Balancing Granularity and Context in Designing Evidence Model for Stealth Assessment: A Case of Pause Distribution in Digital Game-based Learning Environment

Zhichun Liu[1], Curt Fulwider[2]

1. Department of Human Development, Teachers College, Columbia University

2. Department of Educational Psychology and Learning Systems, Florida State University

Author Note

Correspondence should be addressed to Zhichun Liu, Department of
Human Development, Teachers College, Columbia University,
525 W 120 Street, New York, NY, 1002. Email: liulukas91@gmail.com

Balancing Granularity and Context in Designing Evidence Model for Stealth Assessment: A Case of Pause Distribution in Digital Game-based Learning Environment

The idea of a stealth form of assessment has attracted the attention of educators and educational researchers in recent years (Shute et al., 2020; Georgiadis, Van Lankveld, Bahreini, & Westera., 2020). The term *stealth assessment* first appeared in Shute (2011), but the principle and the motivation behind it has been driven by a long-standing issue in teaching and learning: How can learning be seamlessly assessed and supported in an unobtrusive and, most importantly, enjoyable way? Offline, in the physical world, assessment and support of learning rely on experienced instructors making observations and real-time instructional decisions. However, this in-person approach is only feasible at a small scale—the days of one-room school houses are long past—and the instructor requires time and effort to learn about their students, which is also an unscalable solution. Nowadays, a lot of learning is happening either via digital tools (e.g., instructional video, educational robots) or in digital environments (e.g., educational games and simulations, learning management systems). As a result, a large amount of digital data can be used to help make inferences about learning. Hence, stealth assessment becomes possible (McCreey, Krach, Bacos, Laferriere, & Head, 2019; Min et al., 2019; Smith, Shute, & Muenzenberger, 2017).

Stealth Assessment

The goal of stealth assessment is to make “quiet” inferences of learners’ competencies based on their observed behaviors and provide “powerful" instructional supports based on the inferences (Shute, Ventura, Bauer, & Zapata-Rivera, 2009; Shute., 2011). This approach can be very effective both in terms of assessment and learning. First, in terms of assessments, measurement occurs behind-the-scenes or stealthily. Learners do not need to know that they are “being observed” or “taking a test”. They can continue with the learning activity uninterrupted while the assessment occurs unobtrusively and continually in the background. This helps to reduce test anxiety, which affects both test performance and measurement validity (Gaye-Valentine, 2013; McDonald, 2001). Not knowing that they are taking a test will help learners perform and interact with the learning environment in a much more natural way, by reducing potential Hawthorne effects. In addition, stealth assessment adopts the idea of *evidence-centered design* (ECD, see Mislevy, Almond, & Lukas, 2003) and uses an iterative process to elicit evidence. The more learners interact with the learning system, the more evidence will be fed into the assessment component of stealth assessment, and the inference becomes more precise. Second, in terms of learning, stealth assessment provides a seamless learning experience and a sense of autonomy to learners, both are keys to achieving the *flow* states in learning (Csikszentmihalyi, 2009). And because the task selection is driven by ECD, the task difficulty is tailored to the learners' recent performance, so that (a) the completion result will provide the most information (Van der Linden & Boekkooi-Timminga, 1989) and (b) the task will be within the zone of proximal development (ZPD, see Vygotsky, 1978).

Stealth assessment has a few important components. First, ECD aims to establish an empirical model and build connections between what we can observe and what we want to measure (Almond, Kim, Velasquez, & Shute, 2014). Second, instructional supports (e.g., feedback) should be delivered formatively based on the results of the assessment. In an assessment-only context, no intervention is given. In contrast, under the vision of stealth assessment, formative instructional supports are important as they are the source of learning through digital tools or environments without interruptions. Combining the ECD assessment and formative instructional support, a learning system with stealth assessment will be able to interweave both assessments and learning at the same time and create a seamless and enjoyable learning experience.

Evidence-centered Design

An evidence-centered assessment design is the engine of stealth assessment. The ECD framework consists of a *competency model*, an *evidence model*, and a *task model* (Mislevy et al., 2003). The *competency model* (CM) describes the set of target knowledge or skills to be assessed. It acts as a profile of learners but at a fine-grain level. The CM (sometimes also called a student or proficiency model, see Mislevy, Steinberg, & Almond, 1999; Almond, Mislevy, Steinberg, Yan, & Williamson, 2015) needs to be well-operationalized and have high construct validity to avoid potential confusion or vagueness. The *evidence model* (EM) is the operationalization of the CM and it defines the connection between what and how performances observed within the activity indicate the level of competence defined in the CM. The *task model* (TM) is used to select the appropriate task contexts where certain performances and actions in EM can be elicited. Tasks are mapped to the CM via the definitions outlined in the EM. As a result, we can systematically infer learners' target competency through observable behaviors that occur within a predefined task context. The three models work together to connect the shreds of evidence we collect to the target competency we propose to measure.

Among the three models, the EM is particularly important. The CM contains conceptual information about what we want to measure, and the TM contains practical information about what we can collect. The EM aims to bridge the CM and the TM so that we can (a) use evidence in the TM to infer competency defined in the CM and (b) use the competency map in CM to inform the task selection in TM.

Challenges in Designing Observables

A challenge the original ECD model faces is in connecting the observed behaviors with complex competencies such as problem solving or collaboration (Arieli-Attali, Ward, Thomas, Deonovic,

& Von Davier, 2019). It is hard to capture the variations in learning progression and accurately interpret the meaning of learning activities based on the context. The assessment was largely task-driven, so when the contexts and interactions become complex, the ability to infer from data at a finer granularity becomes limited (Behrens, Mislevy, DiCerbo, & Levy, 2010). In one of our *Physics Playground* projects (e.g., Shute et al., 2020), we built a stealth assessment model based primarily on summary statistics such as time spent on task and overall in-game performance (e.g., level of reward earned at the end of each level) to represent learners' competencies. The results were promising in terms of construct validity and convergent validity on a broad level (i.e., force and motion, linear momentum, energy, and torque), but the model had limitations in differentiating more specific physics concepts nested within the broader constructs (e.g., conservation of momentum and property of momentum) because the nodes in the CM are interconnected and/or overlap at the lower level. Therefore, the EM, at lower levels, may have limited capacity to interpret variations in learning progressions.

Behrens et al. (2010) has suggested using "trace data " to unpack this complexity because many learning systems including digital games and simulations have brought the ability to capture highly detailed behavioral data as performance traces, which contains vast amounts of information. In a different example of stealth assessment, Shute, Wang, Greiff, Zhao, & Moore (2016) designed specific rules in the EM to link behaviors with target competencies (i.e., problem-solving skills) in the context of a modified version of the popular mobile game *Plant vs. Zombie*. For example, one observable was the use of "plant food when there are <3 zombies on the screen ". This observable is theoretically tied to effective tool usage and leverages the power of game-based behavioral data, However, this type of observable is specific to the game itself, which means it lacks transferability to other games or other instructional contexts. In addition, learners may demonstrate effective tool usage that is not fully captured by one or two predetermined observables—especially in highly interactive and adaptive problem-solving environments. Both ends of the spectrum propose a challenge to designing EMs for stealth assessment: We need to balance the granularity in creating observables. Not too coarse that we will lose the valuable information to unpack the complexity; nor too fine-grained that the observables lack transferability and flexibility.

Pause: A Simple Behavior in Complex Contexts

One possible way of addressing this type of challenge is to design observation-based behavioral interactions that are common to most learning challenges (i.e., transferability) and unpack the complex meanings of the interactions under different interactive contexts (i.e., fine granularity). In this study, we investigate the pause behavior, a very common action to almost all multi-step cognitive tasks, in a game-based learning environment and demonstrate its complex meaning under different contexts as well as the different potential implications on learning.

Complex meanings of pauses in game. While many assessments tend to focus on the final product (e.g., completion, accuracy, speed), learning in a self-governed interactive system such as games emphasizes the problem-solving processes (Eseryel, Law, Ifenthaler, Ge, & Miller., 2014). Problem-solving processes and behaviors have been used to both measure the proficiency level of learners (Jitendra, Sczesniak, Deatline-Buchman, 2005; Stoeffler, Rosen, Bolsinova, & von Davier, 2020) and predict the learning outcomes (Taub et al., 2017). One important indication of problem solving is how someone monitors and regulates the problem-solving processes (Bransford & Stein, 1984; Gick, 1986). The regulation does not have to happen visibly as a behavior because reflection can be an internal and mental process (McAlpine, Weston, Beauchamp, Wiseman, & Beauchamp, 1999). Active reflection (in the form of a pause from the gameplay) can indicate a metacognitive strategy on managing limited cognitive resources, which is crucial in complex problem-solving environments (Rhodes, 2019; Wright, 1992). In addition, pauses can also indicate cognitive reappraisal strategies such as affective regulation (Spann, Shute, Rahimi, D'Mello, 2019; Gross, 1998). Studies have also shown that even a short passive pause from cognitive tasks can increase subsequent cognitive performance (Ariga & Lleras, 2011; Jansseen et al., 2014). However, on the other hand, pauses can also simply indicate disengagement. Learners might be distracted by other irrelevant tasks or completely overwhelmed and get bored with what they are asked to do. However, no matter if the pause is cognitive reflection, emotional regulation, or disengagement from the task, pauses are a rich source of data about the learning process. Therefore, in this study, we propose to unpack the complexity of pause behavior in an educational game with the purpose of designing a balanced game trace observable for stealth assessment.

A conceptual framework of pauses in game. Although sometimes problem-solving processes in cognitive tasks can be diverse across different individuals, researchers have managed to break the process down into stages. Many theoretical frameworks have been proposed to interpret the process. For example, Polya (1957) proposed a four-step model (i.e., understanding problem, strategy design, strategy execution, evaluation of the strategies). Further, Schweizer, Wüstenberg, & Greiff (2013) proposed a simplified two-step model of problem-solving: rule-acquisition and rule-application. Both the understanding and evaluation phases of the problem-solving process are for acquiring rules; while both design and implementation are for applying the rules to solve the problem. Learners may not necessarily be doing anything while in the rule acquisition state (e.g., analyzing a problem, evaluation on a solution), where more cognitive and metacognitive resources are engaged, hence fewer observable actions and longer pauses. On the other hand, the rule application state requires more active interactions and implementation, hence shorter pauses. Therefore, within the game problem-solving context, there should exist both active and inactive stages in terms of pauses. In addition, considering some students may occasionally disengage from the problem-solving, a disengaged stage is also plausible.

Research Questions

1. What does the pause behavior look like in an educational game? Do there exist various types of pauses in gameplay?

2. What is the relationship among various types of pauses, learning, and game enjoyment?

Method

Data

Participants for this study were a convenience sample of 199 students between 9th to 11th grade from a large K-12 school in southeastern United States. The characteristics of the sample are listed in Table 1. It is worth noting that the data was drawn from an experimental study where students were randomly assigned into one of the three game conditions. Each condition had a different sequence of gameplay (i.e., the order in which new levels were presented to the player). In our current study, we are not addressing the performance difference between groups because all students' received the same experience within each level. The performance differences between groups are described in Shute et al., (2020).

All students participated in six 50-minute sessions playing *Physics Playground*, a 2D digital game designed to develop middle and high school students' conceptual physics learning (see Figure 1 for example levels). The goal of the game is to use different physics agents (e.g., lever, pendulum, gravity) to direct a green ball to hit a red balloon. In Physics Playground, students can interact with the game in various ways: from drawing physics agents to changing environmental parameters (e.g., gravity, air resistance), from resetting the ball's position to accessing learning supports. The game engine helps us to record every game interaction that happens over the course of someone’s gameplay. Based on timestamp information, we operationalize the pause behavior as the interval between any two consecutive user-initiated game events:

$$Pause_i \ = Time_{i+1} - Time_i$$

We first removed all the non-user-initiated game events from the log files (e.g., earning a trophy, snapshots of objects positions) because they were created either in conjunction with a user-initiated event or auto-generated throughout gameplay. We then removed all the between-session and between-level pauses because they are out outside the scope of our analysis—pause behavior *within* a game task. And finally, we removed extremely small outliers that were identified as either system or user errors (i.e., pauses under 0.08s, which is more than 12 clicks within 1 second, see Click Speed Test, n.d.). After removing sources of possible noise, we were left with a log file containing only the pauses each participant made throughout gameplay. Pause data was extremely skewed given the nature of pause time within gameplay where short pause between typical gameplay behaviors occur much more frequently and long pauses (e.g., > 1 minute) are rare. Therefore, we applied a natural log transformation to all pause data.

Table 1. Demographic Information of Participants by Group

Condition	*n*	Boy	Girl	Other or prefer not to say
1	64	32	31	1
2	68	40	27	1
3	67	32	33	2
Total	199	104	91	4

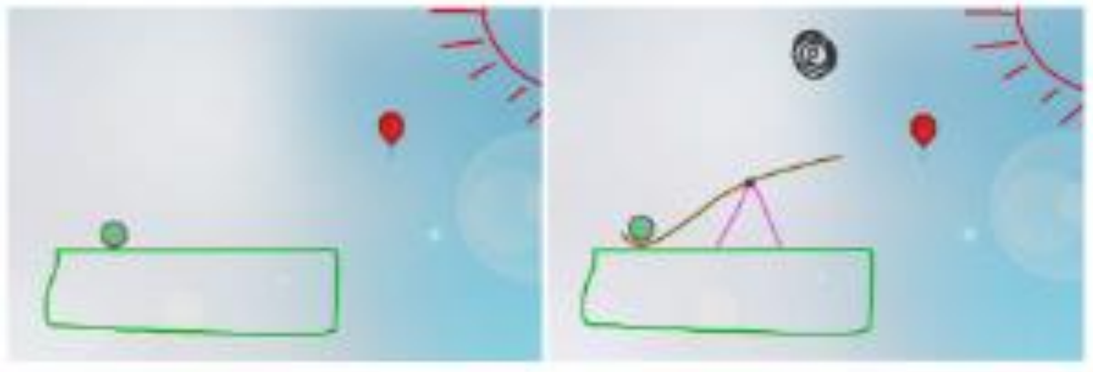

Figure 1. One level in Physics Playground. In this level, students will need to draw a lever to lunge the ball to the balloon.

Analysis

Because of their complexity, pause distributions are believed to be a combination of potential behaviors from normal delays between game actions to disengagement. In other words, a pause of around a single second may be the movement from one side of the screen to another. More than five minutes of no activity may indicate disengagement. Therefore, we chose to use a finite mixture model to fit the pause distribution data.

Multi-modal distributions are an indication that multiple latent classes may be present within a single continuous distribution. Therefore, we begin our analysis using descriptive data (e.g., median values, density plots) to understand the potential mixture distribution of all pauses. We then use K-Means to identify potential cut-off points between modes within each distribution. The groups formed around the cut-off values are labeled (i.e., active, inactive, and disengaged) and used as priors in fitting the finite latent-class mixture model with the flexmix R package (Leisch, 2004). The initial labels are then verified and refined based on the results of the flexmix model. By labeling and refining the mixture distribution, we then obtain each students' potential frequency and proportion of different pause types.

Finally, we examine correlations between pause behaviors and the following gameplay and assessment variables used as indicators of student performance: pretest scores, gain score, the total number of levels completed, and a self-reported measure of enjoyment. The pretest was administered prior to playing and focused on the targeted physics concepts. The gain scores were the difference between the posttest (a parallel form of the pretest) and the pretest (for more information see Shute et al., 2020). The total number of levels solved is self-explanatory. Players played at their own pace to complete as many levels as possible in the time allowed. And, finally, enjoyment was measured as a component of a concluding survey. Participants reported their level of enjoyment of the game on a 5-item questionnaire.

Estimation Methods

The goal of the estimation is to identify the finite latent classes of a continuous pause distribution. The estimation was done using the expectation-maximization algorithm with maximum likelihood and MCMC sampling under the Bayesian framework implemented by flexmix. A single type of pause distribution in cognitive tasks (e.g., writing tasks) can be simulated by a log-normal distribution which typically has long tails (e.g., Almond, Deane, Quinlan, Wagner, & Sydorenko, 2012; Guo, Deane, vanRijin, Zhang, & Bennett, 2018). Therefore, after log transforming the pause intervals, we can assume each potential component follows a univariate normal distribution, which makes the estimation into a latent class regression (DeSarbo & Cron, 1988). Flexmix can help us identify the component parameters including the means and variances.

Results

Descriptives of Pause Distribution

We first obtained the descriptive data of all 199 students' pause distributions (Table 2). Students in all three conditions behaved similarly. Overall pauses ranged from 0.08 seconds to 336.14 seconds with a median pause at 0.5 seconds across all participants. Although students in condition one (i.e., game levels were presented in a set sequential order) demonstrated slightly longer pauses in relation to the overall median, in general, students did not differ substantially across conditions.

Table 2. Descriptive data of pause distribution by condition

Condition	n	Min	1st Q	Median	Mean	3rd Q	Max
1	64	.080	.18	.58	2.68	2.49	358.48
2	68	.081	.18	.46	2.39	2.31	315.68
3	67	.080	.17	.48	2.34	2.20	335.67
Total	199	.08	.18	.50	2.47	2.33	336.14

We then produced density plots of the log-transformed distributions (Figure 2). The density plots show that the distribution of pauses was multimodal—primarily bi-model but occasionally tri-modal or multimodal. This indicates that the continuous distribution of pause times may be a result of mixture distributions—different states possibly contributing to the different durations of pause.

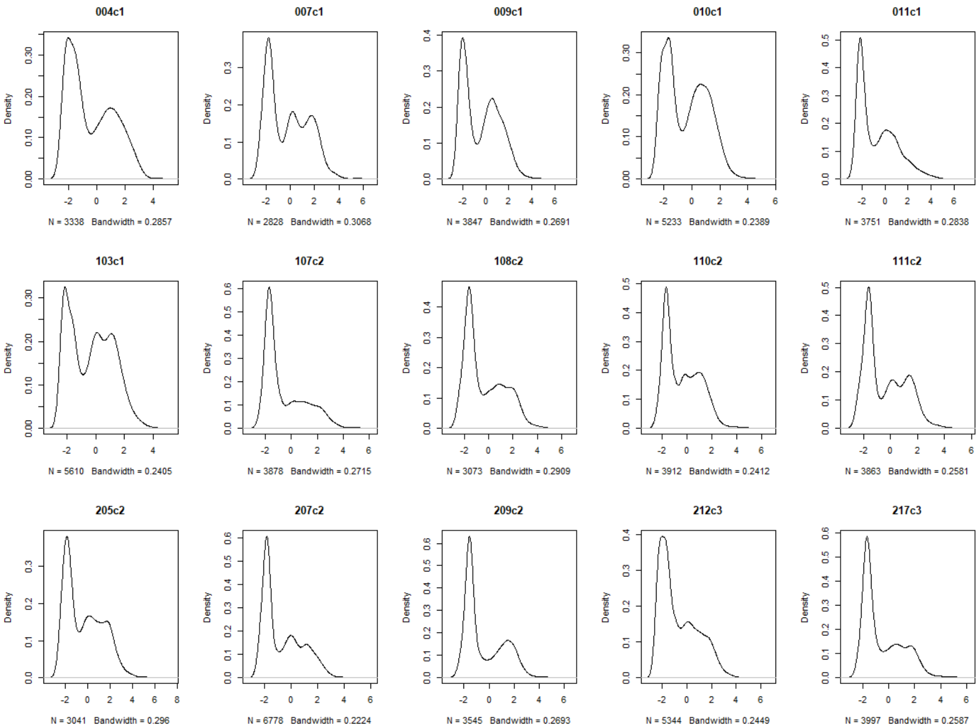

Figure 2. Sample density plots of pause times.

Determining The Number of States in Pause Distribution

The literature provides the theoretical support for constructing a mixture of “active-inactive” pause distribution. In addition to the “active” and “inactive” states, learners could have been completely disengaged, which is out of the scope of regular problem solving in cognitive tasks—they could have taken much longer pauses because they are distracted (e.g., chat with friends, on their phone, bathroom break). As a result, we propose that a potential mixture distribution of pause times consists of three possible states: (a) active, (b) inactive, and (b) away-from-keyboard (i.e., disengaged, AFK). Depending on how much a learner is engaged, AFK may not exist throughout the gameplay. Figure 4 shows all pause behaviors demonstrated by all students. As the conceptual framework suggests, we identified three possible states. The shortest pause state (i.e., active states) possibly centered around 0.1 second (around e^-2 seconds); The intermediate pause state (i.e., inactive) possibly centered around 1 to 3 seconds (around e^0 and e^2 seconds), while the longest pause state (i.e., AFK) possibly centered more than 50 seconds (around e^4 seconds).

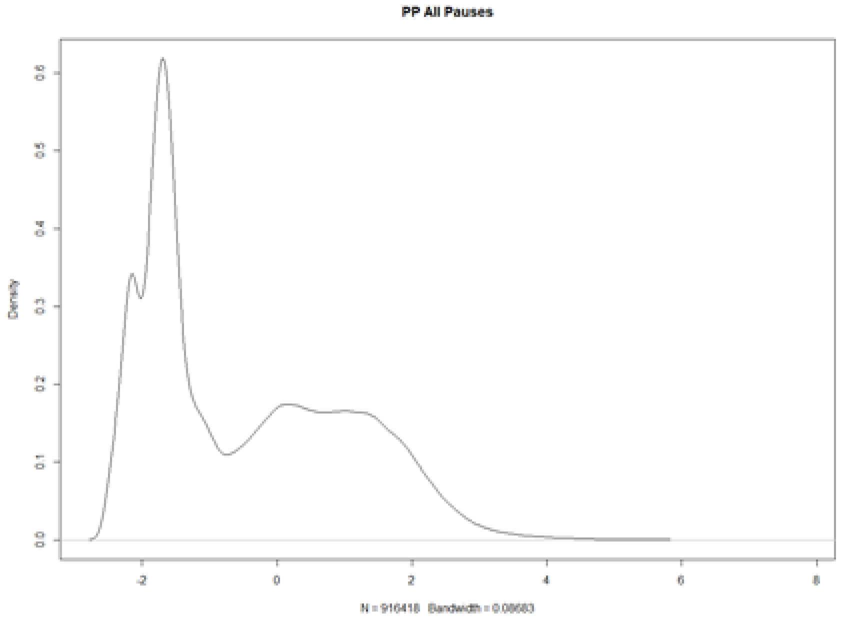

Figure 4. Pause times of all students on a natural log scale.

K-Means

Fitting a mixture model with latent class regression using the flexmix package takes a Bayesian approach, which requires a prior distribution. The priors were created from the results of the K-means analysis. Based on the conceptual framework we have discussed; we intend to identify three groups within students' overall pause distribution. We used the K-Means algorithm to find three possible centers based on all pause times within each students' pause distribution (Table 3) and the distribution of possible centers is shown in Figure 5.

Table 3. Descriptive data for centers identified by K-Means of each student's pause distribution

Center	Min	1st Q	Median	3rd Q	Max
1	0.08	0.18	0.40	2.15	39.57
2	39.63	47.69	60.87	88.21	200.44
3	14.17	71.75	149.56	321.27	1968.54

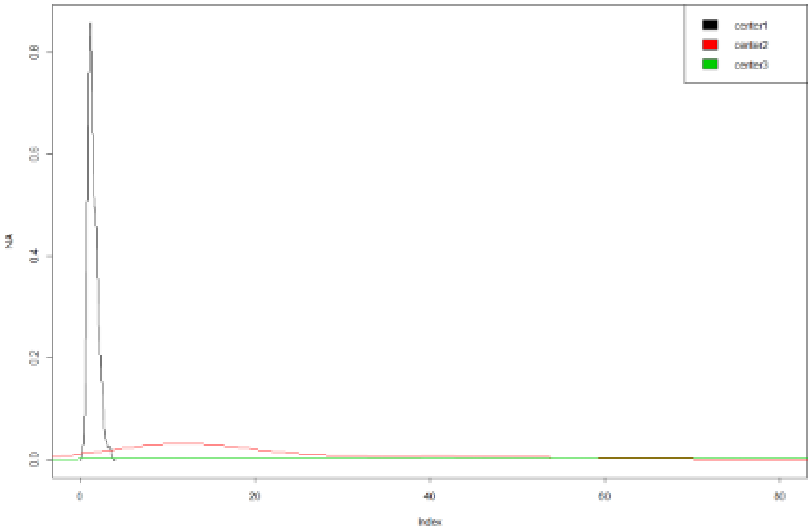

Figure 5. The distribution of centers identified by K-Means of each student's pause times.

Using the median of possible centers may not capture the majority of diverse students, therefore, we decided to use the third quartile of the first (i.e., 2.15 seconds) and second centers (i.e., 88.21 seconds) as the preliminary cutoff for active, inactive, and AFK states. It is worth noting that AFK states vary and are quite infrequent during gameplay. Therefore, this type of behavior is not likely to be picked up by the flexmix model. On the other hand, AFK states are very different from active and inactive pauses, but disengagement is important to identify. Therefore, we hand-labeled the AFK states using the cutoff point of 88.21 seconds based on the results of the K-means analysis. In other words, inactive and active states were identified using flexmix (discussed in the next section), but AFK states were identified based on the results of K-means.

Flexmix

Because we manually labelled the AFK states, the task of the flexmix mixture model became differentiating active and inactive states. We used the aggregated results from K-Means classification as priors to train a mixed model with the expectation-maximization algorithm to label active and inactive states for each individual player. With flexmix, we were able to refine the initial labels of active and inactive pause times. With the refined labeling, we were able to calculate the individual students' proportional feature of pause behaviors (i.e., proportion of active states, inactive states, and AFK states among all pauses).

Relationship between Learning, Enjoyment, and Pause States

To validate the relationship between mined features, we considered four types of performance measures: (a) incoming knowledge, (b) level completion, (c) learning gain, and (d) the game enjoyment. The results suggest that prior knowledge was negatively related to the proportion of active states (r = -16, p < .05), positively related to the proportion of inactive states (r = .17, p < .05), and negatively related to the proportion of AFK states (r = -.30, p < .001). Level completion, on the other hand, was positively related to the proportion of active states (r = .19, p < .001), negatively related to the proportion of inactive states (r = -.17, p < .05), but again, negatively related to the proportion of AFK states (r = -.53, p <.001). Learning gain was not correlated to neither active nor inactive states, but it was negatively related to AFK states (r = -.17, p < .05). Finally, enjoyment was not related to active (r = .-11, p = .12), inactive (r = .11, p = .11), or AFK states (r = -.10, p = .18). The correlational matrix is displayed in Table 6 below.

Table 6. Means, standard deviations, and correlations with confidence intervals

Variable	*M*	*SD*	1	2	3	4	5	6
1. active	0.70	0.09						
2. inactive	0.29	0.09	-1.00**					
			[-1.00, -1.00]					
3. AFK	0.00	0.00	-.14*	.10				
			[-.28, -.00]	[-.04, .24]				
4. Pretest	11.82	3.53	-.16*	.17*	-.30**			
			[-.29, -.02]	[.03, .30]	[-.43, -.17]			
5. Gain	0.63	2.89	.08	-.08	-.17*	-.28**		
			[-.06, .22]	[-.21, .07]	[-.30, -.03]	[-.41, -.15]		
6. Level Completion	45.86	16.05	.19**	-.17*	-.53**	.44**	.21**	
			[.05, .32]	[-.30, -.03]	[-.62, -.42]	[.33, .55]	[.07, .34]	
7. Game Enjoyment	18.32	4.64	-.11	.11	-.10	.15*	.21**	.09
			[-.25, .03]	[-.03, .25]	[-.23, .05]	[.01, .28]	[.07, .34]	[-.05, .23]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates *p* < .05. ** indicates *p* < .01.

Discussion

RQ1: What Does the Pause Behavior Look Like in An Educational Game?

We have shown, as a result of our analysis, that pause behaviors in Physics Playground are diverse. Pauses range from as short as .08 seconds to as long as 5 minutes. The multi-modal distribution (shown in Figure 3) indicates that there may exist multiple latent states behind the observed pause times. However, students may have utilized multiple approaches to solve in-game problems. Based on the previously defined framework and the present results, gameplay may be divided into a *problem-solving state* where students are engaged with the game—actively trying to complete the task— and an *AFK state* where students are disengaged from the problem-solving activity. The former state produces relatively shorter pauses while the latter produces much longer pauses. The problem-solving state can be further broken down into two types of shorter pauses: *active pauses* which are natural transitions between normal gameplay behaviors (e.g., moving objects on the screen, exploring learning supports, building structures, etc.) and inactive pauses which may manifest when students are engaged in thinking processes such as planning, prediction, and evaluation that do not result in observable behaviors.

Fitting the distribution to a mixture model, we were able to distinguish three possible states across the continuous pause distribution and identify the pause states for each student. This process is critical for modeling students' game behavior because it represents each students' distinctive gameplay pattern. Some students may be active in the game in terms of creating different physics agents or maneuvering the ball (e.g., consecutive nudging), while some students may have relatively parsimonious game actions because more time is devoted to planning or evaluation. Because these unique gameplay patterns can be considered behaviors elicited from unique problem-solving states, we can use observed pause behavior patterns to infer each student's in-game problem-solving states as the basis for stealth assessment design.

RQ2: The Relationship Between Pause, Learning, And Game Enjoyment

We observed a consistent negative correlation of AFK behavior and all performance indicators, meaning that extended pause durations are likely indications of disengagement from the game. However, this disengagement may not necessarily harm the enjoyment because disengagement can sometimes be a strategy for emotional regulation from difficult tasks. On the other hand, active states and inactive states have an interesting relationship between learning: Constant active gameplay may help the student finish more levels (r = .19, p < .001), but it did not necessarily transfer to learning nor enjoyment (r = .-11, p = .12) nor was it related to higher incoming knowledge. On the other hand, inactive gameplay may not help learners to solve more game levels (r = -.17, p < .05) within a given time, these learners could be learners with higher incoming knowledge and are more engaged with predicting/evaluation, which could take more time. Overall, we demonstrated the potential underlying meaning of different types of pause behaviors, which supported the classification. Although they are all pauses, each pause behavior may carry a different meaning depending on its context.

Implications To Stealth Assessment: Finding Balance in Granularity

Leveraging data mining methods to unpack the complexity. In this study, we have shown that the pause between gameplay actions, albeit less than a second at times, can be an ample source of meaning depending on the context. A brief pause within a cognitive task may be the natural transition from one behavior to another. However, the pause may be an indication of a person's problem-solving approach. Are they taking a moment to predict the next outcome? Perhaps they are considering a new solution design. Or, perhaps they have become distracted by another irrelevant matter—total disengagement. This is information we can distill from mundane game actions to gain a better understanding of the entire gameplay at a finer granularity.

Therefore, different educational data mining methods may be beneficial for designing observables in stealth assessment. As discussed earlier, one challenge in stealth assessment is to shift the focus from gross performance (e.g., overall task performance) to trace data, a ripe source of underutilized data. However, trace data itself may not convey much information because of the fine granularity of the data. At first glance, trace data may appear too chaotic to yield any meaningful insight. However, in this study, we have demonstrated a potential means to differentiate the types of pause behavior related to problem solving by fitting a mixture distribution model. This approach is similar to Grover et al. (2017)'s proposal in extending the classic ECD framework. In their work, data-driven techniques such as clustering, and pattern recognition are used to help the development of scoring rules. This approach is in addition to the traditional theory-based hypothesis-oriented way of defining scoring rules, which can be limited when the learning system becomes dynamic and complex.

Balancing the granularity. It is important to balance the granularity issue in designing stealth assessment (Figure 6). At the least granular level, we need gross performance and summary statistics to capture students' general proficiencies. This approach has been well-established and strongly supported by past research (e.g., Shute et al., 2020). However, this approach overlooks valuable interaction data (i.e., trace data generated in a digital game system). At a moderate grain size, we need to apply contextual meaning to trace data. This could be a complete exploratory approach with unsupervised machine learning approaches (e.g., clustering or dimension reduction) or it could also be done in a theory-driven supervised approach (e.g. the classification approach demonstrated in this current study). Ultimately, the goal is to find emerging patterns within trace data to make better sense of the higher level (i.e., larger grain-size) performance data. Trace data can be a magnifier to zoom into a specific context and understand the variability in learning performances. At the finest grain size, we can focus on the individual behaviors to design observables for use in stealth assessment (as illustrated in Shute et al., 2016), but we may lose the flexibility and transferability. Therefore, we must find a balance of granularity for designing observables. On the one hand, we can leverage various data mining methods to design observables that are detailed enough to capture the learning trajectory but also generic enough to be generalized to other tasks. On the other hand, we can gain understandings of how learning happens in a complex and interactive environment. The result of this type of knowledge could directly benefit both the instructional design and learning science communities.

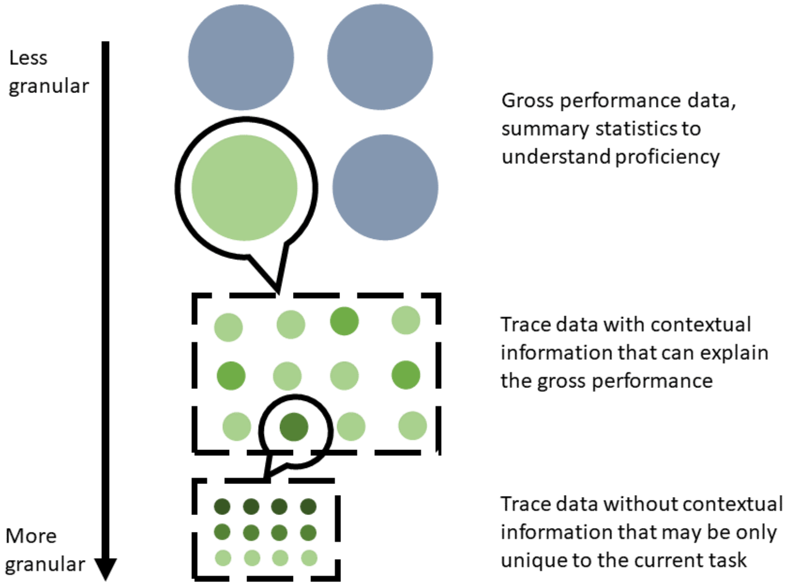
Less granular
More granular
Gross performance data, summary statistics to understand proficiency
Trace data with contextual information that can explain the gross performance
Trace data without contextual information that may be only unique to the current task

Figure 6. The granularity in designing observables for stealth assessment

Validity issues. However, validity is a potential caveat in leveraging data mining methods for designing stealth assessment. First, at a less-granular level, defining and operationalizing the construct can be relatively straightforward (e.g., through creating a Q-matrix to link the performance in task model and competency model). As the granularity becomes finer, we as designers are trying to link more detailed interactions with cognitive tasks to the competency model. This linkage may be dynamic and complex, which presents a potential threat to construct validity of the observables—are we really sure what we see means what we think? Second, although the data mining approach can be theory-driven (as illustrated in this current study), the exploratory nature of data mining may mislead researchers and designers by showing a temporal correlation, which leads to a potential threat to internal validity. Third, a natural threat to external validity will arise: How much would the conclusion hold within cognitive tasks across different contexts and outside out of cognitive tasks altogether? Therefore, it is crucial to (a) test the construct validity including convergent and divergent validity with other possible measures, (b) triangulate the claim through multiple methods (e.g., observations, interviews, or other analysis approaches) to enhance the internal validity, and (c) use conjecture mapping (Reimann, 2016) and iterations (Anderson & Shattuck, 2012) to test the extent to which the claim will hold in other settings as suggested in design-based research.

Limitations And Future Directions

This study has a few potential limitations. First, as discussed in the previous section, this study did not have the chance to comprehensively examine potential validity issues. Although we attempted to provide explanations to the pause behavior with various learning and enjoyment measures, it is recommended that future research focus on either replicating the process in a different context and triangulate the conclusion with other quantitative or qualitative analytical tools. Second, we only have selected one behavior—pausing—to illustrate how to leverage data mining methods and find a balance granularity in designing stealth assessments. Many more cognitive-task-related behaviors such as resetting, revisiting, or help-seeking are worth investigating. And to this end, future research should focus on building a more generalized framework of stealth assessment for a variety of cognitive tasks. Third, we have not fully unpacked the AFK behavior because of its sparsity. Anecdotally, during the experiment sessions, we noticed that some students break from the game as an emotional regulation strategy, which allows them to come back to the game with a fresher mind. This type of AFK is different from complete disengagement. However, compared to both active and inactive pauses, neither type of AFK behaviors have enough data for thorough analysis. As a result, we have chosen a hard cut-off to label this type of behavior. Yet, we recognize that there is still much information buried within the game traces we have collected.

Conclusion

In this study, we have discussed some challenges in the design of stealth assessment. We used a case study to demonstrate how to unpack the hidden information in game trace data with a simple yet complex behavior—pausing. We have identified multiple possible meanings of pausing in the game and also shown its relationship with performance and enjoyment data that we have collected from the students. The findings of this study will help researchers and practitioners to design evidence models that can understand the complex learning behaviors under various contexts and shed light on implementing stealth assessment in educational settings.

Acknowledgement

We thank Dr. Valerie Shute for her vision and insights in stealth assessment and building up an amazing Physics Playground team. We also would like to thank the entire Physics Playground team for the dedication. Special thanks go to Dr. Russell Almond for his advice on the mixture model. We also appreciate Wenting Song for helping us identify a missing citation. This research was supported by the US National Science Foundation (award number #037988) and the US Department of Education (award number #039019).

References

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://dx.doi.org/10.1080%2F15366367.2014.910060

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., and Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer. https://doi.org/10.1007/978-1-4939-2125-6

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task*. *ETS Research Report Series*, *2012*(2), i-61. https://doi.org/10.1002/j.2333-8504.2012.tb02305.x

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research?. *Educational researcher*, *41*(1), 16-25. https://doi.org/10.3102%2F0013189X11428813

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & Von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in psychology*, *10*, 853. https://doi.org/10.3389/fpsyg.2019.00853

Ariga, A., & Lleras, A. (2011). Brief and rare mental “breaks” keep you focused: Deactivation and reactivation of task goals preempt vigilance decrements. *Cognition*, *118*(3), 439-443. https://doi.org/10.1016/j.cognition.2010.12.007

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://files.eric.ed.gov/fulltext/ED520431.pdf

Bransford, J., & Stein, B.S. (1984). *The IDEAL problem solver: A guide for improving thinking, learning, and creativity*. W. H. Freeman.

Click Speed Test (n.d.). *Clicks per second*. https://clickspeedtest.com/clicks-per-second.html

Csikszentmihalyi, M. (2009). *Flow: The psychology of optimal experience* (Nachdr.). Harper & Row.

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification, 5*(2), 249-282. https://doi.org/10.1007/BF01897167

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Journal of Educational Technology & Society, 17*(1), 42-53.

Gaye-Valentine, A. (2013). Assessing the construct validity of test anxiety: The influence of test characteristics and impact on test score criterion validity. *TPM - Testing, Psychometrics, Methodology in Applied Psychology, 2*, 117–130. https://doi.org/10.4473/TPM20.2.2

Georgiadis, K., Van Lankveld, G., Bahreini, K., & Westera, W. (2020). On the robustness of stealth assessment. *IEEE Transactions on Games, 13*(2), 180-192. https//doi.org/ 10.1109/TG.2020.3020015

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist, 21*, 99-120. https://doi.org/10.1080/00461520.1986.9653026

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of general psychology, 2*(3), 271-299. https://doi.org/10.1037%2F1089-2680.2.3.271

Grover, S., Bienkowski, M., Basu, S., Eagle, M., Diana, N., & Stamper, J. (2017). A framework for hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming. In M. Hatala (Ed.), *Proceedings of the*

Seventh International Learning Analytics & Knowledge Conference 2017. Association for Computing Machinery, 530–531. https://doi.org/10.1145/3027385.3029440

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, *55*(2), 194-216. https://doi.org/10.1111/jedm.12172

Janssen, M., Chinapaw, M. J. M., Rauh, S. P., Toussaint, H. M., Van Mechelen, W., & Verhagen, E. A. L. M. (2014). A short physical activity break from cognitive tasks increases selective attention in primary school children aged 10–11. *Mental health and physical activity*, *7*(3), 129-134. https://doi.org/10.1016/j.mhpa.2014.07.001

Jitendra, A. K., Sczesniak, E., & Deatline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review*, *34*(3), 358-371. https://doi.org/10.1080/02796015.2005.12086291

Leisch, F. (2004). *Flexmix: A general framework for finite mixture models and latent glass regression in R*. https://cran.r-project.org/web/packages/flexmix/vignettes/flexmix-intro.pdf

McAlpine, L., Weston, C., Beauchamp, C., Wiseman, C., & Beauchamp, J. (1999). Building a metacognitive model of reflection. Higher education, 37(2), 105-131. https://doi.org/10.1023/A:1003548425626

McCreery, M. P., Krach, S. K., Bacos, C. A., Laferriere, J. R., & Head, D. L. (2019). Can video games be used as a stealth assessment of aggression?: A criterion-related validity study. *International Journal of Gaming and Computer-Mediated Simulations*, *11*(2), 40-49. https://doi.org/10.4018/IJGCMS.2019040103

McDonald, A. S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology*, *21*(1), 89–101. https://doi.org/10.1080/01443410020019867

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2019). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies, 13*(2), 312-325. https://doi.org/10.1109/TLT.2019.2922356

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *On the roles of task model variables in assessment design*. National Center for Research on Evaluation, Standards, and Student Testing. https://files.eric.ed.gov/fulltext/ED431804.pdf

Polya, G. (1957). *How to Solve It?*. Princeton University Press.

Reimann, P. (2016). Connecting learning analytics with learning research: The role of design-based research. *Learning: Research and Practice*, *2*(2), 130-142. https://doi.org/10.1080/23735082.2016.1210198

Rhodes, M. G. (2019). *Metacognition. Teaching of Psychology*, *46*(2), 168-175. https://doi.org/10.1177%2F0098628319834381

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. Learning and Individual Differences, 24, 42-52. https://doi.org/10.1016/j.lindif.2012.12.011

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Information Age Publishers.

Shute, V. J., Rahimi S., Smith, G., Ke, F., Almond, R., Dai, C-P, Kamikabeya, R., Liu, Z., Yang, X., & Sun, C. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in Physics Playground. *Journal of Computer-Assisted Learning*, *37*, 127–141. 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. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295-321). Routledge, Taylor and Francis.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. R. (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

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. http://www.jattjournal.com/index.php/atp/article/view/142702

Spann, C. A., Shute, V. J., Rahimi, S., & D'Mello, S. K. (2019). The productive role of cognitive reappraisal in regulating frustration during game-based learning. *Computers in Human Behavior.* https://doi-org.proxy.lib.fsu.edu/10.1016/j.chb.2019.03.002

Stoeffler, K., Rosen, Y., Bolsinova, M., & von Davier, A. A. (2020). Gamified performance assessment of collaborative problem solving skills. *Computers in Human Behavior, 104*, 106036. https://doi.org/10.1016/j.chb.2019.05.033

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. *Computers in Human Behavior*, *76*, 641-655. https://doi.org/10.1016/j.chb.2017.01.038

Van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for IRT-based test design with practical constraints. *Psychometrika*, *54*(2), 237-247. https://psycnet.apa.org/doi/10.1007/BF02294518

Von Wright, J. (1992). Reflections on reflection. Learning and instruction, 2(1), 59-68. https://doi.org/10.1016/0959-4752(92)90005-7

Vygotsky, L. S. (1978). Interaction between learning and development. In M. Gauvain & M. Cole (Eds.), *Readings on the development of children* (2nd ed., pp. 33–40). New York: Scientific American Books. https://doi.org/10.2307/j.ctvjf9vz4.11