU, ranging from 1 to 5) indices of each game level. The composite difficulty score was the sum of GM and PU (discussed in more detail later). Summing the number of entries in each column provides a quick estimate of the amount of evidence available for the corresponding competency, towards ensuring complete coverage of the competency model. Including the difficulty information in the Q-matrix allows the assessment designers to ensure that the levels span the relevant depths of competencies of interest and that the game will not be too challenging or too easy. After finalizing our CM for Study 2, we evaluated our game levels to see if we had enough game levels per low-level competencies (e.g., Newton's laws, properties of momentum, conservation of momentum, etc.). We found that the Newton's Second and Third laws did *not* have enough game levels directly related to them. Therefore, we expanded our TM (Step 4 in Table 1) and designed a second level type called manipulation levels.

Fig. 8 Learning supports available in *Show me the Physics* (top) and example of the learning support for an interactive Definition (bottom)

In parallel with expanding our TM, we designed multiple type of learning supports in the game (see Fig. 8). The learning supports included: *Animations* which contained videos presenting physics concepts (e.g., properties of torque) in the game environment relevant to a student's current game level; *Definition* which included physics terms applicable to the game's content and students after students watched a short animation illustrating the term in the game environment (e.g., gravitational potential energy) and complete the term's definition through a fill-in-the-blank, drag-and-drop interaction; *Formula* presents the physics concept's formula (if applicable) and defines the associated variables; *Hewitt video* contains cartoon animations explaining various physics concepts, originally developed by Paul Hewitt, and edited with permission; and *Glossary* support contains brief explanations of a set of physics terms relevant to the game (to read more about the software architecture of *PP*, please see Rahimi et al., 2023c; to read more about a series of studies we conduced to evaluate the effectiveness of the

learning supports in *PP*, please see Bainbridge et al., 2022; Kuba et al., 2021; Rahimi et al., 2021; Rahimi et al., 2022a; Yang et al., 2021a, 2021b; Yang et al., 2022).

We included 91 game levels, and asked the following three research questions in Study 2:

1. Is the stealth assessment of physics understanding valid?
2. Which delivery method of game levels (i.e., adaptive, linear, or free choice) is more effective for improving students' physics understanding when controlling for incoming knowledge?
3. Which type of embedded learning support most effectively enhances learning and game performance?

In the current paper, we only focus on research question 1 and describe how we designed a learner-facing dashboard to show the real-time estimates of competencies to each learner. To read about what we found regarding research questions 2 and 3, please see Shute et al. (2020).

Participants and research design

We recruited 280 9th—11th grade students in a large K-12 school in the southeastern U.S. We included the data from 263 students who had completed both the pretest and posttest, submitted their parental consent forms, and signed the assent form in this study. We had the same number of students self-identify as male ($n=128$) and female ($n=128$), with a wide range of ethnicities. Self-reported ethnicities representing more than 1% of the respondents were: Asian ($n=8$), Black or African American ($n=77$), Hispanic ($n=23$), White (n = 114), Other ($n=7$), Black or African American and White ($n=6$), Black or African American and Hispanic ($n=3$), and Hispanic and White ($n=9$).

Procedure

The experiment spanned six days of class time, comprising six sessions (50 min per session). On Day 1, participants completed a demographic survey and the pretest of physics knowledge (18 items), followed by an introduction to *PP* gameplay. Days 2 to 5 consisted of gameplay throughout the session. Day 6 consisted of gameplay followed by the posttest, the game and learning support satisfaction survey, and receipt of the gift card.

External measure: physics understanding test

To validate our stealth assessment of physics understanding, we created and pilot tested 36 multiple-choice items covering the nine physics competencies in the game, counterbalanced between two equivalent forms for a pretest and posttest (pretest = 18 items, $\alpha=0.77$; posttest = 18 items, $\alpha=0.82$). Each form included two items per competency. These were matched forms, with multiple-choice formatted items accompanied by relevant pictures. The tests have been revised across two years of testing, and the current reliabilities (Cronbach's α values) were: pretest = 0.77; posttest = 0.82. The items were (a) designed in the context of *PP* (i.e., including a video or an image from the game environment), (b) developed with the help of two physics experts, and (c) subjected to several pilot tests before administration in Study 2.

Table 4 Fine-grained validation of stealth assessment estimates

Stealth assessment estimates	Pretest	Posttest
Force and motion	0.29**	0.30**
Linear momentum	0.27**	0.27**
Energy	0.22**	0.35**
Torque	0.14*	0.18**

*$p<0.05$; **$p<0.01$

Results

To address research question 1 concerning validity, we computed the correlation between our overall stealth assessment estimate of "physics understanding" with our external physics test scores. Results showed that both the pretest ($r=0.36$, $p<0.01$) and posttest ($r=0.40$, $p<0.01$) scores significantly correlated with the overall stealth assessment estimate. We were also able to test specific correlations involving each one of our mid-level competency estimates (e.g., Torque) with the score of relevant subscale items on the pretest and posttest (e.g., average score of students on the items related to energy were correlated with their stealth assessment score of energy). Like the overall physics understanding estimate, the mid-level stealth assessment estimates significantly correlated with their associated external measures both on pretest and posttest (see Table 4).

In summary, we found that our stealth assessment of physics understanding was valid—both overall as well as at a more granular, diagnostic level. Note that the BN scores were based only on the experts' original estimates. Refining the model using the data from the field test from Study 2 should yield even better measures of physics competency (Step 9 in Table 2) in future studies. Next, we discuss how we used these estimates in the game.

Using the stealth assessment estimates in PP

We used the validated stealth assessment estimates in the game in two ways: (1) making the game adaptive, and providing learning supports in a personalized way; and (2) showing the estimates to the students in an in-game, student-facing dashboard. One of the conditions in this study was adaptive. We used an algorithm to determine the order to present the levels to best match students' physics understanding competency level. For example, if the stealth assessment estimate of a student's current level of physics understanding of a competency (e.g., Newton's Third Law) was satisfactory (e.g., >0.33), the adaptive algorithm would deliver more difficult levels associated with that competency. If the competency estimate was too low (e.g., <0.33), the algorithm would show a relevant learning support to the student before playing the next level.

Although this assessment method is called *stealth* assessment, the fact that students are being assessed is not hidden from them. To facilitate students' understanding of their progress in playing game levels and their mastery of various physics concepts we designed an in-game dashboard (Fig. 9; see Rahimi & Shute, 2021 for a report on learning analytics dashboard in educational games and how we designed this dashboard

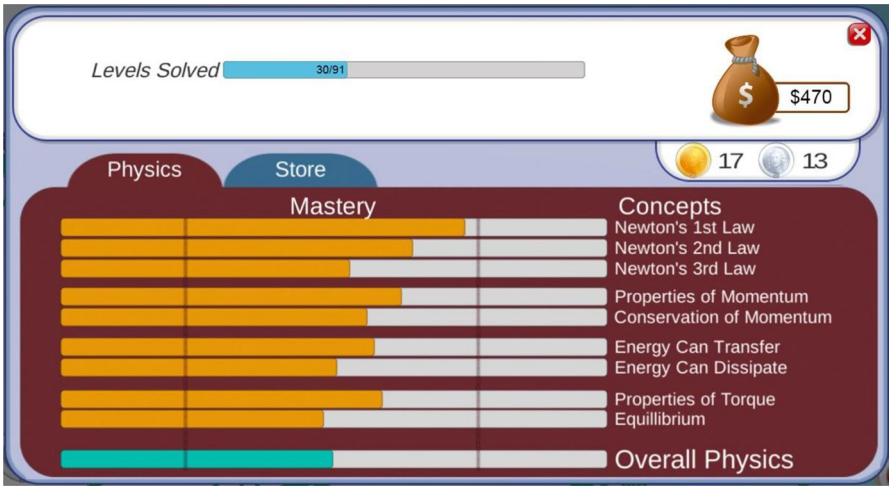

Fig. 9 *My backpack*—students can check their gameplay progress, money balance, and physics understanding

for *PP*). The orange bars show students' level of mastery on different physics concepts and the cyan bar (at the bottom of the dashboard) show students' overall physics understanding. These bars were all connected to students' real-time stealth assessment estimates.

Brief discussion

In summary, Study 2 was the most comprehensive study of stealth assessment we conducted to date. It included (1) designing and developing the stealth assessment machinery based on ECD in the game, (2) designing and developing learning supports, (3) validation of the stealth assessment measure, and (4) using the stealth assessment estimates to make the game adaptive, providing learning supports in a personalized way, and showing learners their progress as they played the game and learned the concepts (the four-stage process shown in Fig. 1). Stealth assessment studies in the literature (some of them shown in Table 1) may or may not include all these components (Rahimi et al., 2023a). However, we emphasize that having all these components operating within a game or other learning environment is what makes the stealth assessment loop complete.

Conclusion

In this paper, we discussed the need for new methodologies that allow researchers to look closely into learning processes. Advances in learning sciences, instructional technologies, and psychometrics have made it possible to design, develop, and evaluate digital learning environments that can collect, analyze, and interpret learners' interaction data in real-time. We introduced stealth assessment as one of the innovative methods that is based on learning, engagement, and motivational theories, and uses a psychometrically sound framework

of ECD. Stealth assessment has been adopted in other fields (e.g., computer science, health, etc.) in the past ten years and we expect to see it expand in the future.

There are two limitations that need to be discussed regarding the design of stealth assessment: (1) the quality of assessment model depends on the information (accuracy) of the indicators (observables); and (2) stealth assessment is a labor-intensive process that requires various types of expertise that may seem difficult to bring together in one project (e.g., computer sciences, learning sciences, psychometrics, instructional design, game design). While these two limitations are valid, to address (1) we recommend that researchers follow the steps of designing stealth assessment in an iterative process. The indicators need to be tested and modified in multiple pilot studies. Later, new indicators may be added to enhance the accuracy of the models. To address (2), researchers in the field of computer science are trying to automate some of the processes of stealth assessment to reduce the labor-intensity of the design process. For example, Min et al. (2015) used deep learning to analyze log data and automatically create accurate evidence models (useful indicators as proxies for learning). In general, a diverse team of experts working toward a common goal is important for a successful stealth assessment design.

So far, most of the applications of this methodology have been limited to research laboratories. However, we expect to see the design and development of learning technologies equipped with stealth assessment at a larger scale (e.g., in schools across the country or even the world and with thousands of learners using these environments). Finally, while we see stealth assessment as a top-down approach, bottom-up approaches (e.g., educational data mining, machine learning) can be used to enhance the accuracy of stealth assessment estimates in future learning environments. For instance, machine learning can be used to identify new patterns in learners' learning and gaming behaviors, as well as new variables (observables) that can be included to improve stealth assessment's accuracy in the future (e.g., Rahimi et al., 2022b). Additionally, machine learning can be used for scoring the appropriate observables. Using methods such as stealth assessment can open many doors to us to improve how learners learn and to provide personalized learning experiences for all learners at scale.

Acknowledgements This work was supported by the US National Science Foundation [award number #037988] and the US Department of Education [award number R305A170376].

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest We wish to confirm that there are no known conflicts of interest associated with this manuscript.

References

Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018). Improving stealth assessment in game-based learning with LSTM-based analytics. *International Conference on Educational Data Mining*. Retrieved from https://par.nsf.gov/biblio/10100664-improving-stealth-assessment-game-based-learning-lstm-based-analytics

Almond, R., Tingir, S., Lu, X., Sun, C., & Rahimi, S. (2017). An elicitation tool for conditional probability tables (CPT) for physics playground. *Uncertainty in Artificial Intelligence*.

Antoniou, P. E., Siountas, A., Zilidou, V. I., & Bamidis, P. D. (2017) Virtual scenarios for stealth assessment of the elderly Perceptions and acceptance of technology-based health and wellness interventions. IEEE

30th International Symposium on Computer-Based Medical Systems (CBMS), https://doi.org/10.1109/CBMS.2017.115

Bainbridge, K., Shute, V. J., Rahimi, S., Liu, X., 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 & Instruction, 77*, 1–11. https://doi.org/10.1016/j.learninstruc.2021.101547

Brusilovsky, P. (2002). Adaptive hypermedia: From intelligent tutoring systems to web-based education. In *Intelligent Tutoring Systems: 5th International Conference, ITS 2000 Montréal, Canada, June 19–23, 2000 Proceedings* (pp. 1–7). Berlin, Heidelberg: Springer Berlin Heidelberg.

Capuano, N., & King, R. (2015). Knowledge-based assessment in serious games: An experience on emergency training. *Journal of E-Learning and Knowledge Society*, *11*(3). https://www.learntechlib.org/p/151927/

Chen, F., Cui, Y., & Chu, M.-W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education, 30*(3), 481–503. https://doi.org/10.1007/s40593-020-00202-6

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

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper and Row.

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

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 1, pp. 416–436). Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n21

DeRosier, M. E., CraigAshley, B., & Sanchez, R. P. (2012). Zoo U: a stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction*. https://doi.org/10.1155/2012/654791

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society, 17*(1), 17–28.

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

DiCerbo, K. E., Xu, Y., Levy, R., Lai, E., & Holland, L. (2017). Modeling student cognition in digital and nondigital assessment environments. *Educational Assessment, 22*(4), 275–297. https://doi.org/10.1080/10627197.2017.1382343

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. Routledge.

Georgiadis, K., Faber, T., & Westera, W. (2019). Bolstering stealth assessment in Serious Games. In A. Liapis, G. N. Yannakakis, M. Gentile, & M. Ninaus (Eds.), *Games and learning alliance* (pp. 211–220). Springer International Publishing. https://doi.org/10.1007/978-3-030-34350-7_21

Glaveanu, V. P., Hanchett Hanson, M., Baer, J., Barbot, B., Clapp, E. P., Corazza, G. E., Hennessey, B., Kaufman, J. C., Lebuda, I., Lubart, T., Montuori, A., Ness, I. J., Plucker, J., Reiter-Palmon, R., Sierra, Z., Simonton, D. K., Neves-Pereira, M. S., & Sternberg, R. J. (2020). Advancing creativity theory and research: A socio-cultural manifesto. *The Journal of Creative Behavior, 54*(3), 741–745. https://doi.org/10.1002/jocb.395

Gray, A. (2016). The 10 skills you need to thrive in the Fourth Industrial Revolution. *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/

Gupta, A., Carpenter, D., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2021). Multimodal multi-task stealth assessment for reflection-enriched game-based learning. *MAIED@ AIED*, 93–102.

Halverson, R., & Owen, V. E. (2014). Game-based assessment: An integrated model for capturing evidence of learning in play. *International Journal of Learning Technology, 9*(2), 111–138. https://doi.org/10.1504/IJLT.2014.064489

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2020, July). Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. *International Educational Data Mining Society*. Retrieved from https://eric.ed.gov/?id=ED607823

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four-c model of creativity. *Review of General Psychology, 13*(1), 1–12. https://doi.org/10.1037/a0013688

Keller, J. M. (1987). The systematic process of motivational design. *Performance + Instruction, 26*(9–10), 1–8. https://doi.org/10.1002/pfi.4160260902

Kiili, K., & Ketamo, H. (2018). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies, 11*(2), 255–263. https://doi.org/10.1109/TLT.2017.2687458

Mayer, I. (2018). Assessment of teams in a digital game environment. *Simulation & Gaming, 49*(6), 602–619. https://doi.org/10.1177/1046878118770831

Messick, S. (1994). Validity of psychological assessment: Validation of inferences from persons' responses and performancesas scientific inquiry into score meaning. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Wiebe, E., Boyer, K. E., & Lester, J. C. (2015). Deep-Stealth: 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

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*. https://doi.org/10.1207/S15366359MEA0101_02

Phobun, P., & Vicheanpanya, J. (2010). Adaptive intelligent tutoring systems for e-learning systems. *Procedia-Social and Behavioral Sciences, 2*(2), 4064–4069.

Rahimi, S., & Shute, V. (2021). Learning analytics dashboards in educational games. In M. Sahin & D. Ifenthaler (Eds.), *Visualizations and dashboards for learning analytics* (pp. 527–546). Springer International Publishing. https://doi.org/10.1007/978-3-030-81222-5_24

Rahimi, S., Shute, V. J., Kuba, R., Dai, C.-P., Yang, X., Smith, G., & Alonso Fenandez, C. (2021). The 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., Fulwider, C., Jiang, S., & Shute, V. J. (2022a). Predicting learning gains in an educational game using feature engineering and machine learning. In C. Chinn, C. Tan, C. Chan, & Y. Kali (Eds.), *ICLS Proceedings—International Collaboration toward Educational Innovation for All* (pp. 2124–2125).

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

Rahimi, S., Shute, V. J., Khodabandelou, R., Kuba, R., Babaee, M., Esmaeiligoujar, S. (2023a). Stealth Assessment: A Systematic Review of the Literature. *the ICLS proceedings*. Montreal,

Rahimi, S., Almond, R. G., Shute, V. J., & Sun, C. (2023b). 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).

Rahimi, S., Almond, R. G., & Shute, V. J. (2023c). Stealth assessments' technical architecture. In M. P. McCreery & S. K. Krach (Eds.), *Games as Stealth Assessments* (pp. 1–25).

Reeves, T. C., & Lin, L. (2020). The research we have is not the research we need. *Educational Technology Research and Development, 68*(4), 1991–2001. https://doi.org/10.1007/s11423-020-09811-3

Reeves, T. C., & Oh, E. G. (2017). The goals and methods of educational technology research over a quarter century (1989–2014). *Educational Technology Research and Development, 65*(2), 325–339. https://doi.org/10.1007/s11423-016-9474-1

Resnick, M. (2018). *Lifelong kindergarten: Cultivating creativity through projects, passion, peers, and play* (Reprint edition). MIT Press.

Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan, 42*(7), 305–310.

Richards, R. (2010). Everyday creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (pp. 189–215). NY: Cambridge University Press.

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Shute, V. J. (1995). SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction*, *5*, 1–44.

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

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

Shute, V. J., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Springer International Publishing.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016a). 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*. https://doi.org/10.1111/jcal.12473

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. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Routledge.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and Learning of qualitative physics in newton's playground. *The Journal of Educational Research, 106*(6), 423–430. https://doi.org/10.1080/00220671.2013.832970

Shute, V. J., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs in video games. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment* (pp. 535–562). Wiley Inc.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016b). 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. (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).

Smith, G., Shute, V., & Muenzenberger, A. (2019). Designing and validating a stealth assessment for calculus competencies. *Journal of Applied Testing Technology, 20*(S1), 52–59.

Smith, G., Shute, V. J., Rahimi, S., Kuba, R., & Dai, C.-P. (2023). Stealth assessment and digital learning game design. In M. P. McCreery & S. K. Krach (Eds.), *Games as Stealth Assessments*.

Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). The dynamical analysis of log data within educational games. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 81–100). Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_4

Snow, E. L., Likens, A. D., Allen, L. K., & McNamara, D. S. (2016). Taking control: Stealth assessment of deterministic behaviors within a game-based system. *International Journal of Artificial Intelligence in Education, 26*(4), 1011–1032. https://doi.org/10.1007/s40593-015-0085-5

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

Verma, V. (2021). *Content agnostic game based stealth assessment* [Ph.D., Arizona State University]. Retrieved from https://www.proquest.com/docview/2533327267/abstract/DB15E052087447APQ/1

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

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children*. Holt, Rinehart & Winston.

Yang, D., Zargar, E., Adams, A. M., Day, S. L., & Connor, C. M. (2021b). 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

Yang, X., Rahimi, S., Fulwider, C., Smith, G., & Shute, V. J. (2022). Exploring students' behavioral patterns when playing educational games with learning supports at different timings. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-022-10125-9

Yang, X., Rahimi, S., Shute, V. J., Kuba, R., Smith, G., & Alonso Fernadndez, C. (2021a). The relationship among prior knowledge, accessing learning supports, learning outcomes, and game performance in educational games. *Educational Technology Research and Development, 69*, 1055–1075. https://doi.org/10.1007/s11423-021-09974-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Seyedahmad Rahimi Ph.D., is an assistant professor of Educational Technology in the School of Teaching and Learning at the University of Florida.

Valerie J. Shute Ph.D., is the Mack & Effie Campbell Tyner Endowed Professor Emerita in Education in the Department of Educational Psychology and Learning Systems at Florida State University.