Nuances in Efficacy of Computer-Based Learning Platforms

Understanding the Interplay between Instructional Design, Engagement, and Learning

Dissertation

Kirk P. Vanacore

Dissertation Committee:

Adam Sales, Worcester Polytechnic Institute (Chair)
Neil Heffernan, Worcester Polytechnic Institute
Erin Ottmar, Worcester Polytechnic Institute
Stacy Shaw, Worcester Polytechnic Institute
Ken Koedinger, Carnegie Mellon University

School of Arts and Sciences
Worcester Polytechnic Institute
Worcester
August 20, 2024

Contents

* **

[*] Submitted Works

[**] Published Works

Abstract

Assessing the impact of educational programs on student learning is a vital part of educational and learning sciences research. Randomized control trials are often employed to evaluate the effectiveness of these programs, with results typically presented in binary terms – effective or ineffective. However, programs and practices may be effective for some students and not for others, with the program's impact possibly dependent on specific implementation parameters and students' attributes. One under-explored dimension that likely influences the efficacy of educational programs is student engagement.

This dissertation presents research on two opposing engagement behaviors – gaming the system and productive persistence – commonly exhibited in computer-based learning platforms. The studies included in this dissertation explore how programs' instructional designs can influence students' behavior and how students' behavioral tendencies moderate the programs' efficacy. For students who game the system, The evidence suggests that the most argent 'gamers' may benefit from delayed access to hints and feedback, with mixed evidence on the impact of gamification. Although one gamified CBLP likely negatively impacted 'gamers' learning, another had no interaction with students' propensity to game the system, indicating a complex relationship between disengagement and instructional designs. On the other end of the spectrum, gamified performance-based feedback positively affected students' productive persistence behaviors. Furthermore, the impact of various programs differed based on students' tendencies toward engaging in productive persistence. This finding suggests that the ways in which programs allow for and encourage productive persistence may influence whether and how students learn.

Overall, this work highlights the need to consider behavior heterogeneity —- how a program's effect varies based on participants' behavioral tendencies -— when evaluating whether and for whom programs are impactful. If a program's impact depends on students having propensities towards or away from specific behaviors, instructional designers should consider how programs can be developed to increase positive learning behaviors and provide students with the educational experiences that will benefit them most.

Chapter 1

Introduction

Researchers, governments, and funding agencies have put substantial effort and resources into assessing the impact of educational programs on student learning [16, 19, 24, 39, 43]. Typically, this involves running efficacy studies.* Randomized control trials have been lauded as the best study design for evaluating effectiveness as randomization allows for estimating unbiased and unconfounded treatment effects [7, 38]. The results of these studies are often presented as a binary. An educational program may be effective or ineffective. A teaching practice works, or it does not work. However, some programs and practices are effective for some students and not for others [18, 40]. Furthermore, a program's impact may depend on specific implementation parameters [15, 28]. Effectiveness may also vary continuously along measures of participants' attributes, like students' prior knowledge of the content [13]. Therefore, to have a clear understanding of efficacy, we must consider many possible dimensions in which a program can be more or less impactful [8, 10].

Engagement is one understudied dimension that likely influences the impact of an educational program. Figure 1.1 presents a conceptual model of how student engagement is involved in mediating and moderating the effects of a program on learning outcomes. In this heuristic, the program components are divided into two aspects: the content that is being taught and the design of the program that delivers the content. Instructional designs can include different problem types, adaptive features, on-demand assistance, and various forms of gamification. Both aspects influence students' in-program processes. These behaviors include how students attend to the learning activity (*i.e.,* engagement) and their cognitive processes underlining improvements in understanding and abilities that occur while students are using the programs (*i.e.,* cognitive processes). Engagement behaviors such as diligence or persistence may manifest as continued work within a learning activity even after low performance [25] or appropriate pauses for thinking [12, 27]. Conversely, students may display disengagement by quitting an activity early (*i.e.,* stopout [9]) or progressing through an activity without learning (*i.e.,* gaming the system [3]). Notably, students' engagement may influence their cognitive processes. For example, attentiveness to aspects of the program may be required for certain learning gains. The cognitive processes may be indirectly observed through behaviors such as performance and, when it is available, shown work (*e.g.,* math deviations, open response explanations). Yet cognitive processes

*Throughout this paper, I use the words efficacy and effectiveness to mean the same thing: whether a program has the intended impact on participants. The nuanced distinctions between efficacy and effectiveness are beyond the scope of this paper, but broadly speaking, the methods involved in this work could apply in both contexts.

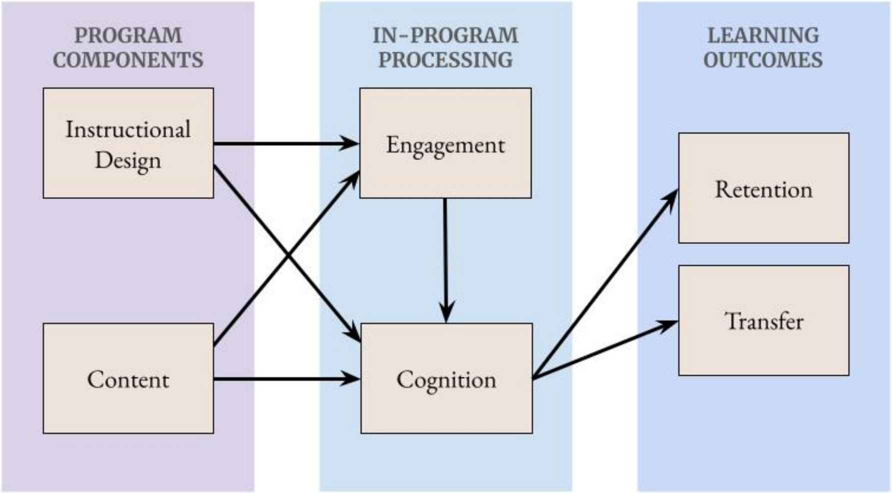


Figure 1.1: Conceptual model for how program components impact learning outcomes through in-program processes.

are internal and are therefore unobservable for researchers. Though these processes may be latent, learning can be estimated through measures of retention and transfer [44].

This model displays how adjustments in the instructional design may influence learning in two ways. First, they can directly affect students' cognitive processes through which learning occurs. For example, including illustrative diagrams of complex concepts as part of the instructional design may help students understand those concepts, thus directly impacting the students' cognitive processes. Alternatively, the instructional design may influence students' cognitive processes by first impacting the way those students engage with the content. Providing gamified rewards, for example, may increase persistence, giving students more learning opportunities and thus influencing their cognitive processes and, by proxy, their learning outcomes. The following work presents studies that delve into these indirect paths by evaluating how engagement interacts with the effects of institutional design on learning.

The following studies and proposed analyses address two juxtaposed behaviors characteristic of different qualities of engagement: gaming the system and productive persistence. Gaming the system is the act of attempting to progress through a learning activity in a CBLP by leveraging the features of that CBLP, such as hints and feedback, to submit correct responses [3]. Alternatively, productive persistence is the effortful continuance of a learning activity to overcome difficulties or misunderstandings. Often, this requires students to deliberately engage in one learning task while intentionally delaying their progress in the program, ostensibly to ensure they have mastered the underlying skills [21, 32]. Notably, these behaviors are orthogonal – one cannot engage in one while also engaging in the other – and they represent two ends of the spectrum of engagement. When students game the system, they exhibit minimal engagement beyond aimless clicking or leaving the activity altogether. Therefore, gaming the system behaviors represent the bottom of the engagement spectrum. On the other hand, when students display productive persistence, they work at the top of the engagement spectrum, as productive persistence requires effortful conscientiousness [32].

Computer-based learning platforms (CBLPs) employ various instructional designs, in part, to help maximize student engagement with learning activities Many programs embed instruction, feedback, and assistance into problem sets to make the content more accessible to students with a wide range of prior knowledge and abilities [11, 29, 30, 35, 42]. Gamification – the inclusion of game-like features into traditionally non-game activities – can increase engagement by making content and concepts more amenable to students who are likely to dislike and even avoid the learning tasks [14, 26]. According to [6]'s taxonomy of game attributes, games generally include action language, assessment, conflict/challenge, control, environment, game fiction, human interaction, immersion, and rules/goals. The definitions of game features can be obtuse, such as features that induce a ”series of interesting choices” [37] to those that inspire challenge, curiosity, control, and fantasy [31]. Generally, the role of gamification is to make an activity fun and engaging, thus changing students' behaviors and attitudes [26].

In theory, instructional design decisions may affect students differently based on their engagement tendencies. For example, a student who is likely to disengage from a static problem set may be inspired to engage in a problem set if there are hints and feedback that make the problems understandable and accurately attainable. Alternatively, levels of assistance that are misaligned with a student's abilities may shield students from the need to participate with the requisite amount of struggle for learning [23]. Furthermore, students who are likely to be disengaged by abusing those hints and feedback could alternatively immerse themselves in a learning game [36]. Thus, the effect of these methods may vary by, or even be continent upon, students' engagement profiles.

Despite the focus on educational efficacy studies from researchers and policymakers [16, 19, 24, 39, 43] as well as the immense effort to predict and understand students' learning-related behaviors within CBLPs (*e.g.,* [2, 5, 17]), there is a gap in understanding the nuanced relationship between how students engage with CBLPs and CBLPs' effectiveness. The following work includes a series of studies that address some of the ways in which engagement interacts with instructional designs within CBLPs, thus potentially influencing their efficacy. As stated above, I focus on two components of engagement—gaming the system and productive presence—in the context of CBLPs with differing instructional designs involving varying forms of assistance delivery and levels of gamification. This work will contribute to our growing understanding of the underlying learning mechanisms as students use CBLPs. For example, if a CBLP only impacts students when they engage with that program in a specific way, that pattern of behaviors and the program features that support them likely represent a learning mechanism. Understanding the interaction between engagement and efficacy can help learning experience designers build more impactful programs by developing features that help students engage in behaviors associated with learning. Furthermore, the heterogeneity identified by these analyses suggests personalized differential impact patterns; some students will benefit from one instructional design, whereas others will benefit from an alternative design. Thus, the analyses point toward the need for dynamic personalization, while also providing some indicators of which students might benefit from which design.

Beyond the theoretical and practical implications, there are methodological components of this work relevant to the field of learning sciences and the study of CBLPs. Each analysis utilizes a causal methodology that allows for nuanced inferences about how CBLPs impact students. May of the following analyses include on a method of causal moderation known as Fully Latent Principal Stratification (FLPS) that allows for the interactions between treatment assignment and latent states that emerge after the intervention begins. I also use a method for analyzing regression discontinuity designs (RDDs), Limitless RDD, to estimate causal effects when treatment assignment is determined by a cutpoint on some scale, even when that scale is not continuous and the outcome is binary. Throughout this proposal, I discussed the implications of how these methods can be used in various contexts to further our understanding of how students learn through CBLPs.

1.1 Dissertation Overview

The following document consists of four sections presenting multiple published and unpublished works, culminating in the proposal of new work. Each work represents a stand-alone manuscript. Therefore, although I have taken steps to prevent self-plagiarism, the descriptions of the study design, data, and methods are, at times, similar or redundant across these documents.

Beyond providing a brief preface to the topic of CBLP effect heterogeneity based on engagement, the Introduction (Chapter 1) includes a book chapter published in *Perspectives on Learning Analytics for Maximizing Student Outcomes*, in which I, along with my coauthors, present work that looks beyond performance orientation of learning analytics towards process base orientation to understanding and impacting student learning. In this work, I discuss the need to consider the impact of CBLP in the context of students' engagement as a part of their learning processes.

In Chapter 2, I present two studies analyzing the effect heterogeneity of CBLPs based on students' tendencies towards gaming the system (*i.e.,* 'gamers') using FLPS. The first study (Section 2.1), *The Effect of Assistance on Gamers: Assessing The Impact of On-Demand Hints & Feedback Availability on Learning for Students Who Game the System*, focuses on how the availability of assistance while solving problems interacts with students' propensity to game the system. The second study (Section 2.2), *Effect of Gamification on Gamers: Evaluating Interventions for Students Who Game the System*, considers the extent to which 'gamers' would benefit from a gamified design. I find that the most ardent 'Gamers' may benefit from delayed access to hints and feedback and less. Furthermore, I found mixed evidence for how 'gamers' are affected by gamification. One of the studied gamified CBLPs likely negatively impacted 'gamers' learning. Yet assignment to another gamified CBLP did not interact with students' propensity to game the system. Overall, these interactions suggest a nuanced relationship between disengagement and each of the CBLPs' instructional designs.

Chapter 3 involves two studies of productive persistence: one estimates the effect of a specific feature within a gamified CBLP on a productive persistence behavior, and another evaluates the effect of the CBLP varies based on the behavior. The first study (Section 3.1), *Effect of Game-Based Failure on Productive Persistence*, examines how gamification can help students respond to failures productively. Although failure typically precedes disengagement in traditional learning environments [1], in the context of games, failure can create experiences of ”pleasant frustration” that engage players to continue attempting to succeed [20]. Using Limitless RDD, I find that game-based failure has a positive impact on students' likelihood to display productive persistence by reattempting problems on which they have performed poorly. Overall, this finding suggests that gamification may be used to help students overcome the typically demotivating nature of failure experiences.

In the final Chapter (Chapter 4), I explore how students' propensity to display productive persistence by reattempting problems after sub-optimal performance is associated with differences in the effectiveness of CBLPs. Using a Fully Latent Principal Stratification (FLPS) model on data from three CBLPs, results show that while reattempting problems boosts performance, its impact varies depending on the CBLP's instructional design. The findings emphasize the importance of designing educational technologies that not only promote persistence but also offer multiple problem-solving opportunities to cater to diverse student needs.

1.2 Programs and Data

As mentioned above, the studies presented in Chapters 2 and 3 and those proposed in Chapter 4 use data from an efficacy study, which evaluated the impact of three CBLPs — ASSISTments, From Here to There (FH2T), and DragonBox-12 (DragonBox) – on US middle school students' algebraic understanding. The study results are reported in [13], and a full description of the data can be found in [33]. Each manuscript presented in this paper provides relevant details of the study design and the data used.

1.3 Beyond Performance Analytics**

See manuscript below.

Beyond Performance Analytics:
Using Learning Analytics to Understand Learning Processes that Lead to Improved Learning Outcomes

Authors

Kirk Vanacore (kpvanacore@wpi.edu); Worcester Polytechnic Institute;

Ji-Eun Lee (jlee13@wpi.edu); Worcester Polytechnic Institute;

Alena Egorova (aegorova@wpi.edu); Worcester Polytechnic Institute;

Erin Ottmar (erottmar@wpi.edu); Worcester Polytechnic Institute;

Citation

Vanacore, K., Lee, J., Egorova, A., & Ottmar, E. (2023). Beyond Performance Analytics: Using Learning Analytics to Understand Learning Processes That Lead to Improved Learning Outcomes. In G. Durak & S. Cankaya (Eds.), *Perspectives on Learning Analytics for Maximizing Student Outcomes* (pp. 168-187). IGI Global. https://doi.org/10.4018/978-1-6684-9527-8.ch009

Abstract

To meet the goal of understanding students' complex learning processes and maximizing their learning outcomes, the field of Learning Analytics delves into the myriad of data captured as students use computer-assisted learning platforms. Although many platforms associated with Learning Analytics focus on students' performance, performance on learning-related tasks is a limited measure of learning itself. In this chapter, we review research that leverages data collected in programs to understand specific learning processes and contribute to a robust vision of knowledge acquisition. In particular, we review work related to two important aspects of the learning process: students' problem-solving strategies and behavioral engagement, then provide an

example of an effective math program that focuses on the learning process over correct or incorrect responses. Finally, we discuss ways in which we can incorporate findings from this research into the development and improvement of computer assisted learning platforms, with the goal of maximizing students' learning outcomes.

1. Introduction

Learning is a complex process that requires exposure to content, thought, struggle, and memory (Bjork & Bjork, 2020; Koedinger et al., 2023; Lynch et al., 2018; Okano et al., 2000). To have learned something, a student must not only be able to perform a task during or directly after instruction; they must demonstrate that they can retain the information or skill and apply it to new situations (Soderstrom & Bjork, 2015). Learning systems in which students encounter desirable difficulties (Bjork & Bjork, 2020) – such as varying presentation of content (Smith et al., 1978; Smith & Handy, 2014), interweaving knowledge components instead of presenting them sequentially (Rohrer et al., 2014), spacing content delivery (Cepeda et al., 2006) and retrieval practice (Karpicke & Zaromb, 2010) – increase the likelihood of learning. Yet, when desirable difficulties are designed into learning activities, students' performance often suffers, even as these design choices can positively affect learning as measured by distal outcomes (Roediger & Karpicke, 2006; Shea & Morgan, 1979; Smith & Rothkopf, 1984).

Despite the potential incongruence between performance and learning, Computer Assisted Learning Platforms (CALP) often rely heavily on students' performance within learning activities as the primary measure to evaluate students. The instructional design in these CALPs may vary greatly; some incorporate game-based learning (Siew et al., 2016), puzzles (Rutherford et al., 2010), and simulations (Martens et al., 2004; McCoy, 1996), while others focus on more traditional problem sets with tutorial instruction (Heffernan & Heffernan, 2014). While specific design and difficulty within these instructional methods may also differ, many of these CALPs use mastery learning (Barnes et al., 2016; Heffernan & Heffernan, 2014; Macaruso & Hook, 2007; Ritter et al., 2016). Mastery learning is based on the premise that students must master a knowledge component prior to progressing to new content (Bloom, 1968). Mastery of a skill is often determined using students' performance within the activities, either by modeling student knowledge or using simple criteria, such as solving three problems correctly in a row (Corbett & Anderson, 1994; Kelly et al., 2015; Reich, 2020; Yudelson et al., 2013). Early education technologies relied on modified versions of Rasch models which estimated the probability a student will get a problem correct based on a function of a problem's difficulty and a student's ability (Reich, 2020). Alternatively, knowledge tracing – which seeks to predict the probability that a student has mastered a knowledge component – has emerged as one of the main methods for assessing students' mastery of learning within a problem set (Corbett & Anderson, 1994; Yudelson et al., 2013). Furthermore, mastery learning systems often include interactive elements that provide assistance, which may further boost performance by reducing the mental effort necessary to produce correct responses, while potentially hindering learning (Koedinger & Aleven, 2007; Koedinger et al., 2008). Overall, these systems rely on students' performance data, most often represented by their binary outcomes on problems within an activity, to evaluate student learning and determine whether students should progress on to the next knowledge component.

Furthermore, a portion of Learning Analytics (LA) research also relies on proximal performance within a program. For example, A/B tests, which are commonly used in LA research both to evaluate features and to study learning mechanisms, often use proximal performance as an outcome measure. A/B tests are experiments that test the effect of conditions on a desired outcome through random assignment to a treatment (e.g., access to feature) and control (e.g., no feature access). They have commonly used the impact of evaluate features - such as hints, feedback, and reward systems - on student learning. Some of these tests rely solely on the students' performance directly after the problem to estimate the effects of features and interventions (Haim et al., 2022; Patikorn & Heffernan, 2020; Prihar et al., 2021; 2022), while others use a variety of performance and behavioral outcomes that were generated during or directly after the experimental problem set (Gurung, Baral, et al., 2023; Vanacore et al., 2023).

Many mastery learning programs are effective at improving students' learning outcomes, and analyses of performance have advanced our understanding of parts of the learning process (Anderson et al., 1995; Hurwitz & Vanacore, 2022; Kulik et al., 1990; Roschelle et al., 2016). Yet, there is evidence that alternative approaches to mastery learning, such as emphasizing students' learning processes as they solve problems, may have a greater impact on learning (Decker-Woodrow et al., 2023). For LA to fulfill the goal of understanding student learning and optimizing instructional systems (Long et al., 2011), the field has probed deeper into students' behaviors in CALP *as they solve problems* while incorporating this understanding into the development of learning systems. This process leverages different problem types which can range from traditional multiple-choice or fill-in-the-answer problems to more complex puzzles and games that provide different data about how students engage with the content. This undertaking requires analyzing and utilizing more than the proximal performance measures within the data, and focusing on the *learning processes* by which students arrive at the answers they submit.

Our chapter presents various LA research that goes beyond performance measures to understand students' learning processes as they solve problems in various CALP. The chapter is divided into two sections (Sections 2 and 3). In Section 2, we provide examples of research on both mastery learning programs and programs that use alternatives to proximal correctness when evaluating student learning, and explore how the field is moving toward a more nuanced understanding of learning. In Section 3, we provide an example of a program, which focuses on students' problem-solving processes, effort, and engagement, and discuss its impact on student learning.

To frame Section 2, we present a conceptual model (presented in Figure 1) in which data generated by CALP are used to understand how constructs that underlie students' learning processes lead to robust learning defined by retention and transferability. This model is not comprehensive, but it is a useful framework for understanding LA research. Studies in LA use a wide variety of log data collected in CALP. For the purpose of this chapter, we have grouped data into two general categories: *action data* and *time data. Action data* varies depending upon the program, but may include log-in/out data, attempts to solve problems, submitted answers, use of hints embedded within the programs, and steps students take to solve the problem (i.e., derivations). *Time data* can be aggregated at different levels to include students' response time for each action, the time taken to complete a problem, or their total time using the program. Section 2 focuses on how various data from CALP is being used in LA research to study two key constructs of learning: *problem-solving strategies* and *behavioral engagement*.

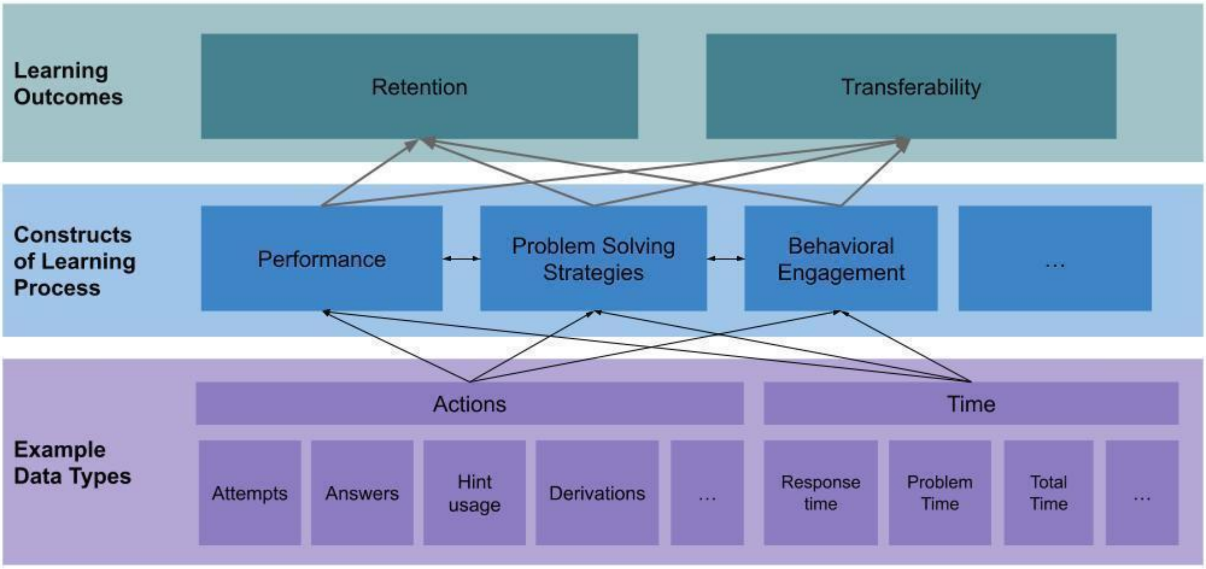


Figure 1. Conceptual model of the components of learning analytics research which include the data captured in CALP systems, the components of the learning process, and learning outcomes. Lines represent the interconnected nature of the components across the model's levels.

In Section 3, we use an online learning game, *From Here To There!* (FH2T) as an example of how systems that emphasize the learning processes can effectively promote students' mathematical learning. We outline the theoretical basis for FH2T and discuss evidence of its efficacy. Finally, we discuss new causal research on understanding key mechanisms and behaviors that drive the efficacy of FH2T as well as ways in which we can encourage effective learning behaviors within learning systems. As a whole, FH2T, and its associated research base, illustrate how CALP can improve our understanding of students' learning processes while positively impacting those processes by looking beyond performance.

2. Beyond Performance: Studying Learners' Problem-solving Strategies and Engagement

2.1 Problem-solving Strategies

In order to solve problems, students need to engage in cognitive processing to understand a problem, derive a solution, and conduct a sequence of steps to arrive at their answers. In many CALP, this process is *unobserved,* as programs simply ask the students to select or input their final answers into the system. In these cases, possible problem-solving paths may be inferred by speculating plausible processes for incorrect answers. While incorrect answers can help provide cues about student errors and inform targeted feedback and guidance, the variability in correct approaches or multiple strategies that students use while problem-solving is largely invisible in the data.

In traditional mastery learning CALP, researchers have attempted to derive students' processes from their given solution. For example, in a series of studies, Gurung and colleagues (2023A; 2023B) had teachers examine common wrong answers in a CALP, speculate how students would produce those answers, and then create automated feedback that directs toward the estimated paths. The feedback had mixed success in affecting students' performance, which may indicate that the teacher's speculations were at times incorrect. Notably, this system only works for incorrect answers and does not accommodate multiple paths to any solution. Furthermore, the students' true path to their solution remains unobserved.

Alternatively, when online learning programs focus on students' problem-solving processes or strategies rather than performance on tasks (i.e., producing correct answers), researchers can utilize *observable* students' actions in the form of log data collected in CALP to unpack traditionally *unobserved* cognitive processes (Rowe et al., 2021; Shute et al., 2016; Sun et al., 2022). For example, Shute and colleagues (2016) applied data mining methods to use student in-game action data as an indicator of problem-solving skills in an online mathematics learning game called "Use Your Brainz". The goal of this game was not just to produce a correct answer, but students needed to apply various problem-solving skills to achieve the goal (i.e., planting special plants on a lawn to prevent zombies from invading their houses). In this study, the authors divided students' use of problem-solving strategies into four sub-constructs: analyzing givens and constraints, planning a solution pathway, using tools and resources effectively, and monitoring and evaluation process. Then, they identified 32 students' in-game actions logged in the system (e.g., planting over three flowers before the zombies arrive), mapped them into each sub-construct of problem-solving skills (e.g., analyzing givens and constraints), and implemented this problem-solving model in the game using Bayesian networking. The results revealed that the in-game indicators of problem-solving skills significantly correlated with the results of two external tests of problem-solving skills, suggesting that the problem-solving skills assessment using in-game metrics is valid.

In another study, Rowe and colleagues (2021) used classification algorithms to create automated detectors of students' implicit problem-solving processes based on gameplay behaviors in an online mathematics game called "Zoombinis." Similar to Shute's (2016) study, they divided students' problem-solving strategies into four different phases (i.e., trial and error, systematic testing, systematic testing with a partial solution, implementing with a full solution) and used both raw log data (e.g., overall gameplay data) collected in the game system and hand-labeled data from observations of students gameplay (e.g., problem decomposition, pattern recognition) to build over 100 features (i.e., detectors) that might be indicative of these strategies. Then, they examined the relationships between these detectors and external post-assessment scores for validation and found that most of the detectors were significantly associated with the post-assessment scores.

Further, when in-game action data is combined with other types of data obtained from external sources (e.g., interaction among students), researchers can measure more complex learning processes, such as how a group of students uses their knowledge and skills to solve complex problems collaboratively. For example, Sun and colleagues (2022) investigated the relationship between collaborative problem-solving behaviors and solution outcomes in a physics learning game, "Physics Playground." They conducted qualitative coding of students' utterances during their gameplay and identified 19 indicators of collaborative problem-solving skills (e.g., constructing shared knowledge, negotiation). Then, they examined the relationship between these indicators and in-game performance (e.g., problem-solving efficiency) measured by action data. The results indicated that "proposing solution ideas contributed to desirable outcomes" was the most influential predictor of in-game performance out of 19 indicators of collaborative problem-solving, suggesting that collaborative problem-solving requires individual contributions as well as collective interactions.

The game-based algebraic CALP *From Here To There!* (FH2T; https://graspablemath.com/projects/fh2t), which was developed by Ottmar and colleagues (2015) utilizes a slightly different approach by directly logging *sequences* of students' problem-solving steps. In each problem of the game, students are asked to transform an algebraic expression or equations (e.g., "11+55+y+89+45" in Figure 2) into a mathematically equivalent but perceptually different goal state (e.g., "100+y+100" in Figure 2) using permissible touch-screen or mouse-based gesture-actions. The sequence of steps a student makes to reach the goal state (e.g., Figure 2a through f) is captured along with its timestamp in the system, allowing researchers and learning analysts to study students' various pathways to reach the solution.

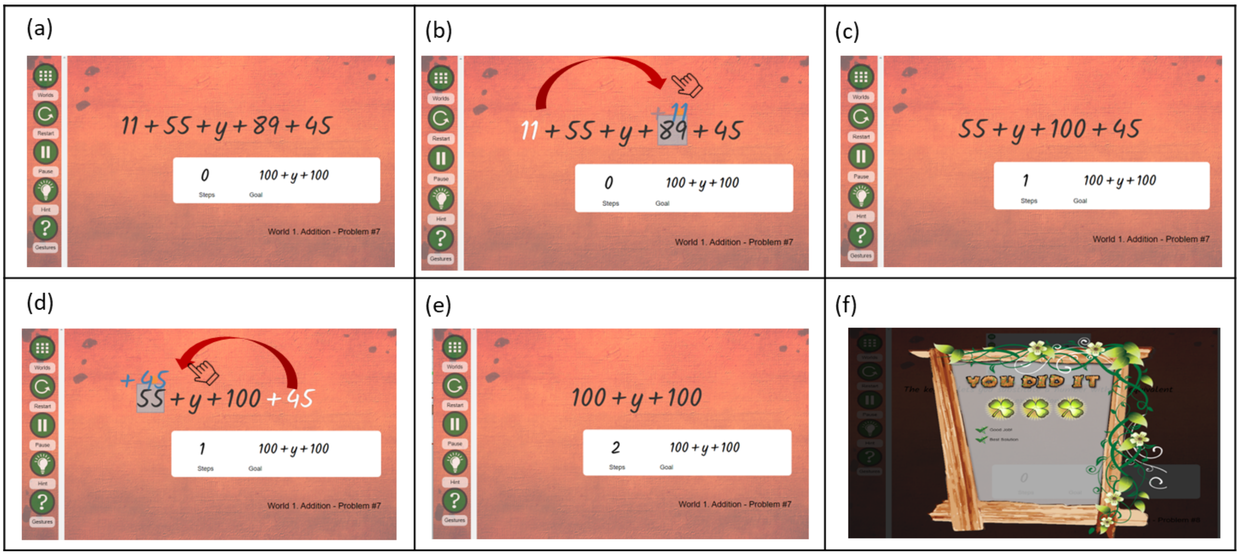


Figure 2. A sample problem in FH2T in which shows the steps students take to manipulate the equation from the start to the goal state.

In contrast to traditional mastery learning programs, FH2T does not evaluate students by producing answers to math problems but by how efficiently they reach the specified goal state in the game. For example, if a student reaches the goal state with the fewest steps possible to complete the problem (i.e., also called the "optimal step") using a two steps strategy (Student A in Figure 3: [start state: 11+55+y+89+45] → [step 1: 11+100+y+89] → [step 2: 100+y+100]), three clovers (i.e., rewards in the game) are given, and an efficiency score (optimal step count/total step count) of 1 is assigned. However, the number of clovers they receive is deducted if the students exceed the minimum required number of steps to reach the goal state (e.g., Student B and Student C in Figure 3). As such, there is no correct/incorrect dichotomy in the game, but there are several different pathways to solve the problems, which allows researchers to explore variations of students' mathematical approaches and problem-solving processes to reach the solution.

Using the data collected in FH2T, we applied data visualization techniques and explored students' algebraic problem-solving processes in the game (Lee, Stalin, et al., 2022). Specifically, the study visualized individual students' step-by-step information about the problem-solving process (e.g., math strategies used, time spent between each action) using Indivisualizer (See Figure 3) and created Sankey diagrams (See Figure 4) to investigate how the productivity of the first step influenced the overall efficiency of problem-solving. The results showed a large variation in students' use of mathematical strategies to solve the problems, with some approaches being more efficient than others, and the productivity of the first step significantly predicted the overall efficiency of problem-solving. The findings suggested that these data visualizations depicting students' problem-solving processes can help unpack individual students' cognitive processing as well as variability in overall students' mathematical problem-solving strategies.

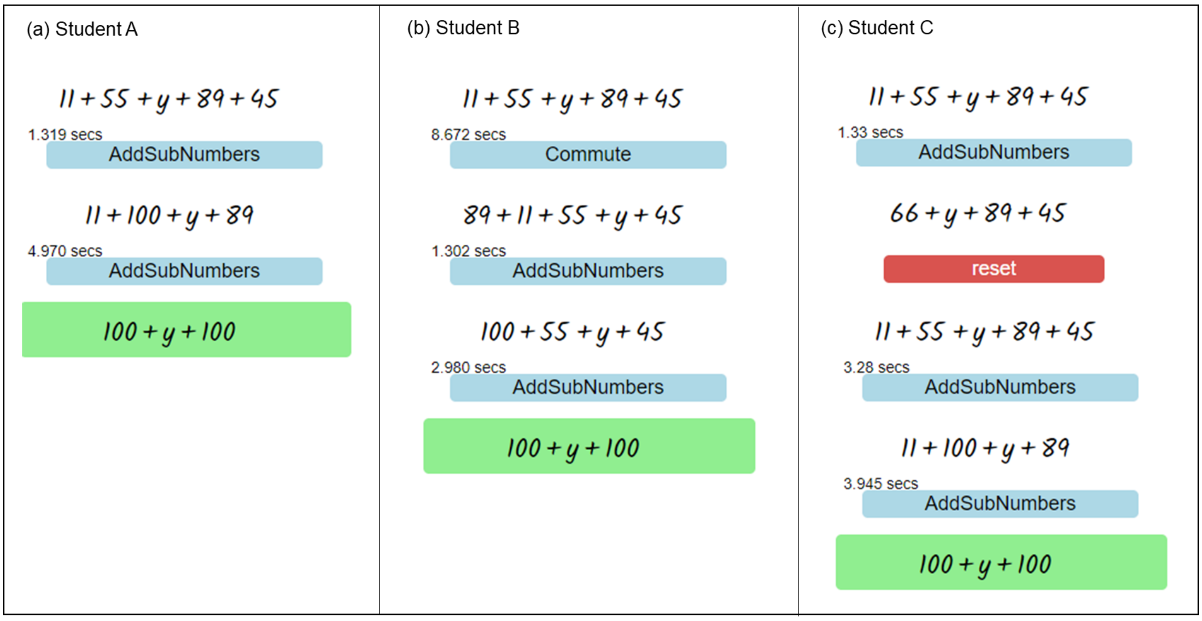


Figure 3. An example of the Indivisualizer

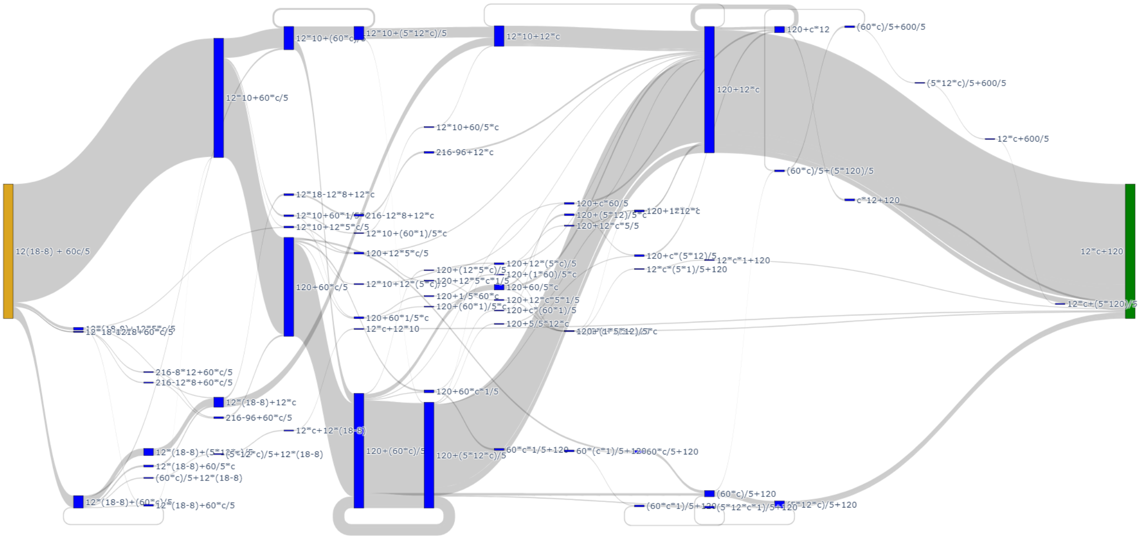


Figure 4. An example of a Sankey diagram showing all student pathways for a given problem

In this case, it is also possible to utilize *time data* to estimate processes, such as the amount of thought a student might be applying to the problem before they produce a solution. Using action data can help us grasp the invisible cognitive processes; however, there is always an amount of thinking ascribed to the action. In order to solve problems properly or efficiently, a student needs to examine a problem, pause, and strategically think about what to do next rather than rushing through the problem. Some studies have used "think aloud" methods to explore students' thinking and pause time (Charters, 2003), but these methods can be time-consuming to collect the data individually for each student. Time data automatically collected in CALP also can be used as a potential indicator of students' cognitive processing. As an example, we used time data to measure students' thinking and pausing before problem-solving in FH2T (Chan, Ottmar, et al., 2022). By computing the time students spent before making a first action on each problem, we found that students' pre-solving pause time was positively associated with strategy efficiency, which suggests that pause time may be a proxy indicator of students' strategic planning (i.e., thinking about pathways and solutions to the specific problem). Note that more details about FH2T are discussed in the later section.

2.2 Students' Behavioral Engagement

Another important aspect of the learning process is students' *behavioral engagement*, which is generally defined as the student's involvement in one's own learning or academic tasks (Sinatra et al., 2015). In classrooms, researchers often measure students' behavioral engagement through class attendance and participation in activities, attentiveness, effort, persistence toward the tasks, or self-directed information-seeking actions when facing obstacles (Sinatra et al., 2015). Studies have found that students' behavioral engagement is significantly associated with positive proximal and distal learning outcomes, such as knowledge gain, better performance (Ladd & Dinella, 2009; Rohrer et al., 2014; Rutherford et al., 2014), and a lower probability of dropout (Archambault et al., 2009).

As with students' problem-solving strategies, behavioral engagement is often observed indirectly in CALP. In mastery learning CALP, LA researchers use process data to assess behavioral engagement, often by attempting to predict whether students will display suboptimal behavioral patterns. For example, researchers have exerted extensive effort to discern whether students are “gaming the system,” which is attempting to succeed by exploiting properties of the system (e.g., requesting all the hints available) instead of actively engaging with the material being taught (Baker et al., 2008). For example, Baker and colleagues (2006) built a detector that identifies students' gaming behaviors using data collected through field observations and students' in-game action data. Alternatively, Paquette and colleagues (2018) identified 13 action patterns of gaming behaviors. Using these patterns, they identified gaming the system behaviors in several learning contexts and platforms with a satisfactory level of reliability. Notably, much of this work focuses on predicting rather than understanding learning behaviors, and when researchers have tried to understand why students game the system, they have looked into self-reports as opposed to utilizing within-program process data (Baker et al., 2008).

There are other examples of using process data to understand behavioral engagement (or disengagement) within mastery learning CALP. For example, “Wheel-Spinning” – spending a substantial amount of time struggling to learn a subject without achieving mastery (Beck & Gong, 2013). As with gaming the system, most of this work has focused on predicting rather than understanding the behavior (Botelho et al., 2019A; Gong & Beck, 2015; Mu et al., 2020; Zhang et al., 2019). Another example is Botelho and colleagues (2019B) exploration of a behavior they called “stopout,” for when students refuse to complete a problem. They used action data prior to the stopout behavior to understand its antecedents.

Although much of behavioral engagement research on mastery learning CALP focuses on negative behaviors of unproductive persistence or disengagement, some researchers have used students’ response time data to estimate attentiveness to problems and hints. One example is Gurung and colleagues (2021) exploration of students’ effort in ASSISTments - a math-focused

CALP that includes mastery learning and traditional problem sets with immediate feedback. They estimated whether students exerted effort after requesting a hint by analyzing students' response times immediately after seeking help in combination with their subsequent action – which they refer to as “response time decomposition.” Shih and colleagues (2008) also examined response times to estimate the time students spent thinking about bottom-out hints (i.e., worked examples that include problems’ answers). Both of these analyses found that more time spent between accessing hints and the student’s next action (requesting another hint or submitting an answer) was associated with better performance, thus they deduced that time to be an indicator of thoughtful effort.

Contrary to much of the research on mastery learning CALP, LA research on behavioral engagement with CALP that focuses on learning processes is often aimed to understand desirable behavioral engagement. Similar to the works of Gurung and Shih, we examined the relationship between pause time and problem solving efficiency in FH2T (Chan et al., 2022). Using students’ response times in conjunction with their derivations, we found that students who spent a higher proportion of pausing time before making a first action used more efficient problem solving strategies.

Another aspect of behavioral engagement explored in LA research is productive persistence in problem-solving. For example, Ventura and Shute (2013) used students’ time on problems to assess their persistence within the physics game, Newton’s Playground, which they validated using an external measure of persistence administered outside of the program. Another way of measuring persistence is to examine how frequently students replay problems. Replayability, which is a feature embedded within many CALP, provides students an opportunity to engage in several attempts to solve problems, often after they have achieved suboptimal performance on prior attempts (Boyce et al., 2011; Liu et al., 2017; Shang et al., 2006).

In FH2T, for example, multiple attempts at each problem are appropriate because there are many different pathways to each solution, and some paths are more optimal than others. This has allowed us to evaluate students' persistence. As mentioned earlier, students receive clovers upon completing a problem, which shows how efficient their solution was, and students are encouraged to replay that problem if their path is suboptimal. Our earlier work found that replaying a problem is associated with a higher likelihood of having optimal performance on the next problem (Liu et al., 2022) and also significantly predicts learning gains (Lee, Chan, et al., 2022).

Further, using FH2T data, we investigated whether specific types of feedback were associated with the number of optional replays students made to solve the problem (Liu et al., 2022; Vanacore, Sales, et al., 2023). In FH2T, students can attempt to make incorrect arithmetic operations or can provide inefficient solutions to the problem. In the first case, the system provides students with error-feedback by shaking the number that the student was trying to put into the incorrect place. In the second case, reward-based feedback – the number of clovers given to students after problem completion – indicates the efficiency of their solution. Our analyses suggest that reward-based feedback can motivate students to attempt problems more than once.

Other studies have used replay behaviors to understand behavioral engagement with the program. For example, Liu and colleagues (2017) examined the relationship between students' replay patterns and their learning gains in a puzzle-like online mathematics game, ST Math. The results provide a nuanced picture of the association between replay behavior and learning. Students who replayed problems immediately after completing the level demonstrated the highest in-game performance, whereas those who replayed problems within their current level had lower learning gains. Similarly, Clark and colleagues (2011) used data on students' replaying levels to understand the nuances of the effects of the physics game on student learning and affect. Notably, they did not find significant correlations between replay behaviors and any of their learning or affective outcomes.

In sum, the studies reviewed in Section 2 provide examples of how LA researchers can use data collected in CALP to understand not only whether the student gets an answer correct but also how the student arrives at a particular solution. When CALP is created to emphasize the process of problem-solving, as opposed to the performance of answering problems, data are produced in a way that helps unpack students' complex cognitive processes when solving problems. Analyses focused on how students arrive at a solution can provide valuable information and insights about implicit students' learning processes.

3. Impact of Focusing on Learning Processes Rather than Mastery: What Works and Why

3.1 Example of Impact: FH2T Efficacy Studies

Throughout the previous section, we have touched upon how LA can help us unpack complexities in students' learning processes using data from both mastery learning and alternative CALP. For this section, we use FH2T as an example of how a CALP focused on students' learning processes has been shown to outperform more traditional methods of instruction focused on mastery learning and performance in improving algebraic knowledge.

As mentioned above, FH2T is a game-based educational technology that aims to improve algebraic understanding (Ottmar et al., 2015). FH2T was developed based on theories of perceptual learning and embodied cognition. Perceptual learning theory posits that algebraic reasoning is inherently perceptual, which involves visual processing: seeing expressions or equations as structured objects, identifying symbols, and organizing them into groups (e.g., parsing 2 × a + b × 3 as (2 × a) + (3 × 4)) (Goldstone et al., 2010). Embodied cognition theories suggest that students' physical experiences influence their thinking and reasoning in mathematics (Abrahamson et al., 2020). Integrating these theories into the game, FH2T made implicit mathematical metaphors and symbols into visual and tactile virtual objects. In this way, students can easily identify implicit algebraic structures through manipulation and transformations of the objects (e.g., touching and moving the symbols) in the game. While many math instructions tend to focus on memorizing abstract and seemingly arbitrary rules (e.g., multiplication and division before addition and subtraction), FH2T provides perceptual-motor experiences that help students acquire not only algebraic knowledge but also appropriate perceptual processing skills essential for fluency in algebraic problem-solving.

One of the integral parts of the game is that students can use any mathematically valid action, resulting in various strategies or pathways to solve each problem. Although there are the most efficient ways that students can solve the problem using the fewest steps possible, students can reach the goal state in a numerous number of mathematically valid ways. In this way, they have the freedom to think flexibly and creatively and realize that math problems can be solved in a large number of ways. Through these interactions, students experience dynamic algebraic transformations, rather than a series of static equations. In sum, FH2T emphasizes the *process of arriving at a mathematical solution* over the correctness of answers or completion of the problems, by providing students with a rich experience with puzzle-like problems that go beyond submitting or selecting correct or incorrect answers.

A number of Randomized Control Trials (RCT) consistently showed that students in the FH2T condition outperformed on algebra assessments compared to students who completed online problem sets, including multiple choice and fill-in-the-blank problem sets with automated hints and feedback (Chan, Lee, et al., 2022; Decker-Woodrow et al., 2023; Hulse et al., 2019). Recently, Decker-Woodrow et al. (2023) conducted a large-scale efficacy RCT to examine whether FH2T improves 7th-grade students' (N = 1,850) algebraic understanding more than two other CALPs: a game-based learning program DragonBox (Kahoot!) and more traditional algebra problems sets presented with hints and immediate feedback (administered through ASSISTments). These conditions were all compared to the active control of traditional algebra problem sets with delayed feedback (also administered through ASSISTments). The results showed that students in the two game-based learning conditions (i.e., FH2T, DragonBox) showed larger learning gains in algebraic understanding compared to the control condition after 4.5 hours of intervention sessions, even after controlling for prior knowledge and demographic variables.

These results indicate that game-based programs which focus on the process of learning have benefits beyond providing more insight into students' learning processes; they also benefit students' learning. Notably, the availability of immediate hints and feedback did not produce significant differences in algebraic knowledge compared with the active control. This may be because hints can alleviate difficulties in the learning process, thus impacting students' performance (Patikorn & Heffernan, 2020; Prihar et al., 2021), while not impacting learning, due to a phenomenon known as the *Assistance Dilemma* (Koedinger & Aleven, 2007). Alternatively, programs that focus on the learning process through game-based tasks may help students engage in productive struggle, by allowing them to explore potential solutions to each problem and seek multiple paths to a solution. Thus, the emphasis on the learning process may create the condition which optimizes learning.

3.2 Providing a Deeper Understanding of Impact

Due to the rich problem-solving data produced as students use FH2T, the data from large efficacy studies provide opportunities to look beyond average treatment effects and understand the causal mechanisms driving these effects. Leveraging quasi-experimental methods, it is possible to use what we have learned about students' processes in correlational research to understand the causal relations between program features and desirable student behaviors as well as between those behaviors and their outcomes. This burgeoning area of work will lead to a better understanding of why some programs outperform others, why some students benefit more than others, and how we might improve programs to maximize learning for all students.

One example of how process data can be used to study the mechanisms of learning is through conducting research on replay behaviors. As explained above, students in FH2T are encouraged to replay problems when they have suboptimal performance, and this behavior is associated with a higher likelihood of performance within the game (Chen et al., 2020; Liu et al., 2022). To go beyond evaluating the association between behaviors and performance, we applied a quasi-experimental method – fully latent principal stratification (Sales & Pane, 2019) – to FH2T efficacy study data in order to estimate the effects of FH2T for students with a high propensity to replay problems (Vanacore et al., 2023). We found that the effect of FH2T was about twice as large for students with a high propensity to replay. This suggests that the ability to replay problems in FH2T is a key mechanism in the effectiveness of the program. Notably, the process orientation of FH2T allows for the replay feature. Unlike traditional mastery learning programs in which students can submit one correct answer, FH2T allows students to take multiple correct paths to the answer. Those who take this opportunity benefit more from the program.

The obvious next step is to understand how to effectively encourage students to replay more problems. To this end, we evaluated the impact of the performance-feedback based systems in FH2T on students' likelihood of replaying each problem after they have suboptimal performance using another quasi-experimental method, regression discontinuity design (Vanacore, Ottmar, et al., 2023). We found that when students received the lowest level performance feedback (a score of one out of three) based on their performance, they were more likely to replay a problem than when they received higher performance feedback (a score of two out of three). We speculate that the students view receiving the lower performance feedback as a game-based -failure, which motivates them to retry the problems. This suggests that adjusting performance-based feedback systems to communicate game-based failures while also providing opportunities to reattempt the problems can encourage students towards productive persistence.

In sum, we have identified a feature of the program (e.g., ability to replay problems) which, when coupled with a behavior (e.g., replaying problems), increases the impact of the CALP on students' learning. Then, we have studied how to influence this behavior within the program. In the future, we can adjust the program in order to encourage replay behavior and test whether this improves the program's efficacy. Thus, this serves as an example of how understanding the learning process can lead to an interactive cycle of improving CALP to help students learn.

4. Conclusions and Future Directions

In order to understand the complexity of students' learning processes, LA has turned to rich data sources produced by CALP to study students' learning processes, including their problem-solving strategies and behavioral engagement. Furthermore, to maximize student outcomes, CALP has provided robust learning environments that go beyond submitting or selecting answers to statically presented problems towards creating dynamic and interactive learning environments. This chapter provided examples of how LA can be used to reveal learning processes and learning-related behaviors. The knowledge gained through this research is then leveraged to create programs that optimize systems for student learning.

Yet to be truly meaningful, this research area must grow in two ways. In their article describing the importance of LA, Wise and colleagues (2021) suggest that impactful LA should focus on (1) "closing the loop" by connecting the learners to actionable interventions through data collection and analyses and (2) creating an iterative cycle which integrates research findings into improving learning systems. They also emphasize that understanding the process of learning using data from programs that emphasize that process is only the beginning for LA. The research on the learning process needs to be connected to interventions for learners and used to improve the programs themselves.

"Closing the loop" involves going beyond learners using a CALP, data collection within the CALP, and analysis/metricization of the data from the CALP. This loop is incomplete without using the analyses and metrics to drive interventions (Clow, 2012). The research presented above does not utilize the analysis of the learning process to differentiate instruction within the programs or help teachers do so in their classrooms. Using our understanding of students' learning processes to drive dynamic systems that tailor instruction to their needs is the necessary and logical next step of this work. Furthermore, helping teachers access understandable data and metrics about their students' strategies and behaviors will help them address their students' misconceptions and provide informed feedback as they teach.

The learning analytics cycle also requires a feedback loop of program improvement (Wise et al., 2021). This is still missing from much of the research on students' learning processes. The field has identified some strategies and behaviors associated with better learning outcomes, and we have started to pinpoint causal factors that connect these strategies with greater impact on learning. Yet, more work should be done to understand how to guide students toward better learning strategies and higher engagement. In Section 3.3, we present one example of a behavior – replaying problems after suboptimal performance – which is both associated with positive outcomes and influenced by a program feature. More work must be done to close the loop between identifying positive learning strategies and behaviors and influencing students such that they develop better strategies and behaviors.

Notably, we focused on problem-solving strategies and behavioral engagement as key constructs of the learning process, as we point out in the introduction, they are not the only relevant constructs. For example, students' affect and mindsets influence how they engage with learning tasks and can be added to the conceptual model proposed in Figure 1. Both of these have been studied at length in LA (e.g., Andres et al., 2019; Baker et al., 2021.; Dillon et al., 2016; Stone et al., 2019; Vanacore et al., 2023; Wang et al., 2015). For example, Baker and colleagues (2012) detected students' affect using derivations of students' action and time data. As with problem-solving strategies and behavioral engagement LA, more work must be done to discern students' affect and mindsets, as well as understand how we might positively influence these aspects of learning.

In conclusion, LA is a field of iterative design and evaluation with the aim of improving learning environments. Researchers and educators know that learning is complex and nuanced. To effectively meet the aim of LA, we must continue to focus on studying learning as a nuanced and complex process. This focus will help researchers and educators understand students as they learn, and ensure that systems are optimally designed to assist the process of learning.

5. Definitions

Action data. Data on the actions students take within a program, including, but not limited to their submission of answers to questions or problems, use of hints and scaffolding, and derivations in problem-solving. These data will vary based upon the actions available in the program.

Computer-assisted learning platforms (CALP). Programs that provide automated learning content and/or problems with the goal of students gaining knowledge or skills. These can include programs with varying levels of adaptivity from intelligent tutors with automated targeted feedback to digitally presented problems.

Desirable difficulties. Elements of learning programs, systems, or courses that create conditions for productive struggle which improve learning.

Learning. Permanent changes in abilities or knowledge, including long term retention of information and transferability of skills outside of the direct context in which they were learned.

Performance. Execution of a task during the learning activity, including whether a problem was answered correctly or a task completed sufficiently.

Randomized Control Trial. A research design in which units (often students) are randomized into conditions, allowing for an evaluation of the effect of those conditions in the unit.

Time data. Data on the time it takes a student to take an action or series of actions.

Quasi-experimental studies. Research designs that estimate the effects of a condition, though the units are not randomized into conditions. This must be done by accounting for confounding that occurs, which influences what units experience what conditions. Common quasi-experimental methods include propensity score matching and regression discontinuity design.

6. References

Adjei, S. A., Baker, R. S., & Bahel, V. (2021). Seven-year longitudinal implications of wheel spinning and productive persistence. *22nd International Conference, AIED 2021*, 16–28. https://doi.org/10.1007/978-3-030-78292-4_2

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, Ray. (1995). Cognitive tutors:

Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207. https://doi.org/10.1207/s15327809jls0402_2

Andres, J. Ma. A. L., Ocumpaugh, J., Baker, R. S., Slater, S., Paquette, L., Jiang, Y., Karumbaiah, S., Bosch, N., Munshi, A., Moore, A., & Biswas, G. (2019). Affect sequences and learning in Betty's brain. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 383–390. https://doi.org/10.1145/3303772.3303807

Archambault, I., Janosz, M., Fallu, J.-S., & Pagani, L. S. (2009). Student engagement and its relationship with early high school dropout. *Journal of Adolescence*, *32*(3), 651–670. https://doi.org/10.1016/j.adolescence.2008.06.007

Baker, R. S. J. D., Corbett, A. T., Koedinger, K. R., & Roll, I. (2006). Generalizing detection of gaming the system across a tutoring curriculum. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 402–411). Springer Berlin Heidelberg. https://doi.org/10.1007/11774303_40

Baker, R. S. J. D, Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G. W., Ocumpaugh, J., & Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra. In *International Educational Data Mining Society.* https://eric.ed.gov/?id=ED537205

Baker, R. S., Nasiar, N., Ocumpaugh, J. L., Hutt, S., Andres, J. M. A. L., Slater, S., Schofield, M., Moore, A., Paquette, L., Munshi, A., & Biswas, G. (2021). Affect-targeted interviews for understanding student frustration. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021 Proceedings, Part I* (pp. 52-63). Cham: Springer International Publishing.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments.

Journal of Interactive Learning Research, *19*(2), 185–224.

Barnes, T., Chi, M., & Feng, M. (2016). MATHia X: The next generation cognitive tutor. *Proceedings of the 9th International Conference on Educational Data Mining*.

Beck, J. E., & Gong, Y. (2013). Wheel-Spinning: Students who fail to master a skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 431–440). Springer. https://doi.org/10.1007/978-3-642-39112-5_44

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. https://doi.org/10.1016/j.jarmac.2020.09.003

Bloom, B. S. (1968). Learning for mastery. Instruction and curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Evaluation Comment*, *1*(2). https://eric.ed.gov/?id=ED053419

Botelho, A. F., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S. A., & Beck, J. E. (2019). Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Transactions on Learning Technologies*, *12*(2), 158–170. https://doi.org/10.1109/TLT.2019.2912162

Botelho, A. F., Varatharaj, V., Van Inwegen, Eric. G., and Heffernan, N. T.. 2019. Refusing to try: Characterizing early stopout on student assignments. In D. Azcona & R. Chung (Eds.), *The 9th International Learning Analytics & Knowledge Conference (LAK19)* (pp. 391-400). ACM. https://doi.org/10.1145/3303772.3303806

Boyce, A., Doran, K., Campbell, A., Pickford, S., Culler, D., & Barnes, T. (2011). BeadLoom Game: Adding competitive, user generated, and social features to increase motivation. *Proceedings of the 6th International Conference on Foundations of Digital Games* (pp. 139–146). https://doi.org/10.1145/2159365.2159384

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Chan, J. Y.-C., Lee, J.-E., Mason, C. A., Sawrey, K., & Ottmar, E. (2022). From Here to There! A dynamic algebraic notation system improves understanding of equivalence in middle-school students. *Journal of Educational Psychology, 114*(1), 56-71. https://doi.org/10.1037/edu0000596

Chan, J. Y.-C., Ottmar, E. R., & Lee, J.-E. (2022). Slow down to speed up: Longer pause time before solving problems relates to higher strategy efficiency. *Learning and Individual Differences, 93*, 102109. https://doi.org/10.1016/j.lindif.2021.102109

Charters, E. (2003). The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education Journal*, *12*(2), Article 2. https://doi.org/10.26522/brocked.v12i2.38

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

Clark, D. B., Nelson, B. C., Chang, H.-Y., Martinez-Garza, M., Slack, K., & D'Angelo, C. M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers & Education*, *57*(3), 2178–2195. https://doi.org/10.1016/j.compedu.2011.05.007

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 134–138.

https://doi.org/10.1145/2330601.2330636

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/BF01099821

Dillon, J., Ambrose, G. A., Wanigasekara, N., Chetlur, M., Dey, P., Sengupta, B., & D'Mello, S. K. (2016). Student affect during learning with a MOOC. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 528–529). https://doi.org/10.1145/2883851.2883960

Fang, Y., Nye, B., Pavlik, P., Xu, Y. J., Graesser, A., & Hu, X. (2017). Online learning persistence and academic achievement. *Proceedings of the 10th International Conference on Educational Data Mining*. EDM 2017, Wuhan, China.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, *2*(2), 265–284. https://doi.org/10.1111/j.1756-8765.2009.01055.x

Gong, Y., & Beck, J. E. (2015). Towards detecting wheel-spinning: Future failure in mastery learning. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (pp. 67–74). https://doi.org/10.1145/2724660.2724673

Gurung, A., Baral, S., Vanacore, K. P., Mcreynolds, A. A., Kreisberg, H., Botelho, A. F., Shaw, S. T., & Hefferna, N. T. (2023). Identification, exploration, and remediation: Can teachers predict common wrong answers? *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 399–410). https://doi.org/10.1145/3576050.3576109

Gurung, A., Lee, M., Baral, S., Sales, A., Vanacore, K., McReynolds, A., Kreisberg, H., Heffernan, C., Haim, A., & Heffernan, N. (2023). How common are common wrong answers? Crowdsourcing remediation at scale. *Learning@Scale 2023*. https://doi.org/10.1145/3573051.3593390

Haim, A., Prihar, E., & Heffernan, N. T. (2022). Toward improving effectiveness of Crowdsourced, on-demand assistance from educators in online learning platforms. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II*, 29–34. https://doi.org/10.1007/978-3-031-11647-6_5

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*, 227–239. https://doi.org/10.1016/j.jml.2009.11.010

Hulse, T., Daigle, M., Manzo, D., Braith, L., Harrison, A., & Ottmar, E. (2019). From here to there! Elementary: A game-based approach to developing number sense and early algebraic understanding. *Educational Technology Research and Development*, 67, 423-441.

Hurwitz, L. B., & Vanacore, K. P. (2022). Educational technology in support of elementary students With reading or language-based disabilities: A cluster randomized control trial. *Journal of Learning Disabilities*, 00222194221141093. https://doi.org/10.1177/00222194221141093

Kelly, K., Wang, Y., Tamisha, T., & Neil, H. (2015). Defining mastery: Knowledge tracing versus consecutive correct responses. *Proceedings of the 8th International Conference on Educational Data Mining*. Educational Data Mining.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, *19*(3), 239–264. https://doi.org/10.1007/s10648-007-9049-0

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, *120*(13), e2221311120.

Koedinger, K. R., Pavlik, P., McLaren, B. M., & Aleven, V. (2008). Is it better to give than to receive? The assistance dilemma as a fundamental unsolved problem in the cognitive science of learning and instruction. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Austin, TX.

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A Meta-Analysis. *Review of Educational Research*, *60*(2), 265–299.

Ladd, G. W., & Dinella, L. M. (2009). Continuity and change in early school engagement: Predictive of children's achievement trajectories from first to eighth grade? *Journal of Educational Psychology*, *101*(1), 190–206. https://doi.org/10.1037/a0013153

Lee, J.-E., Chan, J. Y.-C., Botelho, A., & Ottmar, E. (2022). Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational Technology Research and Development*, *70*(5), 1575–1599. https://doi.org/10.1007/s11423-022-10138-4

Lee, J.-E., Stalin, A., Ngo, V., Drzewiecki, K., Trac, C., & Ottmar, E. (2022). Show the flow: visualizing students' problem-solving processes in a dynamic algebraic notation tool. *Journal of Interactive Learning Research*, *33*(2), 97–126.

Liu, A., Vanacore, K., & Ottmar, E. (2022). *Reward-, but not error-based, feedback systems create micro-failures that support persistence-related learning behaviors [Manuscript Under Review].*

Liu, Z., Cody, C., & Barnes, T. (2017). The antecedents of and associations with elective replay in an educational game: Is replay worth It? *Proceedings of the 10th International Conference on Educational Data Mining*. Educational Data Mining.

Long, P., Siemens, G., Conole, G., & Gašević, D. (2011, February 27). Message from the LAK 2011 General & Program Chairs. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. LAK 2011: 1st International Conference on Learning Analytics and Knowledge, Banff Alberta Canada. https://dl.acm.org/doi/proceedings/10.1145/2090116

Long, P., Siemens, G., Conole, G., & Gašević, D. (2011, February 27). Message from the LAK 2011 General & Program Chairs. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. LAK 2011: 1st International Conference on Learning Analytics and Knowledge, Banff Alberta Canada. https://dl.acm.org/doi/proceedings/10.1145/2090116

Lynch, S. D., Hunt, J. H., & Lewis, K. E. (2018). Productive struggle for all: Differentiated instruction. *Mathematics Teaching in the Middle School*, *23*(4), 194–201. https://doi.org/10.5951/mathteacmiddscho.23.4.0194

Macaruso, P., & Hook, P. E. (2007). Computer assisted instruction: Successful only with proper implementation. *Perspectives on Language and Literacy*, *33*(4).

Martens, RobL., Gulikers, J., & Bastiaens, T. (2004). The impact of intrinsic motivation on e-learning in authentic computer tasks. *Journal of Computer Assisted Learning, 20*(5), 368–376. https://doi.org/10.1111/j.1365-2729.2004.00096.x

McCoy, L. P. (1996). Computer-based mathematics learning. J*ournal of Research on Computing in Education, 28*(4), 438–460. https://doi.org/10.1080/08886504.1996.10782177

Mu, T., Jetten, A., & Brunskill, E. (2020). Towards suggesting actionable interventions for wheel-spinning students. *Proceedings of The 13th International Conference on Educational Data Mining*. EDM 2020.

Okano, H., Hirano, T., & Balaban, E. (2000). Learning and memory. *Proceedings of the National Academy of Sciences*, *97*(23), 12403–12404. https://doi.org/10.1073/pnas.210381897

Paquette, L., & Baker, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. *Interactive Learning Environments*, *27*(5–6), 585–597. https://doi.org/10.1080/10494820.2019.1610450

Paquette, L., Baker, R. S., & Moskal, M. (2018). A System-general model for the detection of gaming the system behavior in CTAT and LearnSphere. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. Du Boulay (Eds.). *Artificial Intelligence in Education: 19th International Conference* (pp. 257–260). Springer International Publishing. https://doi.org/10.1007/978-3-319-93846-2_47

Paquette, L., Baker, R. S., & Ocumpaugh, J. (2014). Reengineering the feature distillation process: A case study in the detection of gaming the system. In J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren (Eds.), *Proceedings of 7th International Conference on Educational Data Mining* (pp. 284–287). ACM.

Pardos, Z. A., Baker, R. S. J. D., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, *1*(1), 107–128. https://doi.org/10.18608/jla.2014.11.6

Park, S. (2023). Discovering unproductive learning patterns of wheel-spinning students in intelligent tutors using cluster analysis. *TechTrends*, *67*(3), 489–497. https://doi.org/10.1007/s11528-023-00847-9

Patikorn, T., & Heffernan, N. T. (2020). Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale* (pp. 115–124). https://doi.org/10.1145/3386527.3405912

Prihar, E., Patikorn, T., Botelho, A., Sales, A., & Heffernan, N. (2021). Toward personalizing students' education with crowdsourced tutoring. *L@S 2021 - Proceedings of the 8th ACM Conference on Learning @ Scale* (pp. 37–45). https://doi.org/10.1145/3430895.3460130

Prihar, E., Syed, M., & Ostrow, K. (2022). *Exploring common trends in online educational experiments*. 12. In A.Mitrovic & N. Bosch (Eds), *Proceedings of the 15th Internationa Conferenceon Educational Data Mining* (pp. 27–38). International Educational Data Mining Society.

Reich, J. (2020). Algorithm-Guided Learning At Scale. In *Failure to Disrupt: Why Technology Alone Can't Transform Education* (pp. 47–76). Harvard University Press.

Richey, J. E., Zhang, J., Das, R., Andres-Bray, J. M., Scruggs, R., Mogessie, M., Baker, R. S., & McLaren, B. M. (2021). Gaming and confusion explain learning advantages for a math digital learning game. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *In Artificial Intelligence in Education: 22nd International Conference* (pp. 342–355). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_28

Ritter, S., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016). How mastery learning works at scale. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 71–79. https://doi.org/10.1145/2876034.2876039

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*(5), 1323–1330. https://doi.org/10.3758/s13423-014-0588-3

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, *2*(4), 2332858416673968. https://doi.org/10.1177/2332858416673968

Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R., Bardar, E., & Gasca, S. (2021). Assessing implicit computational thinking in Zoombinis puzzle gameplay. *Computers in Human Behavior*, *120*, 106707. https://doi.org/10.1016/j.chb.2021.106707

Rutherford, T., Kibrick, M., Burchinal, M., Richland, L., Conley, A., Osborne, K., Schneider, S., Duran, L., Coulson, A., Antenore, F., Daniels, A., & Martinez, M. E. (2010). *Spatial temporal mathematics at scale: An innovative and fully developed paradigm to boost math achievement among all learners*. In Online Submission. https://eric.ed.gov/?id=ED510612

Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J., Richland, L., Tran, N., Schneider, S., Duran, L., & Martinez, M. E. (2014). A randomized trial of an elementary school mathematics software intervention: Spatial-Temporal Math. *Journal of Research on Educational Effectiveness*, *7*(4), 358–383. https://doi.org/10.1080/19345747.2013.856978

Sales, A. C., & Pane, J. F. (2019). The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics*, *13*(1), 420–443. https://doi.org/10.1214/18-AOAS1196

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 179–187. https://doi.org/10.1037/0278-7393.5.2.179

Siew, N. M., Geofrey, J., & Lee, B. N. (2016). Students' Algebraic Thinking and Attitudes towards Algebra: The Effects of Game-Based Learning using Dragonbox 12 + App. *The Research Journal of Mathematics and Technology, 5*(1). https://doi.org/2163-0380

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The Challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition*, *6*, 342–353. https://doi.org/10.3758/BF03197465

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1582–1593. https://doi.org/10.1037/xlm0000019

Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction*, *1*, 341–358. https://doi.org/10.1207/s1532690xci0103_4

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Stone, C., Quirk, A., Gardener, M., Hutt, S., Duckworth, A. L., & D'Mello, S. K. (2019). Language as thought: Using natural language processing to model noncognitive traits that predict college success. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 320–329. https://doi.org/10.1145/3303772.3303801

Sun, C., Shute, V. J., Stewart, A. E. B., Beck-White, Q., Reinhardt, C. R., Zhou, G., Duran, N., & D’Mello, S. K. (2022). The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior*, *128*, 107120. https://doi.org/10.1016/j.chb.2021.107120

Vanacore, K., Ottmar, E., Sales, A., & Liu, A. (2023). Evaluating In-Program Decisions by Leveraging Cut Points and Regression Discontinuity Analysis for Causal Inference. In (Sales, A., Chair) *Causal Modeling of Log Data in Edu Tech Symposium*. National Council on Measurement in Education's Annual Meeting, Chicago, IL. USA.

Vanacore, K. P., Gurung, A., McReynolds, A. A., Liu, A., Shaw, S. T., & Heffernan, N. T. (2023). Impact of non-cognitive interventions on student learning behaviors and outcomes: An analysis of seven large-scale experimental inventions. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 10. https://doi.org/10.1145/3576050.3576073

Vanacore, K., Sales, A., Liu, A., & Ottmar, E. (2023). Benefit of gamification for persistent learners: Propensity to replay problems moderates algebra-game effectiveness. *Tenth ACM Conference on Learning @ Scale (L@S' 23)*. Learning @ Scale, Copenhagen, Denmark. https://doi.org/10.1145/3573051.3593395

Wang, Y., Heffernan, N. T., & Heffernan, C. (2015). Towards better affect detectors: Effect of missing skills, class features and common wrong answers. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 31–35). https://doi.org/10.1145/2723576.2723618

Wise, A. F., Knight, S., & Ochoa, X. (2021). What makes learning analytics research matter. *Journal of Learning Analytics*, *8*(3), 1–9. https://doi.org/10.18608/jla.2021.7647

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial*

Intelligence in Education (pp. 171–180). Springer. https://doi.org/10.1007/978-3-642-39112-5_18

Zhang, C., Huang, Y., Wang, J., Lu, D., Fang, W., Fancsali, S., Holstein, K., & Aleven, V. (2019). Early detection of wheel spinning: Comparison across tutors, models, features, and operationalizations. *Proceedings of The 12th International Conference on Educational Data Mining*. (pp. 468 - 473). International Educational Data Mining Society.

Chapter 2

Heterogeneous Effects for Disengaged Students

2.1 Effect of Assistance on Gaming The System**

See manuscript below.

The Effect of Assistance on Gamers: Assessing The Impact of On-Demand Hints & Feedback Availability on Learning for Students Who Game the System

Kirk Vanacore*
kpvanacore@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

Ashish Gurung*
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
agurung@andrew.cmu.edu

Adam C. Sales
asales@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
nth@wpi.edu

ABSTRACT

Gaming the system, characterized by attempting to progress through a learning activity without engaging in essential learning behaviors, remains a persistent problem in computer-based learning platforms. This paper examines a simple intervention to mitigate the harmful effects of gaming the system by evaluating the impact of immediate feedback on students prone to gaming the system. Using a randomized controlled trial comparing two conditions - one with immediate hints and feedback and another with delayed access to such resources - this study employs a Fully Latent Principal Stratification model to determine whether students inclined to game the system would benefit more from the delayed hints and feedback. The results suggest differential effects on learning, indicating that students prone to gaming the system may benefit from restricted or delayed access to on-demand support. However, removing immediate hints and feedback did not fully alleviate the learning disadvantage associated with gaming the system. Additionally, this paper highlights the utility of combining detection methods and causal models to comprehend and effectively respond to students' behaviors. Overall, these findings contribute to our understanding of effective intervention design that addresses gaming the system behaviors, consequently enhancing learning outcomes in computer-based learning platforms.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**; • **Human-centered computing** → *Empirical studies in interaction design.*

KEYWORDS

Computer-Based Learning Platforms, Gaming the System Detection, Causal Inference, Feedback, Hints

ACM Reference Format:

Kirk Vanacore*, Ashish Gurung*, Adam C. Sales, and Neil T. Heffernan. 2024. The Effect of Assistance on Gamers: Assessing The Impact of On-Demand Hints & Feedback Availability on Learning for Students Who Game the System. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3636555.3636904

1 INTRODUCTION

Researchers in Learning Analytics (LA) and associated fields have exerted immense effort to identify students' latent states as they use computer-based learning platforms (CBLP). For example, LA researchers often seek to determine when students are gaming the system – attempting to progress through a learning activity without learning [11, 19, 44–46]. Furthermore, extensive literature exists on understanding why students exhibit these behaviors [7–9]. Other studies have evaluated interventions that may reduce the frequency and migrate the effects of gaming to the system behaviors [10, 40, 53, 62, 64]. However, many of these interventions focus on changing game behaviors instead of improving learning outcomes. Thus, it is unclear which interventions help students who tend to game the system to engage with and learn from CBLPs.

In the current paper, we seek to isolate the learning impact of immediate feedback on students who game the system. More specifically, this study addresses whether students who game the system in a traditional CBLP that includes multiple-choice and open-response questions with immediate hints and feedback would respond differently to a CBLP in which they do not have access to those hints and feedback while solving problems. Furthermore, the paper provides an example of incorporating predictions from detection models into causal models. The combination of detection and causal models allows us to go beyond identifying and understanding behaviors towards knowing how to respond in a way that positively impacts learning.

2 BACKGROUND

2.1 Gaming The System

Gaming the system behavior is defined as "attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that

*Authors contributed equally to this publication.

LAK '24, March 18–22, 2024, Kyoto, Japan
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636904

knowledge to answer correctly" [8]. This behavior is associated with reduced performance within CBLP [9] and poor learning outcomes [17]. It is also predictive of poor distal outcomes such as state test performance [48] and low college enrollment [2].

The two key behaviors indicative of gaming the system are rapid and repeated requests for help (hint abuse) and submission of answers in a systemic way (guess-and-check) [8, 62]. In sum, these behaviors suggest that a student is trying to get the answer to individual problems and progress through the assignment without employing the requisite effort to learn from the activity. Gaming the system tends to be associated with student frustration rather than general disengagement [8], as it is not correlated with off-task behavior [9].

Some research indicates that specific program features may cause gaming behaviors [7]. Such findings suggest that students move from states of confusion and frustration towards gaming the system behaviors because the CBLP's features do not adequately address their learning needs [7, 54]. Ambiguity and abstractness within CBLP's activities are critical factors associated with gaming behaviors [7]. This evidence, combined with research suggesting that frustration may be causing students to game the system, suggests reducing problem difficulty and adding supports to alleviate frustration may reduce gaming behaviors [8]. Alternatively, additional assistance features, such as hints and feedback, may provide more opportunities to abuse assistance features, further dissuading some students from engaging in a learning process requiring persistence and effort.

2.2 Mitigating Gaming the System Behaviors

Attempts to mitigate gaming behavior generally fall into either proactive or reactive categories. Proactive interventions include dissuading students from unnecessary hint usage [3, 40] or providing visualizations that allow students to see and interpret their behaviors [62, 64]. Reactive options implementation interventions after the gaming behavior has been identified [10].

Research suggests that proactively dissuading students from unnecessarily requesting hints may be too blunt to address a problem from one population of students. Telling students that they should only use hints if they truly needed them while slightly delaying hint availability (10-second pause) reduced hint usage for all students and did not improve overall performance [40]. However, this method may have improved the performance of the lowest-knowledge students; the sample size was small, and the effect was marginally insignificant. Nevertheless, the finding is suggestive that manipulating access to hints could help students who game the system.

In theory, presenting students with information about their gaming-related behaviors can have two potential effects. It can indicate to the student that the system is logging their behaviors, thus creating "panopticon-like paranoia" of constant awareness of potential observation [62]. Visualizing the student's actions may also nudge students to reflect on their learning behaviors [64]. Two studies found that presenting students with graphical representations of their behaviors can reduce gaming the system behaviors [62, 64]. However, neither of these studies evaluated whether their intervention improved students' learning. Furthermore, since students are likely to game the system to progress through learning activities, it is unclear that they will not also game the system to manipulate the outputs of the visualizations.

As a reactive option, [10] placed students who had previously gamed the system on specific content into an intervention focused on that content. In this intervention, students saw an animated dog that displayed emotions that aligned with the student's learning-related behaviors. The animation served as a primitive mirror of emotions that a human tutor might display while working alongside a student (e.g., excitement and positivity if a student exerted effort or frustration if a student gamed the system). The intervention reduced students' gaming behaviors and positively affected their learning outcomes as measured by a post-test compared to a control group who did not have access to the animated dog.

Overall, directly dissuading students from gaming behaviors like hint abuse may have adverse effects on the students' learning in general while potentially benefiting lower-performing students. Alternatively, presenting visualizations that represent students' actions may reduce their gaming behaviors. However, the learning impact of this method has not been explored. Notably, only the study that imposed a direct content-specific intervention showed a significant impact on learning [10]. Thus, none of these studies connected the prevention of gaming behaviors with learning outcomes.

2.3 Immediate Assistance in CBLPs

Immediate assistance meant to provide "just in time" instruction to students is a cornerstone of many CBLPs. The efficacy of immediate hints and feedback has been studied extensively. Timely feedback and support during learning activities benefit students' learning outcomes [14, 56]. Studies suggest that receiving feedback immediately after giving responses or completing problem sets might be effective for improving students' procedural and conceptual knowledge [18, 22, 50]. In CALPs, often, this assistance takes the form of hints and feedback accessible on demand as students work on problems, which can vary in form and focus. Common hints and feedback modes include presenting general topical information [4], worked examples where a student is shown a complete solution to a similar problem [38], providing the complete solution to the given problem [63], being shown similar examples done incorrectly [1, 38], being given targeted feedback based on a students' common wrong answers [27, 28], and being given a series of step-by-step hints [24]. One study found that worked examples improved students' efficacy in learning but not overall learning outcomes [38]. Another study, which used machine learning to generate explanations and deploy them to learners, found that students performed better when presented with explanations as opposed to only receiving the answer [63].

Beyond the benefits to academic performance, optional tutoring strategies in intelligent tutoring systems have the added benefit of encouraging help-seeking behaviors, which can increase students' autonomy and control of their learning [3, 5]. However, the efficacy of optional assistance is contingent upon students having the requisite metacognitive skills to evaluate when they need help. One study found that access to general information aimed at helping students learn concepts, such as a glossary, was often ignored; students prefer specialized hints that focus on their current problem [4]. These findings suggest that students are not focused on learning the broad skills associated with the individual tasks but on getting the information they need to improve immediate performance. Another study found that students required support to utilize on-demand learning assistance but failed to find learning gains even when this support was given [5].

Overall, on-demand instruction is a common component of intelligent tutoring systems, with varied implementations and varied levels of efficacy. Tutoring strategies, which provide targeted support for specific problems, are effective at improving student performance [49, 51]. Research has also shown that this feedback is most effective if provided as the student is answering questions [37]. Although there is ample research on how type and focus influence immediate hints and feedback efficacy, more work is needed to understand who benefits from these resources. Furthermore, the interplay between access to on-demand learning supports, how students use these supports, and learning outcomes has not been fully explored.

3 CURRENT STUDY

The problem of hint abuse and guess-and-check behaviors suggests that some students are not benefiting from the availability of hints and feedback commonly embedded in problems in many CBLPs. However, the solution to this problem is unclear as other research suggests that providing on-demand assistance may alleviate frustration, an ostensible root cause of gaming the system, thus mitigating the behavior [8]. Alternatively, frustration and confusion may be precursors to learning as they represent a student's awareness that they have yet to master the knowledge component being taught. Thus, gaming the system behaviors could be a crutch for some students that allows them to avoid the productive struggle necessary to learn. If this hypothesis is true, removing the program features that allow for hint abuse and guess-and-check may help some students who would have gamed the system engage with the content and learn.

The current study seeks to address this issue directly using a subset of data from a randomized controlled trial that compared methods of teaching pre-algebra concepts of expression equivalence to test whether the effect of feedback varies based on students gaming the system behavior [20]. Our work focuses on two conditions: traditional multiple-choice and open-response problem sets with immediate or delayed hints and feedback (*Immediate Condition* and *Delayed Condition*). In the Immediate Condition, students had access to hints and could see whether their submitted answers were correct while completing the problem sets. In the Delayed Condition, students could only access the hints and feedback after they completed each problem set. (Section 4.1 contains complete descriptions of these conditions.)

This study aims to test whether students who game the system when they access immediate hints and feedback behavior would benefit from delaying access to those resources until they complete the entire problem set. Essentially, the delayed condition removes students' abilities to engage in hint abuse because they did not see the hints during the activity and guess-and-check because they could only submit their answers once. We hypothesize that students with a high propensity to game the system in the Immediate Condition will employ better learning behaviors in the Delayed Condition and thus learn more in that condition. This research question and hypothesis, as well as the methods we employ to answer them, were preregistered on OSF[1]. We will test our hypothesis by estimating whether the effect of immediate feedback varies by student propensity to game the system using a Fully Latent Principal Stratification (FLPS) model, which estimates causal effects for subgroups that emerge during an intervention of a randomized controlled trial [35, 61].

This work provides two contributions to the LA community. First, it addresses when and for whom hints and feedback are effective. Second, the work provides an example of combining the detection of students' behaviors with causal inference in a way that may be leveraged for effective personalization in the future. As the field continues to use artificial intelligence and machine learning to detect and predict students' latent states (e.g., affect, knowledge component mastery, wheel-spinning, etc.), we must also consider how to adjust learning systems based on these predictions. This objective requires understanding which conditions will positively impact students when they are determined to be gaming the system, wheel spinning, confused, etc. Thus, the combination of methods used in this paper, discussed in detail below, may be deployed to help future researchers understand what actionable steps will be impactful after detecting and predicting students' latent states.

4 METHOD

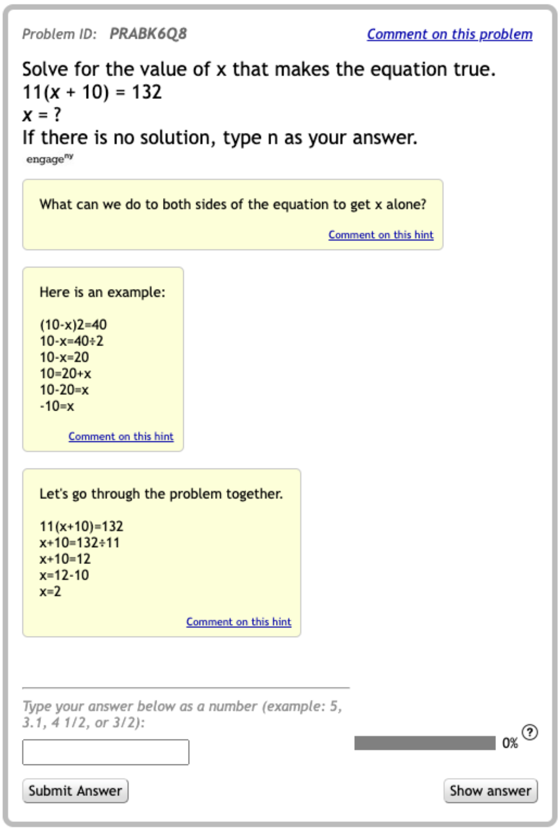
4.1 Conditions

The two conditions included in this paper were administered through ASSISTments [30], an online homework system that provides feedback to students as they solve traditional textbook problems. The problem sets in ASSISTments are adapted from open-source curricula, thus resembling problems students encounter in their textbooks and homework assignments. ASSISTments presents students with problems one at a time on their screen. Each condition included 218 problems of the same problems selected from three curricula – *EngageNY*, *Utah Math*, and *Illustrative Math* – to address specific algebra skills related to procedural knowledge, conceptional knowledge, and flexibility. The problems were divided into nine problem sets and administered in nine half-hour sessions during school hours.

4.1.1 Immediate Feedback and Hints (Immediate Condition). In the Immediate condition, students could request three hints and receive feedback on whether their answers were correct immediately after submitting each answer. Each problem contained a series of hints with a similar structure. An example problem is displayed in Figure 1. The first hint gave the students the first step for answering the problem. The second hint gave the student a worked example of a similar problem. The final hint was a bottom-out hint, which provided the student with the steps to complete the problem as well as the problem's solution. Students could submit as many answers

[1] https://osf.io/jf25x/; This paper only includes one of the analyses/hypotheses included in the preregistration. We are still testing the other hypotheses.

as needed but could not move on until they entered the correct answer.



Problem ID: PRABK6Q8 Comment on this problem

Solve for the value of x that makes the equation true.
11(*x* + 10) = 132
x = ?
If there is no solution, type n as your answer.
engage^{ny}

What can we do to both sides of the equation to get x alone?
Comment on this hint

Here is an example:

(10-x)2=40
10-x=40÷2
10-x=20
10=20+x
10-20=x
-10=x

Comment on this hint

Let's go through the problem together.

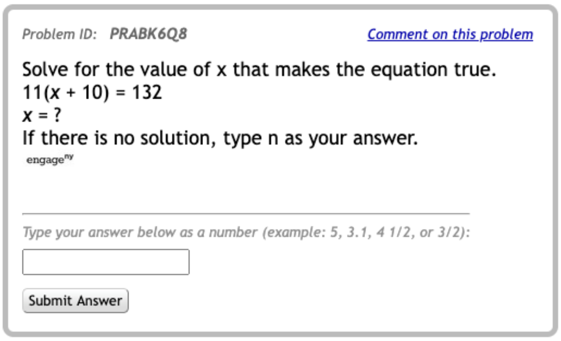
11(x+10)=132
x+10=132÷11
x+10=12
x=12-10
x=2

Comment on this hint

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

Figure 1: Immediate Feedback Condition Example

4.1.2 Problem Sets With Post-Assignment Feedback (Delayed Condition) . The Delayed Condition provided post-assignment feedback rather than immediate hints and feedback. Figure 2 presents an example problem from the Delayed Condition. In this condition, problem sets were administered in "test mode," so students did not receive any feedback or hints during problem-solving. They could only submit one answer and progressed through the problem set without any feedback on their performance. Students received a report with feedback on their accuracy at the end of each problem set, through which they could review their responses, revisit problems, and request hints.



Problem ID: PRABK6Q8 Comment on this problem

Solve for the value of x that makes the equation true.
11(*x* + 10) = 132
x = ?
If there is no solution, type n as your answer.
engage^{ny}

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

Figure 2: Delayed Feedback Condition Example

In the delayed feedback condition, hints and correctness feedback were not provided during the problem set. The students in this condition were provided with a report at the end of each problem set through which they had access to their accuracy and hints on the problems.

4.2 Analysis Plan

The fundamental question of this paper – whether the student who game the system in one condition will benefit from the other condition – poses a methodological difficulty because the conditions likely confound the behavior of gaming the system. In fact, we hypothesize that students who game the system in the Immediate Condition will not game the system in the Delayed Condition and thus benefit from learning when they engage with content. To address this methodological problem, we propose that students have a baseline propensity to game the system in the Immediate Condition that exists prior to randomization, regardless of whether it can manifest itself after treatment assignment. Because it is considered a baseline covariate and potential moderator, similar to students' pretest knowledge, it is independent of the random treatment assignment. However, unlike pretest knowledge, only students in the treatment condition have the opportunity to display the behavior; therefore, its value in the control condition is unknown. Nevertheless, once this latent propensity to game the system is estimated, we can also estimate whether and to what extent it moderates the treatment effect of the various interventions.

Our analyses require two key steps, which are described in depth below. First, we identify instances where students are gaming the system within the Immediate Condition. Then, we use the causal method of Fully Latent Principal Stratification (FLPS), which will allow us to estimate the effect heterogeneity of each condition based on students' latent propensity to game the system [55]. These methods are delineated in the sections below.

4.2.1 Implementation of Gaming The System Detector. This paper employs the rule-based gaming detectors originally proposed by [47] to identify gaming behavior among students working on algebra problems. To the best of the authors' knowledge, this *Cognitive Model* developed by [47] was the first of its kind as the detector was transferable across intelligent tutoring systems from Cognitive Tutor Algebra (CTA) to ASSISTments. The rule-based detector was initially engineered to detect gaming in CTA [12]. [44] extended the implementation of the rule-based gaming detector by relying on human judgments through text replays of logged learner actions [13]. The insights garnered from this cognitive model were subsequently used to develop gaming detectors for CTA and validate the transferability of rule-based gaming detectors by implementing them onto ASSISTments.

We employed the rule-based *Cognitive Model* for identifying gaming behavior as these models mitigate key challenges such as enhancing generalizability and controlling for detector rot. Unlike other system specific gaming detectors, the rule-based model demonstrated cross system generalizability, underscoring its adaptability and reliability in different contexts [47]. Detector rot is a phenomenon that refers to the gradual decline in a model's performance over time. Prior studies have reported that more complex models, using more advanced Machine Learning and Deep Learning algorithms, are more prone to this phenomenon than their simpler counterparts [34, 36]. Given that many gaming the system detection models were developed years before our study, they are at a higher risk of experiencing detector rot. Therefore, we opted for the rule-based detector, in light of its potential for sustained generalizability and resistance to detector rot due to its simple yet effective detection of gaming behavior.

We employed the rule-based detector to identify gaming behavior among students using the Immediate Condition. The rule-based detector system uses log-filed data aggregated in twenty-second clips (Section 5.1 describes the variables used) and indicates whether students were gaming during each time clip. We then used these indicators in the FLPS model described below.

4.2.2 Estimating Effects Using Fully Latent Principal Stratification. FLPS is a variant of Principal Stratification, a causal inference method used in randomized controlled trials for estimating the intervention effects on subgroups that emerge after the treatment has begun [25, 43]. Traditional estimation of effects for subgroups requires that these subgroups are defined before intervention and be independent of any treatment. For instance, in the case of pretest knowledge, simply interacting the treatment with the pretest knowledge score provides information about how the treatment effect varies across subgroups of students with similar prior knowledge. However, subgroups defined based on students' interactions with a treatment program cannot be observed at baseline and are never observed for students randomized to the control condition. Often, program implementation consists of a complex sequence of users' behaviors or choices. FLPS models these behaviors as manifestations of latent student characteristics, which are not directly observed for students randomized to the treatment condition either but must be estimated.

This method is particularly relevant in CBLPs, in which students may display an array of behaviors during the program such as meeting implementation goals [21, 60], productive persistence [61], mastering knowledge components in mastery learning [55], and gaming-the-system behaviors that may be viewed as indicators of latent student characteristics (i.e., high fidelity users, persistent learners, mastery users, gamers). These behaviors are unobserved for the control groups who do not have the same opportunity to use, game, or master as they did not interact with the same CBLP. When observable, they are confounded by the students' condition. Therefore, their underlying latent student characteristics can be interpreted as explaining students' behaviors if randomized to the treatment condition. For example, [55] determined whether the effect of Cognitive Tutor Algebra I on students varies based on whether students were likely to master knowledge components by estimating the likelihood of mastering knowledge components for the treatment group as a latent variable. In the current study, we evaluate whether the effect of the Immediate Condition differs based on students' latent propensity to game the system had they been assigned to that condition.

Let τ_i be subject i's individual treatment effect: the difference between what i's posttest score would be if i were randomized to treatment and their score if randomized to control. Since students' gaming the system behavior was detected in the Immediate Condition, our goal is to know how students with a high propensity to game the system would fare if placed in the Delayed Condition. For simplicity, we refer to the Immediate Condition as the treatment and the Delayed Condition as the control. Let $\mathcal{T}$ and C be the samples of students randomized to treatment and control, respectively. Let α_{ti} be i's propensity to game the system if randomized to the treatment condition. α_{ti} is defined for $i \in C$, as students in the control condition still had a potential to game the system had they been randomized to treatment, even if this potential was never realized. Therefore, α_{ti} is estimated from gaming the system behavior for $\mathcal{T}$ and α_{ti} is imputed for C.

The principal effect is the treatment effect for the subgroup of students with a particular value for α_t:

$$\tau(\alpha) = E[\tau | a_t = \alpha] \tag{1}$$

To estimate the function $\tau(\alpha)$, we (1) estimate α_t for $\mathcal{T}$ as a function of pre-treatment covariates observed in both groups, (2) use that model to impute α_t for C, (3) estimate $\tau(\alpha)$ by including a treatment interaction in a linear regression model. The models are estimated using iterations through these steps in a Bayesian principal stratification model with a continuous variable consisting of measurement and outcome submodels, as outlined by [31] and [42].

4.2.3 Measurement Submodel: Modeling Gaming the System Behavior. First, we estimate α_t by running a multilevel logistic submodel predicting whether the gaming detector identified the students in the treatment condition to have gamed the system on each twenty-second time clip as delineated in the equation 2. Let G_{cji} be a binary indicator of whether student i gamed the system during time-clip c when working on problem j. Let P_{ki} be covariate predictor k of K student-level predictors, which are measured at baseline for both $\mathcal{T}$ and C (described in section 5.2). Let the random intercepts be μ_j for problems, μ_i for students, μ_t for teachers, and μ_s schools, each modeled as independent and with normal distributions with means of 0 and standard deviations estimated from the data.

$$logit(G_{cji}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k \mathrm{P}_{ki} + \mu_j + \mu_i + \mu_t + \mu_s \tag{2}$$

Using the parameters from equation 2, students' propensity to game the system is defined as

$$\alpha_i = \sum_{k=1}^{K} \gamma_k \mathrm{P}_{ki} + \mu_i + \mu_t + \mu_s \tag{3}$$

We impute α_i for C with random draws from a normal distribution with mean $\sum_k \gamma_k \mathrm{P}_{ki} + \mu_{t[i]} + \mu_{s[i]}$, where $\mu_{t[i]}$ and $\mu_{s[i]}$ are the random intercepts for student i's teacher and school, respectively, and standard deviation equal to the estimated standard deviation of μ_i. Note that randomization occurred at the student level (i.e. teachers had students in the treatment and control in their classes). Therefore, we include the random intercepts for schools and teachers from submodel 2 in valuation for α_{ti} for students in the control. However, μ_i is unknown for C, but we can assume its distribution is the same in the two conditions because of the randomization.

4.2.4 Outcomes Submodel: Modeling Posttests $\tau(a)$. To estimate the treatment effect for students with differing propensities to game the system, we run a multilevel linear regression predicting student's post-test algebraic knowledge (Y_i). The submodel includes interaction between α_{ci}—estimated for $\mathcal{T}$ and randomly imputed for C—and Z_i, an indicator of being in the treatment condition (Immediate Condition). Let the random effects for teacher be v_t and for school be v_s.

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 \alpha_{ti} + \beta_3 \alpha_{ti} Z_i + \sum_{k=1}^{K} \lambda_k P_{ki} + v_t + v_s + \epsilon_i \quad (4)$$

Using the parameters from the submodel 4, the treatment effect for students with a particular propensity to game the system is modeled as

$$\tau(\alpha) = \beta_1 + \beta_3 \alpha \quad (5)$$

Submodels (2) and (4) together formed a FLPS model, which we fit using the Stan Markov Chain Monte Carlo software through STAN [6].

5 DATA & VARIABLES

The data for this study exists at two different levels. There are student-level variables, including the student-level predictors and the learning outcome. Alternatively, the data used in gaming the system detector and the detector's output is aggregated in twenty-second clips of the students usage of the program. Only time-clip data from the Immediate Condition is used. All data from the study are available through OSF[2] and a full explanation of the data can be found in [41].

The sample consists of 779 students: 394 in the Immediate Condition and 385 in the Delayed Condition. The students were taught by 34 teachers, in 9 schools. In the Immediate Condition, students completed 96,311 problems across 107,577 clips.

5.1 Gaming the System Detector Data and Output

Gaming labels were generated by adopting the methodology described by [47] to produce action clips for the gaming detector, as explained in Section 4.2.1. These clips capture sequences of actions taken by students in ASSISTments. Each clip contains unique identifiers: the student working on the problem, the problem(s) being worked on, the skill associated with the problem, and the problem type. Additionally, the clips detail the start and end times of actions, their total duration, and, for attempts, the action's correctness and the student's answer. Supplementary data within the clips include indicators for hint requests, the number of hints requested during the clip, the total hints available for the problem, the use of a 'bottom-out' hint, and the total attempts by the student. It is important to note that the dataset had an indicator for scaffolding support as well; however, the problems analyzed in our study did not implement scaffolding support. Gaming labels produced by the *Cognitive Model* detector were generated based on instances where students' actions in the clip met one or more rules indicative of gaming system behavior.

5.2 Pretreatment Predictors

To estimate students' propensity to game the system, we used demographic data and data from assessments administered prior to their use of their assigned condition. Pretest scores were collected by the original studies' researchers: algebraic knowledge, math anxiety, and perceptual processing skills. Algebraic knowledge was measured using a variant of the learning outcome described below (Section 5.3). The math anxiety assessment was adapted form from the *Math Anxiety Scale for Young Children-Revised* [16], which assessed negative reactions towards math, numerical inconfidence, and math-related worrying (Cronbach's α=.87; see the items on OSF[3]). Five items adapted from the Academic Efficacy Subscale of the *Patterns of Adaptive Learning Scale* to assess math self-efficacy ([39] Cronbach's α = .82; see items on OSF[4]). The perceptual processing assessment evaluates students' ability to detect mathematically equivalent and nonequivalent expressions as quickly as possible [23, 32] (see item on OSF[5]). Log forms of assessment test times were also included in the models. We included polynomials of the pretest scores when they improved model fit. The school district in which the original study was conducted provided students' demographic data—race/ethnicity, individualized education plan status (IEP), and English as a second or foreign language (ESOL) status. The district also provided students' most recent standardized state test scores in math. Race and ethnicity were dummy-coded, with white students as the reference category. We standardized (z-score) all continuous scores to improve model fit and ease interpretation. Missing data were imputed using singly-imputation with the Random Forest routine implemented by the missForest package in R [52, 59]. The number of missing data for each student was included as a predictor in each submodel.

5.3 Learning Outcome

The learning outcome for the study is students' algebraic knowledge, which was assessed using ten multiple-choice items from a previously validated measure of algebra understanding (Cronbach's α = .89; see the items on OSF[6]) [58]. Four of the items focused on conceptual understanding of algebraic equation-solving (e.g., the meaning of an equal sign), three focused on procedural skills of equation-solving (e.g., solving for a variable), and three focused on flexibility of equation-solving strategies (e.g., evaluating different equation-solving strategies). These ten items together assessed students' knowledge in algebraic equation-solving, the improvement of which was the goal of the interventions. The assessment

[2] https://osf.io/r3nf2/

[3] https://osf.io/rq9d8

[4] https://osf.io/rq9d8

[5] https://osf.io/r47ev

[6] https://osf.io/uenvg

was taken before and after the intervention. Students' scores were standardized to ease model fit and interpretation.

6 RESULTS

6.1 Gaming Detector

The students in the Immediate feedback condition produced 89,960 twenty-second clips of data. The gaming detector estimated that students gamed the system during 4.62% of the clips. A majority of students (94.18%) gamed the system at least once. Overall, students averaged two gaming clips (mean = 5.85, SD = 4.11), but the distribution is skewed, such that 39.05% of the total gaming system behavior was attributed to the students in the upper quartile of gaming frequency. This suggests that some students have a higher propensity to game the system than others.

Notably, there were many problems (74.35%) in which no student gamed the system at all. There was a small significant negative correlation between the gaming rate on each problem and the average accuracy on the problem (r = -0.21, p < 0.001). Furthermore, some problem types had significantly higher rates of gaming behaviors than others (F[2,244] = 4.69 $p < 0.001$). Students were more likely to game the system on 'check all that apply' problems compared with problems they had to submit a number ($p < 0.001$), variable ($p = 0.038$), an algebraic expression ($p = 0.044$), or submit an open response ($p < 0.029$). Students were also more likely to game the system on a multiple choice question than problems in which they had to submit a numeric answer ($p = 0.035$). Because the problems without gaming behavior would not be informative for assessing students gaming the system, we did not include these problems in the measurement model.

6.2 Fully Latent Principal Stratification

We ran 11,000 iterations of FLPS models using Markov chain Monte Carlo chains calling *Stan* through *rstan* [57] in *R*; code is posted on GitHub[7]. We evaluated convergence using trace plots and $\hat{R}$. The maximum $\hat{R}$ for estimated parameters was 1.01. Table 1 provides parameter estimates and relevant statistics for the measurement and outcome submodels.

6.2.1 *Measurement Submodel.*

Prior to the simultaneous estimation of the FLPS sub-models in STAN, we ran the measurement model using the *stan_glmer* function from the *rstanarm* package [26] with different combinations and transformations of the pretest predictors and demographic variables to find a suitable model for the analysis. While building the measurement model, we split the data into training (80%) and testing (20%) data sets. The model performed best when pretest sub-scores from the tests were included, so we included many pretest sub-scores in the final model. Our final model produced an acceptable AUC of 0.87 on the training data set.

The measurement model provides some notable predictors of gaming the system behavior. Students with lower scores on the math section of the state test were more likely to game the system ($\gamma_{17} = -0.31$, $P(<0) = .99$). This effect was consistent with the albeit smaller associations with the algebraic knowledge sub-scores. There was a nonlinear association between math anxiety and gaming behavior. Students with higher math anxiety were also less likely to game the system ($\gamma_6 = -0.23$, $P(<0) = .97$) and this effect became greater in magnitude as high math anxiety increased, as shown the effect associated with the math anxiety squared ($\gamma_7 = -0.04$, $P(<0) = .92$). Students with a higher negative reaction toward math ($\gamma_8 = 0.19$, $P(>0) = .99$) and with higher numeric confidence ($\gamma_9 = 0.13$, $P(>0) = .96$) were more likely to game the system. Times on the algebraic knowledge test and the perceptual sensitivity learning subtests were all predictive of gaming the system behavior. Students who took less time on these tests were more likely to game the system.

6.2.2 *Outcomes Submodel.*

The outcomes model provided evidence of an interaction between gaming behavior and the feedback conditions, suggesting that while feedback is likely effective for students with a low propensity to game the system, it is likely ineffective for those with a high propensity to game the system. The main effect for the Immediate Condition was likely positive ($\beta_1 = 0.06$, $P(>0) = .90$). Notably, the effect was small – 6% of a standard deviation – suggesting that while hints and feedback play a role in the effectiveness of CBLP other program components also contribute to its effectiveness. As expected, students with a high propensity to game the system performed substantially worse than those with a low propensity to game the system ($\beta_2 = -0.37$, $P(<0) > .99$).

The interaction between students' propensity to game the system and the Immediate Condition was likely negative ($\beta_3 = -0.11$, $P(<0) > .93$). Table 2 presents estimated average effects for students in each quartile of the propensity to game the system (α). The students at the bottom quartile of gaming the system propensity experienced an estimated positive effect from the Immediate Condition of 0.18 SD of algebraic knowledge. In contrast, those at the top quartile of gaming the system propensity experienced an estimated negative effect of -0.02 SD of algebraic knowledge. This finding implies that students who engage in gaming the system behavior may benefit from the delayed condition, whereas those with a lower propensity to game the system likely benefit from the Immediate Condition.

7 DISCUSSION & CONCLUSION

The findings of the study indicate that the impact of on-demand hints and feedback on student performance in a CBLP varies widely. This disparity in outcomes may be attributed to how students utilize assistance features. Those who exploit hints excessively or rely on trial-and-error methods to complete assignments would potentially benefit from restricted or delayed access to immediate hints and feedback. Nevertheless, even when immediate hints and feedback were eliminated (as in the Delayed Condition), the decrease in performance associated with gaming behavior was not completely alleviated. Therefore, these results suggest that while removing on-demand instruction may assist students inclined towards gaming the system, further intervention is required to fully mitigate the detrimental effects of such behavior or address the root causes behind it.

Although the delayed hints and feedback condition was originally intended to be an active control in this study, it can be viewed as a proactive intervention targeting gaming the system behavior. This approach, similar to others mentioned in previous research

[7] https://github.com/kirkvanacore/FLPS_GamingTheSystem

Table 1: Fully Latent Principal Stratification Model Parameter Estimates

	Measurement Submodel				Outcomes Submodel			
Predictors	Estimate	SD	P(>0)	P(<0)	Estimate	SD	P(>0)	P(<0)
Z					0.06	0.05	0.90	0.10
α_t					-0.37	0.15	0.01	0.99
$\alpha_t : Z$					-0.11	0.08	0.07	0.93
Algebraic Procedural Knowledge	-0.04	0.05	0.17	0.83	-0.01	0.04	0.42	0.58
Algebraic Conceptual Knowledge	-0.08	0.06	0.10	0.90	0.13	0.05	0.99	0.01
Algebraic Flexibility Knowledge	-0.03	0.04	0.25	0.75	-0.03	0.04	0.23	0.77
Algebraic Knowledge Items Complete	-0.03	0.08	0.36	0.64	0.03	0.06	0.72	0.28
Algebraic Knowledge Time (Log)	-0.04	0.04	0.14	0.86	-0.04	0.03	0.06	0.94
Math Anxiety	-0.23	0.12	0.03	0.97	0.07	0.10	0.78	0.22
Math Anxiety (Squared)	-0.04	0.03	0.08	0.92	-0.02	0.02	0.23	0.77
Math Negative Reaction	0.19	0.08	0.99	0.01	-0.04	0.07	0.27	0.73
Math Numerical Confidence	0.13	0.08	0.96	0.04	-0.06	0.06	0.16	0.84
Math Self Efficacy	-0.04	0.04	0.21	0.79	0.03	0.04	0.81	0.19
Perceptual Sensitivity Score Part 1	-0.05	0.04	0.11	0.89	0.00	0.03	0.54	0.46
Perceptual Sensitivity Time Part 1 (Log)	-0.10	0.06	0.03	0.97	-0.00	0.04	0.50	0.50
Perceptual Sensitivity Score Part 2	0.01	0.05	0.60	0.40	0.11	0.04	0.99	0.01
Perceptual Sensitivity Time Part 2 (Log)	-0.06	0.05	0.10	0.90	-0.01	0.04	0.34	0.66
Perceptual Sensitivity Score Part 3	0.05	0.05	0.82	0.18	-0.04	0.04	0.15	0.85
Perceptual Sensitivity Time Part 4 (Log)	-0.08	0.06	0.09	0.91	0.07	0.04	0.95	0.05
State Test Score	-0.31	0.06	0.01	0.99	0.03	0.06	0.68	0.32
Female	-0.01	0.04	0.38	0.62	0.02	0.03	0.73	0.27
Hispanic	0.11	0.14	0.78	0.22	0.06	0.10	0.72	0.28
Asian/Pacific Islander	-0.15	0.13	0.11	0.89	0.11	0.10	0.88	0.12
Black	0.32	0.20	0.94	0.06	0.17	0.15	0.87	0.13
IEP	-0.02	0.04	0.28	0.72	0.00	0.03	0.54	0.46
EIP	0.01	0.04	0.55	0.45	0.00	0.03	0.53	0.47
ESOL	0.01	0.05	0.54	0.46	0.01	0.04	0.65	0.35
Gifted	-0.02	0.04	0.33	0.68	0.11	0.03	0.99	0.01
In-person Instruction	0.03	0.05	0.72	0.28	-0.16	0.07	0.02	0.98
Missing Data	0.01	0.07	0.58	0.41	-0.00	0.05	0.49	0.51

Table 2: Mean effects of the Immediate Condition (τ) and each quartile of propensity to game the system (α)

	Mean	
Quartile	α	τ
1	-1.05	0.18
2	-0.23	0.09
3	0.28	0.03
4	0.76	-0.02

[3, 40], employs a standardized approach for all students. The differential effect observed for immediate hints and feedback supports [40]'s suggestion that deterring certain students from using hints may be beneficial for some students despite the overall negative impact on the student population. Our finding supports this hypothesis.

One potential solution to the impact differential could involve disabling immediate hints and feedback for students identified as gaming the system, thereby offering a targeted intervention to redirect their focus toward learning from the activity. This solution would not only allow students who do not game the system to benefit from the on-demand assistance, but it could also allow students who game the system to benefit from this assistance when not gaming the system. This type of adaptive system may also mitigate some of the negative effects of the gaming behavior by allowing those who game to experience the best of both interventions (immediate and delayed). However, it is essential to acknowledge that implementing such an approach may foster frustration and disengagement. Further investigation is needed to test these hypotheses.

The measurement model parameters that predict gaming the system suggest that intricate factors contribute to this behavior. One possible scenario is that students with low knowledge but high confidence resort to gaming the system after encountering failure in the activities, which contradicts their perceived self-efficacy. It may seem contradictory that students' overall math anxiety is negatively associated with gaming the system behavior, whereas negative reactions towards math in general are positively associated with it. However, it is plausible that math-anxious students approach problems cautiously, while those with negative reactions toward math may prioritize completing the assignment quickly. These findings contribute to existing literature, highlighting that attributing gaming the system solely to general disengagement may be too simplistic, as it likely stems from various underlying causes [8].

Time spent on pretests was associated with gaming the system; those who took more time on the pretests were less likely to game the system. This finding is not necessarily surprising as gaming is associated with rapid behaviors [8, 44]. However, it is still notable that this behavior may be evident in the testing context. Thus, gaming the system may be indicative of general rushing behavior. Analyses of pause time have suggested that students who take more time may be exerting more effort [29] and performing better [15, 33] than those respond quickly after starting a problem. Together, these findings indicate that gaming behaviors may be interrelated with other leaner profiles, which are also associated with heterogeneity in learning outcomes.

Finally, this paper showcases the potential of combining detection and causal methods in the field of LA to gain a deeper understanding of appropriate actions following the identification of specific behaviors or latent states. Artificial intelligence driven detection within CBLPs often leaves learning experience designers with "what next" questions (e.g. "What should we do now that we know a student is frustrated?"). FLPS provides one solution by combining the output of detectors with causal models to address which program features will differently benefit students who exhibit specific behavior patterns. Although in the current analysis we use a rule-based detection method, the integration of artificial intelligence prediction systems with FLPS holds promise for not only assessing students' experiences and actions in CBLPs but also suggesting optimal adaptations within these programs to maximize their learning impact.

8 LIMITATIONS & FUTURE DIRECTIONS

Although our findings suggest that the impact of feedback may vary depending on students' inclination to game the system, it is important to acknowledge the limitations of this analysis. First, there remains some uncertainty regarding whether the main and interaction effects in the model significantly differ from zero. Sampling from the posterior distribution indicated a 90% certainty that the main effect (i.e., the effect of Immediate Feedback for those with an α of zero) was greater than zero, and a 93% certainty that the slope associated with the propensity to game the system (α) was less than zero. However, it is still possible that the observed effects may be smaller in magnitude than those presented here. Further research is needed to validate and replicate these results to establish the robustness of our findings.

Although this analysis provides information about who is likely to game the system and under what circumstances, it does not fully address questions related to the profiles of students who game the system. Our measurement model finds some corollaries to gaming behavior but does not provide a robust delineation of learner profiles for students who are likely to game the system. More work is needed in this area. Similarly, we found a negative correlation between the rate of gaming and the average accuracy on each problem, but we need a more robust understanding of why students are gaming on specific problems. There is a notable reciprocal relation between gaming the system and accuracy, which may explain this relationship. Some problem types had higher associations with gaming behavior than others, such as 'check all that apply' and multiple choice problems. Future work should seek to understand how problem difficulty and type influence the likelihood that students will game the system.

Additionally, it is essential to recognize the limitations of FLPS. The effects estimated using FPLS rely on the underlying quality of the model itself, and the extent to which errors in the model estimation may introduce bias to the effect remains unclear. More work is necessary to develop a comprehensive understanding of how to evaluate these models and ensure they provide unbiased estimates of treatment effects.

ACKNOWLEDGMENTS

The authors would like to thank past and current including NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (R305D210036, R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305A120125 & R305R220012), GAANN (P200A120238, P200A180088, P200A150306 & P200A150306), EIR (U411B190024 S411B210024, & S411B220024), ONR (N00014-18-1-2768), NIH (via SBIR R44GM146483), Schmidt Futures, BMGF, CZI, Arnold, Hewlett and a $180,000 anonymous donation. None of the opinions expressed here are those of the funders.

REFERENCES

[1] Deanne M. Adams, Bruce M. McLaren, Kelley Durkin, Richard E. Mayer, Bethany Rittle-Johnson, Seiji Isotani, and Martin van Velsen. 2014. Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior* 36 (Jul 2014), 401–411. https://doi.org/10.1016/j.chb.2014.03.053

[2] Seth A Adjei, Ryan S Baker, and Vedant Bahel. 2021. Seven-year longitudinal implications of wheel spinning and productive persistence. In *22nd International Conference, AIED 2021.* Springer, The Netherlands, 16–28. https://doi.org/10.1007/978-3-030-78292-4_2

[3] Vincent Aleven. 2001. *Helping students to become better help seekers: Towards supporting metacognition in a cognitive tutor.* Technical Report. German-USA Early Career Research Exchange Program: Research on Learning Technologies and Technology-Supported Education, Tubingen, Germany.

[4] Vincent Aleven and Kenneth R. Koedinger. 2000. Limitations of Student Control: Do Students Know when They Need Help?. In *Intelligent Tutoring Systems (Lecture Notes in Computer Science)*, Gilles Gauthier, Claude Frasson, and Kurt VanLehn (Eds.). Springer, Berlin, Heidelberg, 292–303. https://doi.org/10.1007/3-540-45108-0_33

[5] Vincent Aleven, Ido Roll, Bruce M. McLaren, and Kenneth R. Koedinger. 2016. Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education* 26, 1 (Mar 2016), 205–223. https://doi.org/10.1007/s40593-015-0089-1

[6] Dorsa Mohammadi Arezooji. 2020. A Markov Chain Monte-Carlo Approach to Dose-Response Optimization Using Probabilistic Programming (RStan). (2020).

[7] R. Baker, A. Carvalho, Jay Raspat, Vincent Aleven, and K. R. Koedinger. 2009. Educational Software Features that Encourage and Discourage "Gaming the System". In *Proceedings of the 14th international conference on artificial intelligence in education*, Vol. 14. IOS Press, Washington, D.C., 475–482. http://pact.cs.cmu.edu/koedinger/pubs/Baker%2C%20de%20Carvalho%2C%20Raspat%2C%20Aleven%2C%20Corbett%20Koedinger%20AIED09.pdf

[8] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.

[9] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. 2004. Off-task behavior in the cognitive tutor classroom: When students" game the system". In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 383–390.

[10] Ryan S. J. d. Baker, Albert T. Corbett, Kenneth R. Koedinger, Shelley Evenson, Ido Roll, Angela Z. Wagner, Meghan Naim, Jay Raspat, Daniel J. Baker, and Joseph E. Beck. 2006. Adapting to When Students Game an Intelligent Tutoring System. In *Intelligent Tutoring Systems: 8th International Conference*, Mitsuru

Ikeda, Kevin D. Ashley, and Tak-Wai Chan (Eds.). Springer, Jhongil, Taiwan, 392–401. https://doi.org/10.1007/11774303_39

[11] Ryan S. J. D. Baker, Albert T. Corbett, Kenneth R. Koedinger, and Ido Roll. 2006. *Generalizing Detection of Gaming the System Across a Tutoring Curriculum*. Lecture Notes in Computer Science, Vol. 4053. Springer Berlin Heidelberg, Berlin, Heidelberg, 402–411. https://doi.org/10.1007/11774303_40

[12] Ryan S. J. d. Baker, Albert T. Corbett, Ido Roll, and Kenneth R. Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18, 3 (Aug 2008), 287–314. https://doi.org/10.1007/s11257-007-9045-6

[13] Ryan S. J. d. Baker, Antonija Mitrović, and Moffat Mathews. 2010. Detecting Gaming the System in Constraint-Based Tutors. In *User Modeling, Adaptation, and Personalization (Lecture Notes in Computer Science)*, Paul De Bra, Alfred Kobsa, and David Chin (Eds.). Springer, Berlin, Heidelberg, 267–278. https://doi.org/10.1007/978-3-642-13470-8_25

[14] Andrew C. Butler and Nathaniel R. Woodward. 2018. *Toward consilience in the use of task-level feedback to promote learning*. Vol. 69. Academic Press, 1–38. https://doi.org/10.1016/bs.plm.2018.09.001

[15] Jenny Yun-Chen Chan, Erin R. Ottmar, and Ji-Eun Lee. 2022. Slow down to speed up: Longer pause time before solving problems relates to higher strategy efficiency. *Learning and Individual Differences* 93 (Jan 2022), 102109. https://doi.org/10.1016/j.lindif.2021.102109

[16] Lian-Hwang Chiu and Loren L Henry. 1990. Development and validation of the Mathematics Anxiety Scale for Children. *Measurement and evaluation in counseling and development* 23, 3 (1990), 121–127.

[17] Mihaela Cocea and Arnon Hershkovitz. 2009. The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? *Frontiers in Artificial Intelligence and Applications* 200 (2009). https://doi.org/10.3233/978-1-60750-028-5-507

[18] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4, 4 (Dec 1994), 253–278. https://doi.org/10.1007/BF01099821

[19] Steven Dang and Ken Koedinger. 2019. *Exploring the Link between Motivations and Gaming*. https://eric.ed.gov/?id=ED599218 ERIC Number: ED599218.

[20] Lauren E Decker-Woodrow, Craig A Mason, Ji-Eun Lee, Jenny Yun-Chen Chan, Adam Sales, Allison Liu, and Shihfen Tu. 2023. The impacts of three educational technologies on algebraic understanding in the context of COVID-19. *AERA open* 9 (2023), 23328584231165919.

[21] Kevin C. Dieter, Jamie Studwell, and Kirk P. Vanacore. 2020. Differential Responses to Personalized Learning Recommendations Revealed by Event-Related Analysis.. In *International Conference on Educational Data Mining (EDM)*, Vol. 13. ERIC, Online. https://eric.ed.gov/?id=ED607826

[22] Roberta E. Dihoff, Gary M. Brosvic, and Michael L. Epstein. 2003. The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record* 53, 4 (2003), 533–548.

[23] Bye J. K.; Lee J. E.; Chan J. Y. C.; Closser A. H.; Shaw S. T.; Ottmar E. 2022. Toward Improving Effectiveness of Crowdsourced, On-Demand Assistance from Educators in Online Learning Platforms. In *Poster presented at the annual meeting of the American Educational Research Association (AERA)*.

[24] Mingyu Feng and Neil T Heffernan. 2006. Informing teachers live about student learning: Reporting in the assistment system. *Technology Instruction Cognition and Learning* 3, 1/2 (2006), 63.

[25] Constantine E. Frangakis and Donald B. Rubin. 2002. Principal Stratification in Causal Inference. *Biometrics* 58, 1 (2002), 21–29.

[26] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. 2020. rstanarm: Bayesian applied regression modeling via Stan. https://mc-stan.org/rstanarm R package version 2.21.1.

[27] Ashish Gurung, Sami Baral, Morgan P Lee, Adam C Sales, Aaron Haim, Kirk P Vanacore, Andrew A McReynolds, Hilary Kreisberg, Cristina Heffernan, and Neil T Heffernan. 2023. How Common are Common Wrong Answers? Crowdsourcing Remediation at Scale. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*. 70–80. https://doi.org/10.1145/3573051.3593390

[28] Ashish Gurung, Sami Baral, Kirk P Vanacore, Andrew A Mcreynolds, Hilary Kreisberg, Anthony F Botelho, Stacy T Shaw, and Neil T Hefferna. 2023. Identification, Exploration, and Remediation: Can Teachers Predict Common Wrong Answers?. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 399–410. https://doi.org/10.1145/3576050.3576109

[29] Ashish Gurung, Anthony F. Botelho, and Neil T. Heffernan. 2021. Examining Student Effort on Help through Response Time Decomposition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. ACM, Irvine CA USA, 292–301. https://doi.org/10.1145/3448139.3448167

[30] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (Oct 2014), 470–497. https://doi.org/10.1007/s40593-014-0024-x

[31] Hui Jin and Donald B Rubin. 2008. Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* 103, 481 (2008), 101–111.

[32] David Kirshner and Thomas Awtry. 2004. Visual Salience of Algebraic Transformations. *Journal for Research in Mathematics Education* 35, 4 (2004), 224–257. http://www.jstor.org/stable/30034809

[33] Ji-Eun Lee, Jenny Yun-Chen Chan, Anthony Botelho, and Erin Ottmar. 2022. Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational technology research and development* 70, 5 (Oct 2022), 1575–1599. https://doi.org/10.1007/s11423-022-10138-4

[34] MP Lee, E Croteau, A Gurung, A Botelho, and N Heffernan. 2023. Knowledge Tracing Over Time: A Longitudinal Analysis.. In *The Proceedings of the 16th International Conference on Educational Data Mining*.

[35] Sooyong Lee, Sales Adam, Hyeon-Ah Kang, and Tiffany A Whittaker. 2022. Fully Latent Principal Stratification: Combining PS with Model-Based Measurement Models. In *The Annual Meeting of the Psychometric Society*. Springer, 287–298.

[36] Nathan Levin, Ryan Baker, Nidhi Nasiar, Fancsali Stephen, and Stephen Hutt. 2022. Evaluating Gaming Detector Model Robustness Over Time. In *Proceedings of the 15th International Conference on Educational Data Mining, International Educational Data Mining Society*.

[37] Xiwen Lu, Wei Wang, Benjamin A. Motz, Weibing Ye, and Neil T. Heffernan. 2023. Immediate text-based feedback timing on foreign language online assignments: How immediate should immediate feedback be? *Computers and Education Open* 5 (Dec 2023), 100148. https://doi.org/10.1016/j.caeo.2023.100148

[38] Bruce M. McLaren, Tamara Van Gog, Craig Ganoe, Michael Karabinos, and David Yaron. 2016. The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computers in Human Behavior* 55 (Feb 2016), 87–99. https://doi.org/10.1016/j.chb.2015.08.038

[39] Carol Midgley, Martin L Maehr, Ludmila Z Hruda, Eric Anderman, Lynley Anderman, Kimberley E Freeman, T Urdan, et al. 2000. *Manual for the patterns of adaptive learning scales*. Ann Arbor: University of Michigan. 734–763 pages.

[40] R Charles Murray and Kurt VanLehn. 2005. Effects of Dissuading Unnecessary Help Requests While Providing Proactive Help.. In *AIED*. 887–889.

[41] Erin Ottmar, Ji-Eun Lee, Kirk Vanacore, Siddhartha Pradhan, Lauren Decker-Woodrow, and Craig A. Mason. 2023. Data from the Efficacy Study of From Here to There! A Dynamic Technology for Improving Algebraic Understanding. *Journal of Open Psychology Data* 11, 1 (Apr 2023), 5. https://doi.org/10.5334/jopd.87

[42] Lindsay C Page. 2012. Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness* 5, 3 (2012), 215–244.

[43] Lindsay C. Page, Avi Feller, Todd Grindal, Luke Miratrix, and Marie-Andree Somers. 2015. Principal Stratification: A Tool for Understanding Variation in Program Effects Across Endogenous Subgroups. *American Journal of Evaluation* 36, 4 (Dec 2015), 514–531. https://doi.org/10.1177/1098214015594419

[44] Luc Paquette. 2014. Towards Understanding Expert Coding of Student Disengagement in Online Learning.

[45] Luc Paquette and Ryan S. Baker. 2017. Variations of Gaming Behaviors Across Populations of Students and Across Learning Environments *(Lecture Notes in Computer Science)*, Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 274–286. https://doi.org/10.1007/978-3-319-61425-0_23

[46] Luc Paquette and Ryan S. Baker. 2019. Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. *Interactive Learning Environments* 27, 5–6 (Aug 2019), 585–597. https://doi.org/10.1080/10494820.2019.1610450

[47] Luc Paquette, Ryan S. Baker, Adriana de Carvalho, and Jaclyn Ocumpaugh. 2015. Cross-System Transfer of Machine Learned and Knowledge Engineered Models of Gaming the System. In *User Modeling, Adaptation and Personalization (Lecture Notes in Computer Science)*, Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless (Eds.). Springer International Publishing, Cham, 183–194. https://doi.org/10.1007/978-3-319-20267-9_15

[48] Zachary A. Pardos, Ryan S. J. D. Baker, Maria O. C. Z. San Pedro, Sujith M. Gowda, and Supreeth M. Gowda. 2014. Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics* 1, 1 (2014), 107–128. ERIC Number: EJ1127034.

[49] Thanaporn Patikorn and Neil T. Heffernan. 2020. Effectiveness of Crowd-Sourcing On-Demand Assistance from Teachers in Online Learning Platforms. In *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale*. Association for Computing Machinery, 115–124. https://doi.org/10.1145/3386527.3405912

[50] Gary D. Phye and Thomas Andre. 1989. Delayed retention effect: Attention, perseveration, or both? *Contemporary Educational Psychology* 14, 2 (Apr 1989), 173–185. https://doi.org/10.1016/0361-476X(89)90035-0

[51] Ethan Prihar, Thanaporn Patikorn, Anthony Botelho, Adam Sales, and Neil Heffernan. 2021. Toward Personalizing Students' Education with Crowdsourced Tutoring. In *L@S 2021 - Proceedings of the 8th ACM Conference on Learning @ Scale*. Association for Computing Machinery, Inc, 37–45. https://doi.org/10.1145/3430895.3460130

[52] R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[53] J. Elizabeth Richey, Jiayi Zhang, Rohini Das, Juan Miguel Andres-Bray, Richard Scruggs, Michael Mogessie, Ryan S. Baker, and Bruce M. McLaren. 2021. Gaming and Confrustion Explain Learning Advantages for a Math Digital Learning Game. In *In Artificial Intelligence in Education: 22nd International Conference (Lecture Notes in Computer Science)*, Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova (Eds.). Springer International Publishing, Utrecht, The Netherlands, 342–355. https://doi.org/10.1007/978-3-030-78292-4_28

[54] Ma. Mercedes T. Rodrigo, Ryan S. J. d. Baker, Sidney D'Mello, Ma. Celeste T. Gonzalez, Maria C. V. Lagud, Sheryl A. L. Lim, Alexis F. Macapanpan, Sheila A. M. S. Pascua, Jerry Q. Santillano, Jessica O. Sugay, Sinath Tep, and Norma J. B. Viehland. 2008. Comparing Learners' Affect While Using an Intelligent Tutoring System and a Simulation Problem Solving Game. In *Intelligent Tutoring Systems (Lecture Notes in Computer Science)*, Beverley P. Woolf, Esma Aïmeur, Roger Nkambou, and Susanne Lajoie (Eds.). Springer, Berlin, Heidelberg, 40–49. https://doi.org/10.1007/978-3-540-69132-7_9

[55] Adam C. Sales and John F. Pane. 2019. The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics* 13, 1 (Mar 2019), 420–443. https://doi.org/10.1214/18-AOAS1196

[56] Valerie J. Shute. 2008. Focus on Formative Feedback. *Review of Educational Research* 78, 1 (Mar 2008), 153–189. https://doi.org/10.3102/0034654307313795

[57] Stan Development Team. 2016. RStan: the R interface to Stan. http://mc-stan.org/ R package version 2.14.1.

[58] Jon R Star, Courtney Pollack, Kelley Durkin, Bethany Rittle-Johnson, Kathleen Lynch, Kristie Newton, and Claire Gogolen. 2015. Learning from comparison in algebra. *Contemporary Educational Psychology* 40 (2015), 41–54.

[59] Daniel J. Stekhoven and Peter Buehlmann. 2012. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.

[60] Kirk Vanacore, Erin Ottmar, Allison Liu, and AC Sales. 2023. Remote Monitoring of Implementation Fidelity Using Log-File Data from Multiple Online Learning Platforms. (2023). https://doi.org/10.31234/osf.io/7ru2x

[61] Kirk Vanacore, Adam Sales, Allison Liu, and Erin Ottmar. 2023. Benefit of Gamification for Persistent Learners: Propensity to Replay Problems Moderates Algebra-Game Effectiveness. In *Tenth ACM Conference on Learning @ Scale (L@S '23)*. ACM, Copenhagen, Denmark. https://doi.org/10.1145/3573051.3593395

[62] Jason A Walonoski and Neil T Heffernan. 2006. Prevention of off-task gaming behavior in intelligent tutoring systems. In *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings 8*. Springer, 722–724.

[63] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale (L@S '16)*. Association for Computing Machinery, New York, NY, USA, 379–388. https://doi.org/10.1145/2876034.2876042

[64] Meng Xia, Yuya Asano, Joseph Jay Williams, Huamin Qu, and Xiaojuan Ma. 2020. Using Information Visualization to Promote Students' Reflection on "Gaming the System" in Online Learning. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S '20)*. Association for Computing Machinery, New York, NY, USA, 37–49. https://doi.org/10.1145/3386527.3405924

2.2 Effect of Gamification on Gaming the System**

See manuscript below.

Effect of Gamification on Gamers: Evaluating Interventions for Students Who Game the System

Kirk P. Vanacore
Worcester Polytechnic Institute
Worcester, MA, USA
kpvanacore@wpi.edu

Ashish Gurung
Carnegie Mellon University
Pittsburgh, PA, USA
agurung@andrew.cmu.edu

Adam C. Sales
Worcester Polytechnic Institute
Worcester, MA, USA
asales@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, MA, USA
nth@wpi.edu

Gaming the system is a persistent problem in Computer-Based Learning Platforms. While substantial progress has been made in identifying and understanding such behaviors, effective interventions remain scarce. This study uses a method of causal moderation known as Fully Latent Principal Stratification to explore the impact of two types of interventions – gamification and manipulation of assistance access – on the learning outcomes of students who tend to game the system. The results indicate that gamification does not consistently mitigate these negative behaviors. One gamified condition had a consistently positive effect on learning regardless of students' propensity to game the system, whereas the other had a negative effect on gamers. However, delaying access to hints and feedback may have a positive effect on the learning outcomes of those gaming the system. This paper also illustrates the potential for integrating detection and causal methodologies within educational data mining to evaluate effective responses to detected behaviors.

Keywords: gamification, gaming the system, causal inference, computer-based learning platforms

1. Introduction

Gaming the system – attempting to progress through a learning activity without learning (Baker et al., 2008) – is an enduring problem that reduces the efficacy of Computer-Based Learning Platforms (CBLPs). Educational Data Mining (EDM) researchers have made substantial progress in identifying instances of gaming the system behaviors (Paquette et al., 2014; Paquette and Baker, 2017; Paquette and Baker, 2019; Dang and Koedinger, 2019; Baker et al., 2006), and further research has explored the antecedents of these behaviors (Baker et al., 2004; Baker et al., 2008; Baker et al., 2009); however, solutions remain scarce. Although interventions to address gaming the system have been evaluated (Richey et al., 2021; Walonoski and Heffernan, 2006; Murray and VanLehn, 2005; Xia et al., 2020; Baker et al., 2006), most focus on dissuading students from gaming the system instead of improving learning outcomes. Thus, it is unclear which interventions help students who tend to game the system engage with and learn from CBLPs.

This paper explores two types of interventions – gamification and manipulations of access to assistance – which could hypothetically benefit students who tend to game the system. In general, gamification – the act of infusing game features into a system – is seen as a potential method of increasing engagement within CBLPs (Garris et al., 2002; Gaston and Cooper, 2017; Vanacore et al., 2023a). By increasing engagement, gamification could theoretically decrease gaming the system behaviors and help students benefit from the CBLPs. However, features of gamification could exacerbate these behaviors by providing more opportunities to game. Therefore, simpler alternatives may also be effective, such as restricting access to the assistance that is often abused by gamers[1] (Murray and VanLehn, 2005). The current research on gaming the system interventions has not fully addressed which interventions are most effective for gamers. Thus, it is unclear what changes in CBLP environments can positively influence these students' learning behaviors and outcomes.

In this paper, we explore the impact of key differences in learning platforms on students who tend to game the system. More specifically, this study addresses whether students who game the system in a traditional CBLP – one that includes various closed-ended and open-ended questions with immediate hints and feedback – would respond differently to alternative CBLP environments: two gamified CBLPs and a traditional CBLP in which the access to hints and feedback were delayed until the end of each activity. We find that gamification does not consistently mitigate the negative effects of gaming the system on learning, but students who tend to game the system may benefit from delayed hints and feedback.[2]

As a secondary objective, we present an example of integrating prediction from detection models into causal models. We utilize a method of causal moderation – Fully Latent Principal Stratification (Sales and Pane, 2019)– that can leverage detection model outputs to understand heterogeneity in treatment effects. The combination of detection and causal models provides opportunities to go beyond identifying students' behaviors and/or latent states (e.g., confusion or knowledge component mastery) to understand how to respond in a way that positively impacts learning.

2. Background

In this section, we provide literature reviews of the relevant aspects of the study. First, we include a review of the suspected antecedents and outcomes associated with gaming behaviors (Section 2.1). Next, we consider the theoretical and empirical support that immediate assistance features (Section 2.2) and gamification (Section 2.3) influence students' learning-related behaviors and outcomes. Finally, we review research on interventions aimed at mitigating gaming behaviors (Section 2.4).

[1] Throughout this paper, we use the term 'gamers' to mean students who tend to game the system and 'gaming behaviors' as behaviors indicative of gaming the system.

[2] The finding that delayed hints and feedback are beneficial for students who tend to game the system was first reported in Vanacore et al. (2024). The current paper extends these analyses to include two other conditions.

2.1. Gaming The System

Gaming the system is defined as "attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly" (Baker et al., 2008). This definition is operationalized by two key behaviors: rapid and repeated requests for help (hint abuse) and submission of answers in a systemic way (guess-and-check) (Baker et al., 2008; Walonoski and Heffernan, 2006). These behaviors indicate that a student is trying to submit the answers to problems merely to progress through the assignment without deploying the effortful engagement required to learn from the activity.

Therefore, it is not surprising that gaming the system behaviors are correlated with reduced performance within CBLPs (Baker et al., 2004) and poor learning outcomes (Mihaela and Hershkovitz, 2009). Furthermore, gaming is also predictive of low distal outcomes such as state test performance (Pardos et al., 2014) and college enrollment (Adjei et al., 2021). Thus, gaming behaviors may be related to students' broader educational struggles.

However, gaming the system tends to have a stronger association with student frustration rather than general disengagement (Baker et al., 2008), as it is not correlated with off-task behavior (Baker et al., 2004). Specific program features may cause gaming behaviors by fomenting confusion and frustration (Baker et al., 2009). Research suggests that students move from states of confusion and frustration towards gaming the system behaviors because the CBLP's features do not adequately address their learning needs (Rodrigo et al., 2008; Baker et al., 2009). When the CBLP's activities are ambiguous and abstract, students are more likely to game the system (Baker et al., 2009). Thus, it is plausible that decreasing problem difficulty and adding supports to mitigate frustration may reduce gaming (Baker et al., 2008). Alternatively, additional assistance features may dissuade students from persistent and effortful engagement with the content while providing more opportunities to abuse support features.

2.2. On-demand Hints and Immediate Feedback in CBLPs

Although abuse of on-demand hints and immediate feedback are the key characteristics of gaming the system, these are also fundamental features of many CBLPs. Substantial research is dedicated to immediate hints and feedback. Students tend to benefit from timely feedback and support during learning activities (Butler and Woodward, 2018; Shute, 2008; Lu et al., 2023; Razzaq et al., 2007; Lu et al., 2021). Multiple studies present evidence that providing feedback immediately after students respond to or complete problem sets has a positive effect on students' procedural and conceptual knowledge (Corbett and Anderson, 1994; Dihoff et al., 2003; Phye and Andre, 1989). In many CBLPs, students may also request assistance, which includes hints and feedback as students work on problems. These features have varying implementations. Common hints and feedback methods include presenting general topical information (Aleven and Koedinger, 2000), worked examples that present a complete solution to a similar problem (McLaren et al., 2016), providing the complete solution to the given problem (Williams et al., 2016), being shown similar examples done incorrectly (McLaren et al., 2016; Adams et al., 2014), providing targeted feedback based on a student's common wrong answer (Gurung et al., 2023; Gurung et al., 2023), and being given a series of step-by-step hints (Feng and Heffernan, 2006).

Overall, these features have differing levels of efficacy. For example, tutoring strategies, which provide problem-specific hints, are effective at improving student performance (Prihar et al., 2021; Patikorn and Heffernan, 2020). Williams et al. (2016) found students performed better when presented with Machine Learning-generated explanations as opposed to only receiving the answer. Research has also shown that feedback is more effective if provided as the student is answering questions than if it is delayed until after an activity (Lu et al., 2023), which supports the argument that assistance should be on-demand and immediate. Yet, there are also implementations of this assistance that show mixed or no efficacy. For example, McLaren et al. (2016) found that worked examples did not improve learning outcomes. Aleven and Koedinger (2000) found that students typically ignored assistance that involved general information aimed at helping students learn concepts; these students preferred hints that focused on the problem on which they were working. Another study found that students required support to utilize on-demand learning assistance but failed to find learning gains even when this support was given (Aleven et al., 2016). Overall, these findings suggest that students want assistance and feedback focused on the problem in which they are struggling; they are less engaged with assistance that provides general information about the concepts they are learning.

In theory, on-demand assistance tutoring strategies in intelligent tutoring systems should encourage help-seeking behaviors, which can increase students' autonomy and control of their learning (Aleven et al., 2006; Aleven et al., 2016). Yet, students can abuse this control, as exhibited by gaming behaviors. Although there is ample research on the efficacy of hints and feedback, more is needed to understand who benefits from these resources. Furthermore, the interactions between access to these learning supports, how students use these supports, and learning outcomes have not been fully explored.

2.3. Gamification & Engagement

Gamification, adopting features of games into non-game contexts, is a common method of increasing engagement in CBLPs (Landers, 2014; Karagiorgas and Niemann, 2017). Gamification includes incorporating "design elements, game thinking, and game mechanisms" (Karagiorgas and Niemann, 2017). In contrast, game-based learning allows students to engage in learning through a playful activity (Karagiorgas and Niemann, 2017).

According to the *Theory of Gamified Learning*, gamification impacts students' learning outcomes by influencing their behaviors as they learn (Landers, 2014). For example, games can influence students' learning behaviors by turning activities that may be exasperating into ones that are "pleasantly frustrating" (Gee, 2005). This pleasant frustration is often a product of the balance of success and failures necessary for an engaging game (Juul, 2009). It is plausible that a frustrating problem that might cause an uptick in gaming behaviors in a traditional CBLP could produce engagement in a gamified one. As frustration is a likely antecedent of gaming behaviors, gamification could theoretically reduce instances of gaming the system by reframing typically frustrating experiences as enjoyable. This idea is supported by evidence that students respond positively to game-based failure in a gamified learning environment by exhibiting productive performance (Vanacore et al., 2024). Thus, students who might disengage with a traditional CBLP by gaming the system would possibly instead engage with a gamified version of the CBLP; however, the extent to which gamification can produce these "pleasantly frustrating" experiences associated with games is unclear, as engaging with a gamified CBLP is not the same as playing a game. Furthermore, which features of a learning engagement are gamified may be as important as whether a learning experience is gamified at all.

Research on gamification has shown mixed effects on students' behaviors in learning contexts. Some have found a positive effect of gamified features on student engagement (Garris et al., 2002; Gaston and Cooper, 2017; Vanacore et al., 2023a), while others have not (Malkiewich et al., 2016; McKernan et al., 2015). Garris et al. (2002) found that reward-based systems caused students to take more time and replay problems more frequently. Previous work has also found that gamified performance-based feedback increases students' persistence behaviors associated with greater impacts on student learning (Vanacore et al., 2023a; Vanacore et al., 2023b). Alternatively, Malkiewich et al. (2016) found that some gamified features, such as narrative elements, did not positively impact students' persistence within the game. McKernan et al. (2015) also found that reward systems improved students' perceptions of their experiences, but not their behaviors or outcomes. As gaming the system can be seen as the antithetical behavior of persistence and engagement, it is plausible that gamification could mitigate these behaviors; however, the mixed results suggest that more research is necessary to evaluate this hypothesis.

2.4. Mitigating Gaming the System Behaviors

Interventions for gaming behaviors can be either proactive or reactive. Proactive interventions involve attempting to curtail the gaming behaviors before students have the opportunity to exhibit said behaviors (Aleven et al., 2006; Murray and VanLehn, 2005; Walonoski and Heffernan, 2006; Xia et al., 2020). Reactive options require targeted interventions after the gaming behavior has been identified (Baker et al., 2006).

Research suggests that proactively discouraging students from superfluous hint requests may be an imprecise method of addressing gaming behaviors. Advising students to request hints only when they truly needed them while delaying hint availability by ten seconds reduced hint usage and did not impact overall performance (Murray and VanLehn, 2005). However, Murray and VanLehn (2005) found some evidence that this method may have improved the performance of students who might have otherwise gamed the system; they found a marginally insignificant positive effect on the performance of students with low prior knowledge. The lack of statistical significance may be attributed to the study's small sample. Thus, the finding suggests that manipulating access to hints could help students who game the system.

Another proactive method includes presenting students with information about their gaming-related behaviors. Researchers have theorized that this intervention may influence students in one of two ways. They make the student aware that the system is logging their behaviors, thus creating "panopticon-like paranoia" of constant awareness of potential observation (Walonoski and Heffernan, 2006). Alternatively, visualizing students' actions may also nudge students to reflect on their learning behaviors (Xia et al., 2020). Two studies found that providing graphical representations of learning-related behaviors can reduce instances of gaming the system (Walonoski and Heffernan, 2006; Xia et al., 2020). However, neither of these studies evaluated this method's impact on learning. Furthermore, students who are likely to game the system to progress through learning activities may also game the system to manipulate the outputs of the visualizations.

As a reactive option, Baker et al. (2006) placed students who had previously gamed on specific content into an intervention focused on the content that the students had gamed. In this intervention, students saw an animated dog that displayed emotional responses to their learning-related behaviors. The animation embodied an exaggerated mimic of a teacher's emotions in response to student behavior (e.g., excitement and positivity if a student exerted effort, or frustration if a student gamed the system). This intervention significantly decreased students' gaming behaviors and positively affected their learning outcomes, as measured by a post-test, compared to a control group that did not have access to the animated dog.

Overall, there is suggestive evidence that directly dissuading students from gaming behaviors like hint abuse may benefit some students, yet it is unclear whether it has adverse effects on the non-gaming student population. Alternatively, presenting visualizations that represent students' actions may reduce instances of gaming; however, the learning impact of this method has not been explored. Notably, only the study that imposed a direct content-specific intervention showed a significant impact on learning (Baker et al., 2006). Thus, none of these studies connected the prevention of gaming behaviors with learning outcomes.

One alternative method is to change the learning environment substantially. Richey et al. (2021) evaluated the impact of a gamified CBLP by comparing it to an equivalent non-gamified CBLP and found that the gamified version reduced students' gaming behaviors, thus improving their learning outcomes. The gamified CBLP included interactive characters that prompted students with feedback and provided "narrative context for why they are performing various problem-solving activities." The reduction in gaming behavior completely mediated the effects on the posttests, suggesting the gamified elements' primary mechanism for impacting outcomes is reducing gaming behaviors.

3. Current Study

In the current study, we seek to understand the interplay between gaming the system behaviors, CBLP instructional design, and student learning by evaluating the following research question: *Do students who tend to game the system in a traditional CBLP with on-demand hints and immediate feedback (i.e., 'gamers') benefit from alternative CBLPs?*

We address this question using open-source data[3] from an efficacy study conducted to evaluate the effects of different CBLPs on middle school students' (U.S. Grade Seven) algebraic knowledge (Decker-Woodrow et al., 2023). The original study included four conditions (discussed in detail in Section 4.2) consisting of two traditional CBLP conditions administered through ASSISTments and two gamified CBLP conditions (From Here to There! and DragonBox Algebra 12+). In one of the ASSISTments conditions, students worked through traditional problem sets with access to hints and automated immediate feedback (Immediate Condition). In the other ASSISTments condition, access to both hints and feedback was delayed until after students completed problem sets (Delayed Condition). This question was preregistered[4] along with the hypotheses that gamers would benefit from both the Delayed Condition condition and gamified conditions relative to the Immediate Condition.

The original efficacy study found that both gamified conditions caused higher performance in algebraic knowledge compared to the delayed condition, but the effect of the Immediate Condition was not significant after controlling for student-level covariates (Decker-Woodrow et al., 2023). The current study is an extension of a previous analysis presented at the *14th International Learning Analytics and Knowledge Conference* (Vanacore et al., 2024). In that analysis, we compared the Immediate and Delayed conditions to understand the impact of feedback delivery on learning for students with a higher propensity to game the system in the Immediate

[3] The data can be accessed on the Open Science Foundation (OSF) repository after filling out a data exchange agreement: https://osf.io/r3nf2/

[4] https://osf.io/kfq2s

Condition. We reported that although students with a low propensity to game the system benefited from the Immediate Condition, those with a high propensity to game the system may have benefited from the Delayed Condition.

Here, we extend this analysis by including gamified conditions to fully explore how differences in CBLP institutional design may produce impact differentials for gamers. These gamified conditions share many key features: dynamic manipulation of equations, performance-based feedback, multiple paths to solutions, and the ability to replay problems. But they also have key differences (further described in Section 4.2), including narrative goals and the way puzzles are used to teach mathematical concepts. The following study does not allow us to understand the effects of individual gamified features, yet we can better understand how different students interact with gamified systems generally and whether these systems have consistent or inconsistent patterns of treatment effect heterogeneity.

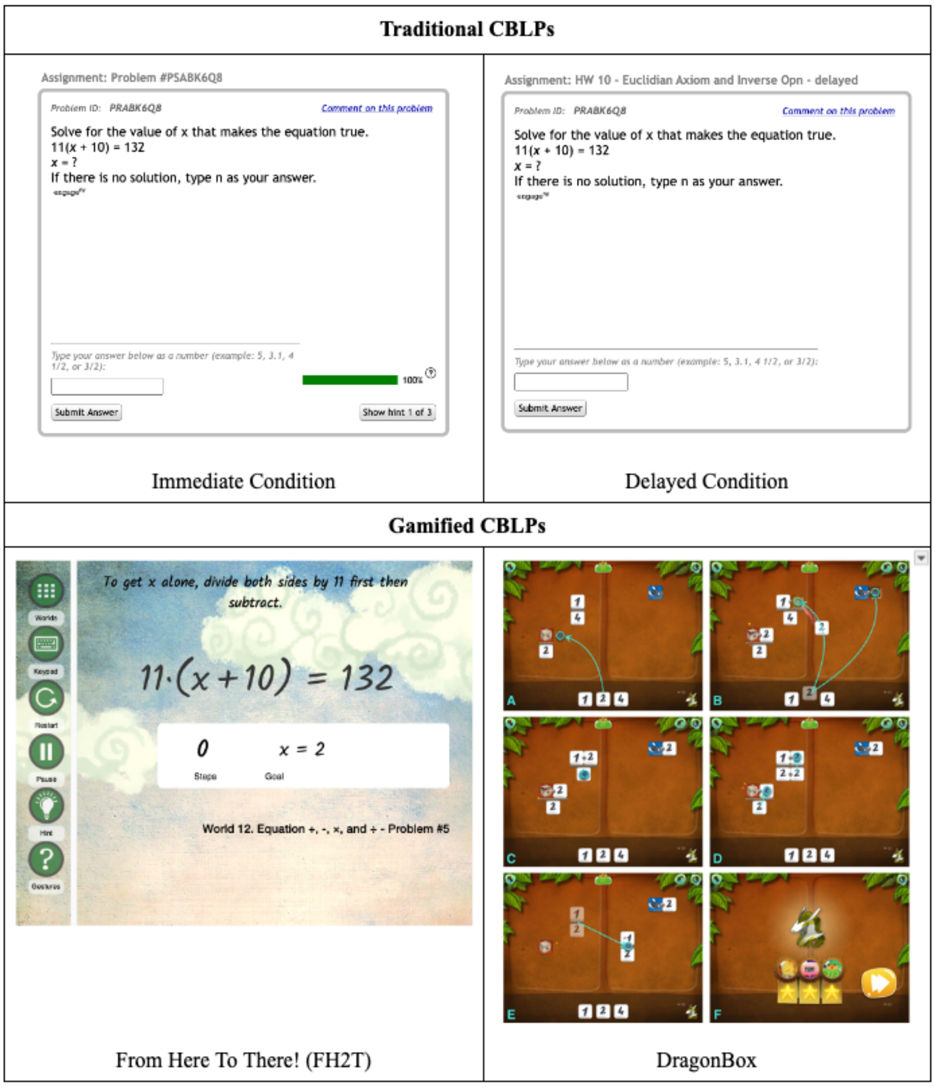
4. Method

4.1. Design

This study consists of a randomized controlled trial conducted in a school district in the Southeastern United States. The study consisted of 52 teachers in 10 schools. Students were ranked within classrooms based on their prior state mathematics assessment scores, blocked into sets of five (i.e., quintets), and then randomly assigned into either the From Here to There (FH2T; 40%), DragonBox (20%), Immediate Feedback (20%), or Delayed Feedback (20%) conditions. The disproportionate weighting was intended to allow researchers to focus their work on evaluating and understanding FH2T. The implementation of the study included nine weekly half-hour sessions in which the students worked in their assigned program. A full description of the study design can be found in Decker-Woodrow et al. (2023), an analysis of the study implementation is reported in Vanacore et al. (2024), and a description of the data set can be found in Ottmar et al. (2023).

4.2. Conditions

As explained in Section 3, the study consists of four conditions: two traditional and two gamified CBLPs. Example problems from each condition are presented in Figure 1. All conditions focused on teaching procedural ability, conceptional knowledge, and flexibility in algebraic equation equivalency.



Traditional CBLPs

Assignment: Problem #PSABK6Q8

Problem ID: PRABK6Q8

Solve for the value of x that makes the equation true.
11(*x* + 10) = 132
x = ?
If there is no solution, type n as your answer.

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

Immediate Condition

Assignment: HW 10 - Euclidian Axiom and Inverse Opn - delayed

Problem ID: PRABK6Q8

Solve for the value of x that makes the equation true.
11(*x* + 10) = 132
x = ?
If there is no solution, type n as your answer.

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

Delayed Condition

Gamified CBLPs

To get x alone, divide both sides by 11 first then subtract.

$11\cdot(x+10) = 132$

0 Steps

$x = 2$ Goal

World 12. Equation +, -, ×, and ÷ - Problem #5

From Here To There! (FH2T)

DragonBox

Figure 1: Example problems from each condition.

Both traditional problem sets' conditions (Immediate and Delayed Conditions) were administered through ASSISTments (Heffernan and Heffernan, 2014), an online homework system that assists students as they solve traditional problem sets. The problem sets in ASSISTments are adapted from open-source curricula, thus resembling problems students encounter in their textbooks and homework assignments. ASSISTments presents students with problems one at a time on their screen. Each condition included 218 problems of the same problems selected from three curricula – *EngageNY*, *Utah Math*, and *Illustrative Math* – to address specific algebra skills. The problems were divided into nine problem sets and administered throughout nine half-hour sessions during school hours.

4.2.1. Immediate Hints and Feedback (Immediate Condition)

In the Immediate Condition, students could request hints while solving problems. They also received automatic feedback on whether their answer was correct or incorrect after every submitted answer. Each problem contained a series of hints with a similar structure. The first hint gave the students the first step to answering the problem. The second hint gave the student a worked example of a similar problem. The final hint provided the student with the steps to complete the problem as well as the problem's solution. Students could submit as many answers as needed but could not move on until they had entered the correct answer. This condition was shown to be effective when implemented as a substitute for paper homework assignments (Roschelle et al., 2016). However, this condition provides both of the features that enable gaming behaviors: a sequence of hints that eventually provide the answer, which students can use for *hit abuse*, and immediate feedback, which allows for *guess-and-check* approaches to submitting responses.

4.2.2. Post-Assignment Hints and Feedback (Delayed Condition)

The Delayed Condition provided post-assignment assistance rather than on-demand hints and immediate feedback. In this condition, problem sets were administered in "test mode," so students did not receive any feedback or hints while submitting answers in each problem set. They could only submit one answer and progressed through the problem set without any feedback on their performance. Students received a report with feedback on their accuracy at the end of each problem set, and they could review their responses, revisit problems, and request hints. This condition was used as an active control in the original study, but we re-conceptualize it here as an intervention for gamers.

4.2.3. From Here to There! (FH2T)

From Here to There! (FH2T) takes a nontraditional approach to algebra instruction by applying aspects of perceptual learning (Goldstone et al., 2010) and embodied cognition (Abrahamson et al., 2020). Instead of presenting students with traditional algebra equations and expressions that students solve or simplify, FH2T gives students access to a starting expression (start state) that they must dynamically transform into a mathematically equivalent expression (goal state). Students can manipulate the expression by dragging numbers and symbols from one position to another on the screen or using a keypad when expanding terms. Only mathematically valid manipulations are accepted. Each valid manipulation counts as a step, which is logged and used to evaluate how efficiently the student transforms the expression from the start to the goal state. FH2T has 252 problems that are presented sequentially by mathematical content and complexity. Students must complete one problem in the sequence before advancing to the next problem.

4.2.4. DragonBox Algebra 12+ (DragonBox)

DragonBox Algebra 12+ [5] (DragonBox) is an educational game that provides instruction in algebraic concepts to secondary school students (ages 12-17). For each problem, students are asked to isolate a box containing a dragon—equivalent to solving an equation for x. This design incorporates research-based pedagogical methods, including discovery-based learning, embedded gestures, diverse representations of concepts, immediate feedback, and adaptive difficulty (Cayton-Hodges et al., 2015; Torres et al., 2016). The game's key innovation is that students

[5] https://dragonbox.com/products/algebra-12

learn the rules of algebra without using or manipulating numbers or traditional algebraic symbols. Thus, students engage with the algebraic concepts as if they are puzzles. Numbers and traditional algebraic symbols are introduced gradually, presumably after the student has learned the underlying concepts. Furthermore, DragonBox applies a narrative goal to the learning: students must allow the dragon to come out of the box by isolating it in the equation. Previous analyses found that DragonBox positively affects engagement and attitudes toward math, but findings are mixed in regard to its efficacy in improving learning outcomes (Siew et al., 2016; Dolonen and Kluge, 2015).

4.3. Analysis Plan

Our goal is to assess whether the students who gamed the system in a traditional CBLP would have benefited from a different CBLP. This poses a methodological difficulty because estimating the effect of randomizing gamers to the Immediate condition requires us to contrast their posttest scores with scores from comparable students in the other conditions; however, we only observe gaming behavior in the Immediate condition. Even if we did observe gaming in another condition, we cannot know if a student who gamed the system in one condition would also game in another. In fact, we hypothesize that students who would game the system in the Immediate Condition would not engage in similar behaviors in the other conditions and benefit from the learning therein.

To address this methodological problem, we propose that students have a baseline propensity to game the system in the Immediate Condition that exists before randomization, regardless of whether it can manifest itself after treatment assignment. Because it is considered a baseline covariate, similar to students' pretest knowledge, it is independent of the random treatment assignment and can serve as a moderating variable. However, unlike pretest knowledge, only students in the Immediate Condition have the opportunity to display the behavior; therefore, its value in the other conditions is unknown. Nevertheless, once this latent propensity to game the system is estimated, we can evaluate whether it moderates the effects of the various interventions.

Our analysis requires two key steps, which are described in depth below. First, we use a gaming the system detector to identify instances where students are gaming the system within the Immediate Condition. Next, we use the causal method of Fully Latent Principal Stratification (FLPS, (Sales and Pane, 2019)), which will allow us to estimate the effect heterogeneity of each program based on students' latent propensity to game the system in the Immediate condition. Each of these methods is delineated in the sections below.

4.3.1. Implementation of Gaming The System Detector

This paper employs the rule-based gaming detectors originally created by Paquette et al. (2015) to identify gaming behavior among students working on algebraic problems. To the best of the authors' knowledge, the *Cognitive Model* developed by Paquette et al. (2015) was the first detector that was transferable across intelligent tutoring systems, from Cognitive Tutor Algebra (CTA) to ASSISTments. The original rule-based detector was engineered to detect gaming in CTA (Baker et al., 2008). Paquette et al. (2014) further developed the rule-based gaming detector by relying on human judgments through text replays of logged learner actions (Baker et al., 2010). The insights gained from human judgments were used to develop gaming detectors for CTA. Subsequently, the transferability of this rule-based gaming detector was validated in ASSISTments.

We elected to use the rule-based *Cognitive Model* for identifying gaming behavior as rule-based models mitigate key challenges to more complex detectors: model generalizability and detector rot. Unlike other system-specific gaming detectors, the rule-based model demonstrated cross-system generalizability, underscoring its adaptability and reliability in contexts beyond the original training data set (Paquette et al., 2015). More complex models are also susceptible to detector rot, a phenomenon that refers to the gradual decline in a model's performance over time. Prior studies have reported that more complex models, using more advanced ML and DL algorithms, are more prone to this phenomenon than their simpler counterparts (Lee et al., 2023; Levin et al., 2022). Given that many gaming the system models were developed years before our study, they are at a higher risk of experiencing detector rot. Consequently, we implemented the rule-based detector, in light of its potential for sustained generalizability and resistance to detector rot.

To implement the rule-based detector for identifying gaming behavior among students, logged student actions are first aggregated into twenty-second clips. These clips are then analyzed alongside additional information to contextualize the actions into more meaningful categories, as outlined in Table 1. This process allows for the identification of specific sequences of action patterns considered indicative of gaming behavior, with a total of 13 patterns detailed in Table 2. For example, analyzing the first pattern, "incorrect → [guess] & [same answer/diff. context] & incorrect" indicates a scenario where the student entered an incorrect answer and re-entered the same incorrect answer within 5 seconds of making the first attempt. For the students in the Immediate Condition in this study, the rule-based detector employs this structured approach, leveraging log-file data compiled in twenty-second intervals (as described in Section 4.4.1) to determine gaming instances during each clip. These determinations are then integrated into the FLPS model, enhancing its ability to accurately reflect student engagement and behavior.

4.3.2. Estimating Effects Using Fully Latent Principal Stratification

FLPS is a variant of Principal Stratification, a causal inference method used in randomized controlled trials for estimating the intervention effects on subgroups that emerge after the treatment has begun (Frangakis and Rubin, 2002; Page et al., 2015). Generally, to estimate subgroup effects, the subgroups must be defined prior to the intervention and be independent of treatment assignment. For example, to test whether an effect varies based on pretest knowledge, interacting the treatment with the pretest knowledge score estimates how the treatment effect differs across different subgroups of students with similar prior knowledge. However, subgroups identified based on students' interactions with a treatment program cannot be observed at baseline. Furthermore, they may only be observed for students randomized in some conditions (in the present case, the Immediate Condition). Even if they could be observed in multiple conditions, the student's behavior is confounded by their condition (i.e., some students may game the system in one condition but would not in another condition). FLPS provides a solution to this methodological problem by modeling these behaviors as manifestations of latent student characteristics, which are not observed for students randomized to some conditions but must be estimated.

Often, a student's interactions with programs are categorized by a complex series of behaviors. This fact is particularly true in CBLPs, where students may display an array of behaviors during the program, such as meeting implementation goals (Dieter et al., 2020; Vanacore et al.,

Table 1: List of contextually interpretable actions that are utilized to develop the rules for identifying gaming the system behavior.

Identifier	Description
[did not think before help request]	Pause smaller or equal to 5 seconds before a help request
[thought before help request]	Pause greater or equal to 6 seconds before a help request
[read help messages]	Pause greater or equal to 9 seconds per help message after a help request
[scanning help messages]	Pause between 4 and 8 seconds per help message after a help request
[searching for bottom-out hint]	Pause smaller or equal to 3 seconds per help message after a help request
[thought before attempt]	Pause greater or equal to 6 seconds before step attempt
[planned ahead]	Last action was a correct step attempt with a pause greater or equal to 11 seconds
[guess]	Pause smaller or equal to 5 seconds before step attempt
[unsuccessful but sincere attempt]	Pause greater than or equal to 6 seconds before a bug
[guessing with values from problem]	Pause smaller than or equal to 5 seconds before a bug
[read error message]	Pause greater than or equal to 9 seconds after a bug
[did not read error message]	Pause smaller than or equal to 8 seconds after a bug
[thought about error]	Pause greater than or equal to 6 seconds after an incorrect step attempt
[same answer/diff. context]	Answer was the same as the previous action, but in a different context
[similar answer]	Answer was similar to the previous action (Levenshtein distance of 1 or 2)
[switched context before right]	Context of the current action is not the same as the context for the previous (incorrect) action
[same context]	Context of the current action is the same as the previous action
[repeated step]	Answer and context are the same as the previous action
[diff. answer AND/OR diff. context]	Answer or context is not the same as the previous action

2024), productive persistence (Vanacore et al., 2023a), mastering knowledge components in mastery learning (Sales and Pane, 2019), and gaming-the-system behaviors that may be viewed as indicators of latent student characteristics (i.e., high fidelity users, persistent learners, mastery users, gamers). Understanding which latent behavioral characteristics are associated with effect heterogeneity can help provide a better understanding of how CBLPs impact students.

However, these latent behavioral characteristics are likely context-dependent (e.g., a student may persist in one CBLP and not another). These behaviors are unobserved for the students randomized to conditions where they do not have the same opportunity to display the relevant behaviors as they did not interact with the same CBLP. For example, some students might request hints in a condition that provides them. However, other conditions may not provide the opportunity to request hints, so we cannot observe whether students in this condition would have requested them. Furthermore, when the relevant behaviors are observable, they are confounded by the students' assigned conditions, which presumably influence their behaviors. For example, suppose students are randomized between receiving optional hints or explanations. In that case, students who prefer hints may be unlikely to request help when they are randomized to explanations and vice versa. Therefore, their underlying latent student characteristics that define the subgroups in the FLPS can be interpreted as explaining students' behaviors if randomized to a specific condition.

Sales and Pane (2019) used this method to determine whether the effect of Cognitive Tutor Algebra I on students varies based on whether students were likely to master knowledge components in Cognitive Tutor by estimating the likelihood of mastering knowledge components for the treatment group as a latent variable. In the current study, we evaluate whether the effect of each alternative condition (Delayed Condition, FH2T, DrangonBox) differs based on students' latent propensity to game the system had they been assigned to Immediate Condition. This propensity to game the system in the Immediate Condition is independent of students' treatment

Table 2: List of contextually interpretable actions that are utilized to develop the rules for identifying gaming the system behavior.

Action patterns considered gaming behavior.
incorrect → [guess] & [same answer/diff. context] & incorrect
incorrect → [similar answer] [same context] & incorrect → [similar answer] & [same context] & attempt
incorrect → [similar answer] & incorrect → [same answer/diff. context] & attempt
[guess] & incorrect → [guess] & [diff. answer AND/OR diff. context] & incorrect → [guess] & [diff. answer AND/OR diff. context & attempt
incorrect → [similar answer] & incorrect → [guess] & attempt
help & [searching for bottom-out hint] → incorrect → [similar answer] & incorrect
incorrect → [same answer/diff. context] & incorrect → [switched context before correct] & attempt/help
bug → [same answer/diff. context] & correct → bug
incorrect → [similar answer] & incorrect → [switched context before correct] & incorrect
incorrect → [switched context before correct] & incorrect → [similar answer] & incorrect
incorrect → [similar answer] & incorrect → [did not think before help] & help → incorrect (with first or second answer similar to the last one)
help → incorrect → incorrect → incorrect (with at least one similar answer between steps)
incorrect → incorrect → incorrect → [did not think before help request] & help (at least one similar answer between steps)

assignment, as it exists prior to any treatment and regardless of whether it has the opportunity to manifest.

Since students' gaming behavior was detected in the Immediate Condition, our goal is to know how students with a high propensity to game the system in the immediate condition would fare if placed in each of the other conditions: the Delayed Condition, FH2T, or DragonBox. For simplicity, we refer to the Immediate Condition as the control and each of the other conditions as treatments. We present the following model as if there is only one treatment for simplicity, but the model estimated in this paper includes multiple treatment conditions, and therefore, multiple main and moderating effects.

Let τ_i be subject i's individual treatment effect: the difference between i's posttest score if i were randomized to treatment and their score if randomized to control. Let $\mathcal{T}$ and $\mathcal{C}$ be the samples of students randomized to treatment and control, respectively. Let α_{ci} be i's propensity to game the system if randomized to the control condition. α_{ci} is defined for $i \in \mathcal{T}$, as students in the treatment conditions still had a potential to game the system had they been randomized to control, even if this potential was never realized. Therefore, α_{ci} is estimated from gaming the system behavior for $\mathcal{C}$ and α_{ci} is imputed for each $\mathcal{T}$. This propensity to game the system is condition-specific and represents an estimate of whether students would have gamed the system if randomized to the control condition (i.e., Immediate Condition).

The principal effect is the treatment effect for the subgroup of students with a particular value for α_c:

$$\tau(\alpha) = E[\tau|a_t = \alpha] \tag{1}$$

To estimate the function $\tau(\alpha)$, we (1) estimate α_c for $\mathcal{C}$ as a function of pre-treatment covariates observed in both groups, (2) use that model to impute α_C for each $\mathcal{T}$, (3) estimate $\tau(\alpha)$ by including a treatment interaction in a linear regression model. The models are estimated using iterations through these steps in a Bayesian principal stratification model with a continuous variable consisting of measurement and outcome submodels, as outlined by (Jin and Rubin, 2008) and (Page, 2012).

Measurement Submodel: First, we estimate α_c by running a multilevel logistic submodel predicting whether the gaming detector identified the students in the treatment condition to have gamed the system on each twenty-second time clip as delineated in the equation 2. Let G_{cji} be a binary indicator of whether student i gamed the system during time-clip c when working on problem j. Let P_{ki} be covariate predictor k of K student-level predictors, which are measured at baseline for both $\mathcal{T}$ and $\mathcal{C}$ (described in section 4.4.2). Let the random intercepts be μ_j for problems, μ_i for students, μ_t for teachers, and μ_s schools, each modeled as independent and with normal distributions with means of zero and standard deviations estimated from the data.

$$logit(G_{cji}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k \mathrm{P}_{ki} + \mu_j + \mu_i + \mu_t + \mu_s \tag{2}$$

Using the parameters from Equation 2, students' propensity to game the system in the immediate condition is defined as:

$$\alpha_i = \sum_{k=1}^{K} \gamma_k \mathrm{P}_{ki} + \mu_i + \mu_t + \mu_s \tag{3}$$

We impute α_i for $\mathcal{T}$ with random draws from a normal distribution with mean $\sum_k \gamma_k \mathrm{P}_{ki} + \mu_{t[i]} + \mu_{s[i]}$, where $\mu_{t[i]}$ and $\mu_{s[i]}$ are the random intercepts for student i's teacher and school, respectively, and standard deviation equal to the estimated standard deviation of μ_i. Note that randomization occurred at the student level (i.e., teachers had students in the treatment and control conditions in their classes). Therefore, we include the random intercepts for schools and teachers from submodel 2 in valuation for α_{ci} for students in the control. However, μ_i is unknown for $\mathcal{T}$, but we can assume its distribution is the same in the two conditions because of the randomization.

Prior to the simultaneous estimation of the FLPS sub-models, we estimated the measurement submodel separately using the *stan_glmer* function from the *rstanarm* package (Goodrich et al., 2020) with different combinations and transformations of the pretest predictors and demographic variables to find a suitable model for the analysis. To build the measurement model, we randomly split the data into training (80%) and testing (20%) data sets.

Outcomes Submodel: To estimate the treatment effect for students with differing propensities to game the system, we run a multilevel linear regression predicting student's post-test algebraic knowledge (Y_i). The submodel includes interaction between α_{ci}—estimated for $\mathcal{C}$ and randomly imputed for $\mathcal{T}$—and each Z_i indicator of being in the treatment condition. In practice there is one Z_i indicator for each treatment condition. Let the random effects for teacher be ν_t and for school be ν_s.

$$Y_i = \beta_0 + \beta_1 \mathrm{Z}_i + \beta_2 \alpha_{ci} + \beta_3 \alpha_{ci} \mathrm{Z}_i + \sum_{k=1}^{K} \lambda_k \mathrm{P}_{ki} + \nu_t + \nu_s + \epsilon_i \tag{4}$$

Using the parameters from the submodel 4, the treatment effect for students with a particular propensity to game the system is modeled as

$$\tau(\alpha) = \beta_1 + \beta_3 \alpha \tag{5}$$

Once again, there is one $\tau(\alpha)$ for each treatment condition, as β_1 and β_3 are estimated for each treatment condition. Submodels 2 and 4 together formed a FLPS model, which we fit using the Stan Markov Chain Monte Carlo software through STAN (Arezooji, 2020). We ran 10,000 iterations of FLPS models using Markov chain Monte Carlo chains calling *Stan* through *rstan* (Carpenter et al., 2017) in *R*; code is posted on GitHub[6]. We evaluated convergence using trace plots and checking whether the $\hat{R}$, which measures convergence by comparing between and within chain estimates of each parameter, was below the recommended threshold of 1.05 (Vehtari et al., 2021). The maximum $\hat{R}$ for the model parameters was 1.04.

4.4. Data & Variables

The data for this study exists at two different levels: student and time clip. The student-level variables include the student pretest data, demographics, roster, and learning outcomes. Alternatively, the data used in gaming the system detector and the detector's output is aggregated in twenty-second clips of the students' usage of the program. Only time-clip data from the Immediate Condition is used, as we are estimating students' propensity to game the system in that condition.

Notably, the study was conducted during the COVID-19 pandemic and involved considerable attrition. An attrition analysis based on United States Institute of Educational Sciences standards found that attrition did not bias the effect estimates of the conditions (Decker-Woodrow et al., 2023). The sample for the current study consists of 1976 students: 402 in the Immediate Condition, 385 in the Delayed Condition, 372 in the DragonBox, and 817 in FH2T. The students were taught by 36 teachers in 10 schools.

4.4.1. Gaming the System Detector Inputs and Output

Gaming labels were generated by adopting the methodology described by Paquette et al. (2015) to produce action clips for the gaming detector as described in Section 4.3.1. These clips capture actions taken by students in ASSISTments. Each clip contains unique identifiers: the student, the problem being worked on at the clip's start, the skill associated with that problem, and the problem type. If the clip spanned multiple problems, the problem the student ended on was also documented. Additionally, the clips detail the start and end times of actions, their total duration, and, for attempts, the action's correctness and the student's answer. Supplementary data within the clips include indicators for hint requests, the number of hints requested during the clip, the total hints available for the problem, the use of a 'bottom-out' hint, and the total attempts by the student. Gaming labels were generated based on instances where the student's actions in the clip met one or more rules indicative of gaming system behavior.

4.4.2. Pretreatment Predictors

To estimate students' propensity to game the system, we used data from assessments administered prior to their use of their assigned condition. Pretest scores were collected by the original studies' researchers: prior algebraic knowledge, math anxiety, and perceptual processing skills. Pretest algebraic knowledge was a variant of the learning outcome described below. The math anxiety assessment was adapted from the *Math Anxiety Scale for Young Children-Revised* (Chiu

[6] https://github.com/kirkvanacore/FLPS_GamingTheSystem/tree/main/code/JEDM_code

and Henry, 1990), which assessed negative reactions towards math, numeric inconfidence, and math-related worrying (Cronbach's α=.87; see the items on OSF[7]). Five items were adapted from the Academic Efficacy Subscale of the *Patterns of Adaptive Learning Scale* to assess math self-efficacy (Midgley et al. (2000); Cronbach's α = .82; see items on OSF[8]). The perceptual processing assessment evaluates students' ability to detect mathematically equivalent and nonequivalent expressions as quickly as possible (see item on OSF[9]). The district also provided metadata on the students, including their demographics and most recent standardized state test scores in math. Demographic data included race/ethnicity, individualized education plan status (IEP), and English as a second or foreign language (ESOL) status. Race/ethnicity was dummy-coded, with white students as the reference category because they were the majority population.

We standardized (z-score) all continuous scores to improve model fit and ease interpretation. Log forms of assessment test times were also included in the models. We include polynomials of the pretest scores when they improved model fit. Missing data were imputed using single-imputation with the Random Forest routine implemented by the missForest package in R (Stekhoven and Buehlmann, 2012; R Core Team, 2016). The number of missing data for each student was included as a predictor in each submodel.

4.4.3. Learning Outcome

The learning outcome for the study is students' algebraic knowledge, which was assessed using ten multiple-choice items from a previously validated measure of algebra understanding (Star et al. (2015); Cronbach's α = .89; see the items on OSF[10]). Four of the items evaluated conceptual understanding of algebraic equation-solving (e.g., the meaning of an equal sign), three evaluated procedural skills of equation-solving (e.g., solving for a variable), and three evaluated flexibility of equation-solving strategies (e.g., evaluating different equation-solving strategies). Together, these ten items assessed students' knowledge in algebraic equation-solving, the improvement of which was the goal of the interventions. The learning outcome was assessed before and after the interventions. Students' scores were standardized (z-score) to ease model fit and interpretation.

5. Results

5.1. Gaming Detector Outputs

The students in the Immediate Condition produced 89,960 twenty-second clips of data. The gaming detector indicated that students gamed the system during 4.62% of the total clips. Most students (94.18%) were detected to have gamed the system at least once. Figure 2 displays a density and boxplot of the percentage of clips in which each student gamed the system during the study. Overall, students were detected to have gamed the system during approximately 5% of clips (mean = 5.10%, median = 4.71%, SD =3.64%,). However, this distribution has a positive skew. Almost a tenth (9.00%) of students were detected to have gamed during at least 10% of the clips. This suggests that while most students engage in some gaming, there is wide variation

[7] https://osf.io/rq9d8
[8] https://osf.io/rq9d8
[9] https://osf.io/r47ev
[10] https://osf.io/uenvg

in student gaming behavior, and some students likely have a substantially higher propensity to game than others.

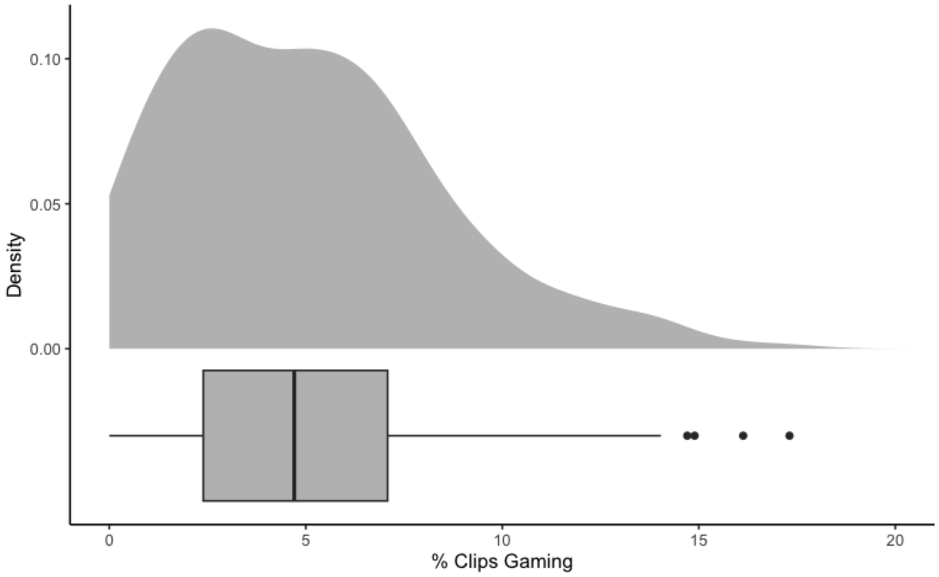


Figure 2: Density and box plots showing the distribution of students' percentage of clips in which the detector indicated gaming behavior.

In almost three-quarters of the problems (74.35%), the detector did not detect any gaming. Notably, there was a small significant negative correlation between the gaming rate on each problem and the average accuracy on the problem (r = -0.21, $p < 0.001$). Furthermore, some problem types had significantly higher rates of gaming behaviors than others (F[5,244] = 6.403 $p < 0.001$). Students were more likely to game the system on 'check all that apply' problems compared with problems for which they had to submit a number ($p < 0.001$), variable (p = 0.016), algebraic expression (p = 0.020), or open response ($p < 0.016$). Furthermore, students were more likely to game the system on a multiple choice question than problems in which they had to submit a numeric answer (p = 0.015). Because the problems without gaming behavior would not be informative for assessing students gaming the system, we did not include these problems when estimating the measurement submodel.

5.2. Fully Latent Principal Stratification

Table 3 provides parameter estimates and relevant statistics for the measurement and outcome submodels.

5.2.1. Measurement Submodel

As described in section 4.3.2, before running the full FLPS model, we tested covariate combinations and transformations and validated the model using training and testing data sets. The model performed best when pretest sub-scores from the tests were included. Math anxiety had a non-linear relation to gaming behavior, so a polynomial transformation was included in the model. Our final model produced an adequate AUC of 0.87 on the testing data set.

Table 3: Fully latent principal stratification model parameter estimates.

	Measurement Submodel				Outcomes Submodel			
Predictors	Estimate	SD	P(>0)	P(<0)	Estimate	SD	P(>0)	P(<0)
α					-0.51	0.13	0.01	0.99
$Z_{DelayedCondition}$					-0.10	0.05	0.02	0.98
$\alpha * Z_{DelayedCondition}$					0.10	0.07	0.90	0.10
Z_{FH2T}					0.01	0.04	0.63	0.37
$\alpha * Z_{FH2T}$					-0.08	0.06	0.09	0.91
$Z_{DragonBox}$					0.08	0.05	0.96	0.04
$\alpha * Z_{DragonBox}$					0.04	0.07	0.71	0.29
Algebraic Procedural Knowledge	-0.06	0.04	0.09	0.91	-0.01	0.03	0.44	0.56
Algebraic Conceptual Knowledge	-0.10	0.06	0.04	0.96	0.15	0.04	0.01	0.99
Algebraic Flexibility Knowledge	-0.05	0.04	0.14	0.86	0.01	0.03	0.58	0.42
Algebraic Knowledge Items Complete	-0.00	0.08	0.49	0.51	0.04	0.05	0.79	0.21
Algebraic Knowledge Time (Log)	-0.08	0.06	0.10	0.90	-0.04	0.04	0.15	0.85
Math Anxiety	-0.18	0.12	0.06	0.94	-0.04	0.08	0.28	0.72
Math Anxiety (Squared)	-0.05	0.03	0.05	0.95	-0.01	0.02	0.24	0.76
Math Negative Reaction	0.16	0.08	0.98	0.02	0.01	0.06	0.59	0.41
Math Numerical Confidence	0.11	0.07	0.93	0.07	0.04	0.05	0.77	0.23
Math Self Efficacy	-0.04	0.04	0.21	0.79	0.03	0.03	0.81	0.20
Perceptual Sensitivity Score Part 1	-0.06	0.04	0.08	0.92	-0.00	0.03	0.48	0.52
Perceptual Sensitivity Time Part 1 (Log)	-0.10	0.06	0.04	0.96	-0.01	0.04	0.42	0.58
Perceptual Sensitivity Score Part 2	0.03	0.04	0.71	0.29	0.08	0.03	0.01	0.99
Perceptual Sensitivity Time Part 2 (Log)	-0.05	0.05	0.16	0.84	-0.03	0.03	0.13	0.87
Perceptual Sensitivity Score Part 3	0.01	0.05	0.60	0.40	0.00	0.04	0.52	0.48
Perceptual Sensitivity Time Part 4 (Log)	-0.06	0.06	0.13	0.87	0.04	0.04	0.86	0.14
State Test Score	-0.32	0.05	0.01	0.99	0.01	0.06	0.57	0.43
Female	-0.03	0.04	0.21	0.79	0.04	0.03	0.92	0.08
Hispanic	0.09	0.13	0.75	0.25	0.16	0.09	0.98	0.02
Asian/Pacific Islander	-0.17	0.12	0.08	0.92	0.18	0.08	0.98	0.02
Black	0.31	0.20	0.93	0.07	0.16	0.13	0.88	0.12
IEP	-0.02	0.04	0.26	0.74	0.02	0.02	0.76	0.24
EIP	0.01	0.04	0.62	0.38	0.01	0.02	0.68	0.33
ESOL	-0.00	0.05	0.47	0.53	-0.03	0.03	0.19	0.81
Gifted	-0.01	0.04	0.43	0.57	0.08	0.03	0.99	0.01
In-person Instruction	0.04	0.13	0.60	0.40	-0.10	0.07	0.06	0.94
Missing Data	0.03	0.09	0.63	0.37	0.02	0.05	0.63	0.37

The measurement submodel's coefficients provide insight into which student characteristics predict gaming behavior in the Immediate Condition. Generally, lower-performing students tended to game the system more frequently. Students with lower scores on the math section of their state test were more likely to game the system (γ_{17} = -0.32, $P(<0)$ = .99). This association was consistent with the albeit smaller associations between the algebraic knowledge sub-scores and the gaming behavior. Furthermore, there was a nonlinear association between math anxiety and gaming behavior. Students with higher math anxiety were also less likely to game the system (γ_6 = -0.18, $P(<0)$ = .94) than those with low math anxiety. This association was even greater for students with highest levels of math anxiety, as shown by the coefficient for math anxiety squared (γ_7 = -0.05, $P(<0)$ = .95). Students with a higher negative reaction toward math (γ_8 = 0.16, $P(>0)$ = .98) and with higher numeric confidence (γ_9 = 0.11, $P(>0)$ = .93) were more likely to game the system. Time spent on the algebraic knowledge test and the perceptual sensitivity learning sub-tests were all predictive of gaming behavior; students who took less time on these tests were more likely to game the system.

5.2.2. Outcomes Submodel

The outcomes submodel provides the parameter estimates that address whether students likely to game the system in the Immediate Condition would benefit from an alternative learning environment. As expected, students' propensity to game the system (α) was negatively associated with the outcome (β_2 = -0.51, $P(<0)$ = 0.99). However, the interactions between α and the conditions varied widely.

First, the Delayed Condition had a negative main effect but a positive interaction. For students who had an average propensity to game the system in the Immediate Condition ($\alpha = 0$), the effect of the Delayed Condition was likely negative ($\beta_{1-Delayed}$ = -0.10, $P(<0)$ = 0.98). The interaction between students' propensity to game the system and the Delayed Condition was likely positive ($\beta_{3-Delayed}$ = 0.10, $P(>0)$ = 0.90). Although the positive effect of the interaction balanced out the negative effect of the Delayed Condition, it did not mitigate the lower performance associated with students' propensity to game the system. That is to say, the negative effect for the Delayed Condition was equivalent in magnitude to the positive effect of the other interaction; thus, a student with a high propensity to game the system ($\alpha = 1$) would have the same benefit from the Delayed Condition as a student with a low propensity to game the system ($\alpha = 0$) would from the Immediate Condition. However, the magnitude of the positive effect for students who game the system in the Delayed Condition does not outweigh the overall negative association of propensity to gaming the system on the student's post-test scores. Therefore, the Delayed Condition likely does not mitigate the gaming behavior. These were consistent with our findings in (Vanacore et al., 2024), thus reproducing those results with a slightly different model.

Second, the results showed varied effects of gamification and the interactions between gamified programs and students' propensity to game the system. The estimated effect of FH2T on students who have an average propensity to game the system in the Immediate Conidtion ($\alpha = 0$) was small, and we have low confidence that it is greater than zero (β_{1-FH2T} = 0.01, $P(>0)$ = 0.63). Yet, the interaction between FH2T and propensity to game the system was likely negative (β_{3-FH2T} = -0.08, $P(<0)$ = 0.91). Thus, contrary to our hypotheses, gamers would likely have been better off in the Immediate Feedback condition than in FH2T; however, these results were not consistent for DragonBox, which had a likely positive main effect ($\beta_{1-DragonBox}$ = 0.08, $P(>0)$ = 0.96), yet there is little evidence of an interaction with students' propensity to game the system in the Immediate Condition ($\beta_{3-DragonBox}$ = 0.04, $P(>0)$ = 0.71).

Figure 3 provides a visual representation of the interactions between α and Z by plotting the effect sizes (τ) for different levels of students' propensity to game the system problem in the Immediate Condition (α). Note that the mean of α is zero and since all the covariates were standardized with means of zero, where the lines cross α = 0 can be interpreted as the estimated effect for the average student with average gaming behavior (*i.e.,* the main effect).

We have little evidence that the slope for the Dragon Box is different than zero. Therefore, it is most likely that students benefited from Dragon Box regardless of their propensity to game the system. However, we have higher confidence that the slopes for the Delayed Condition and FH2T are not zero. The Delayed Condition had an estimated negative impact on students with low or average gaming propensities, but those with very high propensities to game the system in the Immediate Condition may have experienced a positive effect. Note that only 5%

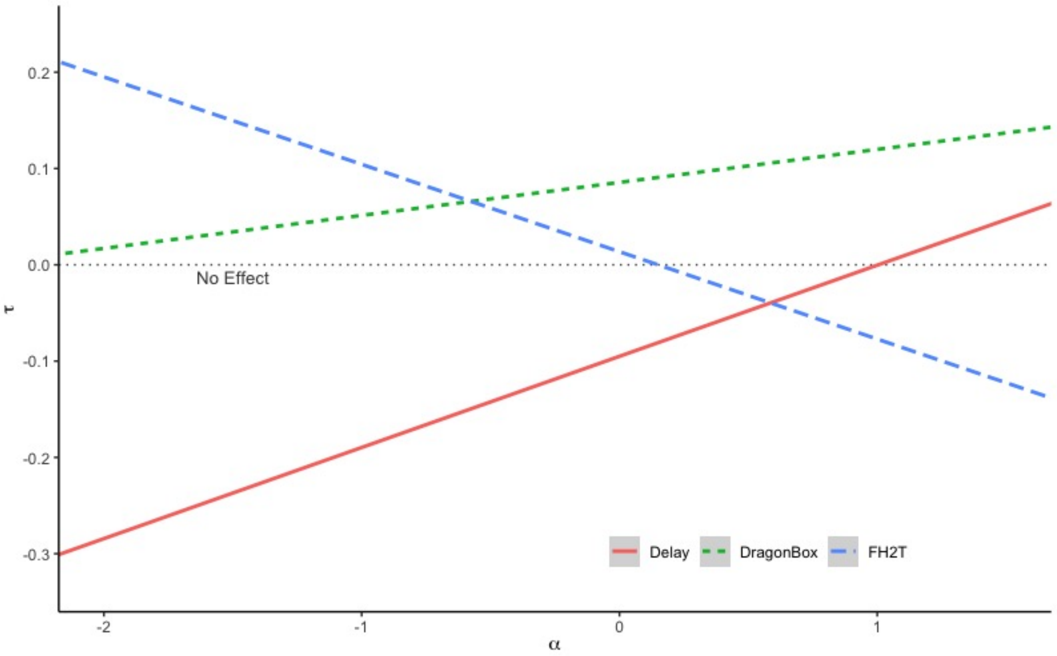


Figure 3: Plot of effect sizes (τ) by propensity to game the system in the Immediate Condition (α).

of the student population had propensities to game the system this high ($\alpha > 1$), so this effect is only relevant for the most ardent gamers. For FH2T, we see the opposite trend: students with the lowest propensity to game the system experience the greatest effects, and those with an average or high propensity to game the system likely experience no effect or negative effects, respectively.

6. Discussion

The results suggest that students who tend to game the system respond differently to various CBLPs, but not in the way we hypothesized. Although the two gamified conditions produced different heterogeneity patterns for students who tend to game the system, neither of these conditions fully mitigated the negative association between gaming the system behaviors and learning. Similarly, delaying access to hints and feedback likely has a positive effect on students who tend to game the system, yet not enough to fully outweigh the potential negative effects of gaming behavior. Although we found evidence that some learning environments may provide marginal benefits to students who tend to game the system, overall, more targeted interventions may be necessary to reengage these students in the learning process so that they fully benefit from the CBLPs' content.

6.1. Effects of Gamification on Students who Game the System

Although neither gamified condition had positive effect differentials on students with a high propensity to game the system, the interactions do provide some insight into how gamers tend to respond to different gamified environments. The effect was likely consistent for one of the gamified conditions (DragonBox), regardless of students' propensity to game the system, whereas gamers in the other gamified condition (FH2T) experienced negative effects. This suggests that gamification alone is not the only factor influencing students' behaviors and outcomes. Previous research on the relationship between gaming the system and gamification found that a gamified version of a CBLP did mitigate the gaming behavior and positively impacted student outcomes (Richey et al., 2021). In our study, the differences in heterogeneity patterns across platforms that shared many features of gamification (e.g., performance-based feedback, multiple paths to solutions, the ability to replay problems, and dynamic manipulation of equations) highlight that even slight variations in how a program is gamified may be more important than whether a program is gamified at all in terms of impacting students' behavior and learning.

It is possible that FH2T did not benefit gamers because the goals are more abstract than traditional problems. Students needed to find the most efficient path to the solution by manipulating the expressions and equations on the screen as opposed to submitting answers to traditional problems. Furthermore, the evaluation system in the game is not delineated prior to playing and is meant to be intuited by the students as they play. As previous research suggests that abstractness and ambiguity may induce gaming behaviors (Baker et al., 2009), one possible explanation is that students who are likely to game the system in the Immediate Condition become even more frustrated by FH2T, causing them to disengage with the activity to a greater extent than they would have in a traditional CBLP environment. This may explain in part why FH2T had the greatest effects on students with the highest pretest scores (Decker-Woodrow et al., 2023), who were the students who were least likely to game the system.

However, DragonBox had a likely consistent positive effect regardless of students' propensity to game the system, despite similar gamification features and a similar dynamic approach in teaching students the concept of equivalency to FH2T. Yet some key differences between these CBLPs may point to reasons for DragonBox's effectiveness among gamers. First, DragonBox allows students to engage with algebraic equivalency without interacting with traditional math symbols. Second, the activity's mathematical goal is embedded in a gamified narrative: freeing the dragon from the box. Finally, students learn the game's goals through animated interactions before learning the algebraic principles. These elements may have alleviated tendencies towards unproductive learning behaviors enough so that students benefited regardless of propensity to game the system. More research is necessary to determine exactly why there were stark differences in the effect trajectories between the two gamified conditions.

6.2. Effect of Assistance on Students who Game the System

The positive interaction between students' propensity to game the system and the Delayed Condition indicates that the impact of on-demand hints and feedback on student performance in a CBLP are heterogeneous. The differences in effects are likely attributed to how students use assistance features. Those who exploit hints excessively or rely on trial-and-error methods to complete assignments would potentially benefit from restricted or delayed access to immediate hints and feedback. Nevertheless, although on-demand hints and immediate feedback were not available in the Delayed Condition, the decrease in performance associated with gaming behavior was not completely mitigated. Therefore, these results suggest that while removing on-demand instruction may assist students inclined towards gaming the system, further intervention is likely required to fully alleviate the detrimental effects of such behavior and address the behavior's root causes.

Delaying hints and feedback can be viewed as a proactive intervention targeting gaming the system behavior. This approach, similar to others employed in previous research (Aleven et al., 2006; Murray and VanLehn, 2005), uses a uniform approach for all students. The differential effect observed for delayed condition supports Murray and VanLehn (2005)'s suggestion that discouraging certain students from using hints may be beneficial for some students despite the overall negative impact on the student population; however, it is notable that only a small minority of the most ardent gamers may have benefited from delaying hints and feedback access.

A possible approach to address the varying effects includes withholding instant hints and feedback when students exhibit gaming behaviors, thus providing a focused corrective measure to steer their attention back to learning from the task at hand. Such a strategy would preserve the availability of real-time help for students who use the system appropriately while potentially benefiting those who exploit the system whenever they engage with it properly. An adaptable system like this could also lessen the damaging impact of manipulative behaviors by permitting gamers to reap the benefits of both instant and delayed hints and feedback. Nevertheless, this is unlikely to mitigate the antecedents of gaming fully without additional remediation. Furthermore, it is crucial to recognize that adopting this method could lead to feelings of frustration and a loss of engagement in some students, making it important to further explore and validate these ideas through additional causal research.

6.3. Combining Detection Models with FLPS

This study also illustrates the possible benefits of integrating detection and causal methods within EDM to explore how systems should effectively respond after identifying certain behaviors or latent states. Using artificial intelligence for detection purposes in CBLPs often leads learning experience designers to confront the dilemma of subsequent steps (for example, "Now that we know a student is frustrated, what do we do?"). FLPS offers a solution by pairing detector outputs with causal models to determine which program features may uniquely benefit students with particular behavioral patterns. Although rule-based detection methods were employed in our current investigation, future research could merge AI prediction systems with FLPS to go beyond assessing student experiences and behaviors in CBLPs to identify optimal alterations within these programs to boost their educational effectiveness.

7. Limitations & Future Directions

Although our findings suggest that the CBLPs' impacts differ based on students' tendencies to game the system, it is important to acknowledge the limitations of this analysis. There remains some uncertainty regarding whether the main and interaction effects in the model significantly differ from zero. Overall, for the estimates we considered likely to be different from zero, there was at most a 10% posterior probability that the true parameter had the opposite sign compared to the estimate presented in this paper. Thus, it is important that future research substantiate these findings to increase the certainty of their veracity.

Furthermore, while this study provides information about how students who game the system in a traditional CBLP respond to alternative CBLPs, it does not fully illuminate why. The patterns of effect sizes found here do not provide a simple principle for addressing gaming the system behaviors. More research is necessary to understand why some interventions work well for gamers whereas others do not.

The current study also does not address whether gaming behaviors are consistent across programs or whether the differential impact of these conditions is a function of changes in gaming behaviors. Future work should consider whether students' gaming behaviors are consistent across programs. This could be done using causal mediation analysis to evaluate whether variation in gaming behavior mediates the relation between conditions and learning outcomes.

Another limitation of this study is that the FLPS results depend on the gaming detector's accuracy. If there are systematic errors in the gaming detector, those errors should be considered while interpreting the FLPS model results. For example, some students may be gaming the system in principle but still have lower action rates. The Cognitive Model may not detect these students, and their propensities to game the system may be underestimated. The effect heterogeneity does not apply to profiles of gamers who are not detected. Conversely, if students are systematically misidentified as gaming, they will also be misidentified as part of gaming strata, which may influence the heterogeneity effects.

Finally, it is important to acknowledge the constraints inherent to the use of FLPS. The estimated impacts using FLPS heavily depend on the inherent quality of the model itself, and it remains uncertain how inaccuracies in estimating the model could potentially lead to biased outcomes. Future research should focus on gaining a more accurate understanding of how to assess these models to confirm that they furnish impartial appraisals of intervention effects.

8. CONCLUSION

The study indicates that students who are likely to game the system respond differently to various CBLPs, with none of the tested conditions fully mitigating the adverse effects of gaming the system. Delayed access to hints and feedback potentially benefits such students; however, the effect size was not large enough to outweigh the effect of their gaming tendencies on learning completely. This highlights the need for targeted interventions to help these students fully benefit from CBLPs' content.

The studied gamified conditions produced inconsistent heterogeneity patterns—one (DragonBox) showed a consistent effect regardless of gaming tendencies, whereas the other (FH2T) indicated adverse effects for gamers despite the fact that these programs employed many similar gamification methods. This emphasizes that how a program is gamified, rather than the mere presence of gamification, might be key in influencing student behavior. It is possible that FH2T's abstract goals caused more student frustration and consequent disengagement for students who were likely to game. On the other hand, DragonBox exhibited a positive effect independent of the student's propensity to game the system, possibly attributable to its emphasis on taking abstract mathematical processes and teaching them through non-mathematical puzzles. However, understanding the differences in heterogeneity patterns across the gamified conditions warrants further research.

The promising interaction between students' gaming tendencies and the Delayed Condition suggests that restricting immediate hints and feedback for those gaming the system might be effective at ameliorating some of the effects of gaming behaviors. Nevertheless, such a strategy could lead to feelings of frustration and decreased engagement in some students. Consequently, more research is necessary to validate these ideas and investigate further interventions.

Acknowledgments

The authors would like to thank past and current funders, including IES (R305D210036, R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305A120125 & R305R220012), NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), GAANN (P200A120238, P200A180088, P200A150306 & P200A150306), EIR (U411B190024 S411B210024, & S411B220024), ONR (N00014-18-1-2768), NIH (via SBIR R44GM146483), Schmidt Futures, BMGF, CZI, Arnold, Hewlett and a $180,000 anonymous donation. The opinions expressed in this paper are those of the authors and do not represent the views of the funders.

References

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., and Alibali, M. W. 2020. The future of embodied design for mathematics teaching and learning. *Frontiers in Education 5*, 1–29.

Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., and van Velsen, M. 2014. Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior 36*, 401–411.

Adjei, S. A., Baker, R. S., and Bahel, V. 2021. Seven-year longitudinal implications of wheel spinning and productive persistence. In *Artificial Intelligence in Education: 22nd International Conference (AIED 2021)*, I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and D. V., Eds. Springer International Publishing, 16–28.

Aleven, V. and Koedinger, K. R. 2000. Limitations of student control: Do students know when they need help? In *Intelligent Tutoring Systems*, G. Gauthier, C. Frasson, and K. VanLehn, Eds. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 292–303.

Aleven, V., McLaren, B., Roll, I., and Koedinger, K. 2006. Toward meta-cognitive tutoring: A model of helpseeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education 16*, 101–130.

Aleven, V., Roll, I., McLaren, B. M., and Koedinger, K. R. 2016. Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education 26,* 1 (Mar), 205–223.

Arezooji, D. M. 2020. A markov chain monte-carlo approach to dose-response optimization using probabilistic programming (rstan). *arXiv Preprint*.

Baker, R., Carvalho, A., Raspat, J., Aleven, V., and Koedinger, K. R. 2009. Educational software features that encourage and discourage “gaming the system”. In *Artificial Intelligence in Education: 14th International Conference (AIED 2009)*, S. D. Craig and D. Dicheva, Eds. Vol. 14. Springer International Publishing, 475–482.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., and Koedinger, K. 2008. Why students engage in “gaming the system” behavior in interactive learning environments. *Journal of Interactive Learning Research 19,* 2, 185–224.

Baker, R. S., Corbett, A. T., Koedinger, K. R., and Wagner, A. Z. 2004. Off-task behavior in the cognitive tutor classroom: when students ”game the system”. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2004).* Association for Computing Machinery, 383–390.

BAKER, R. S. J. D., CORBETT, A. T., KOEDINGER, K. R., EVENSON, S., ROLL, I., WAGNER, A. Z., NAIM, M., RASPAT, J., BAKER, D. J., AND BECK, J. E. 2006. Adapting to when students game an intelligent tutoring system. In *Intelligent Tutoring Systems: 8th International Conference*, M. Ikeda, K. D. Ashley, and T.-W. Chan, Eds. Springer, Jhongil, Taiwan, 392–401.

BAKER, R. S. J. D., CORBETT, A. T., KOEDINGER, K. R., AND ROLL, I. 2006. Generalizing detection of gaming the system across a tutoring curriculum. In *8th International Conference of Intelligent Tutoring Systems (ITS 2006)*, M. Ikeda, K. D. Ashley, and T.-W. Chan, Eds. Spinger, 402–411.

BAKER, R. S. J. D., CORBETT, A. T., ROLL, I., AND KOEDINGER, K. R. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction 18,* 3 (Aug), 287–314.

BAKER, R. S. J. D., MITROVIĆ, A., AND MATHEWS, M. 2010. Detecting gaming the system in constraint-based tutors. In *User Modeling, Adaptation, and Personalization*, P. De Bra, A. Kobsa, and D. Chin, Eds. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 267–278.

BUTLER, A. C. AND WOODWARD, N. R. 2018. Toward consilience in the use of task-level feedback to promote learning. *Psychology of Learning and Motivation 69*, 1–38.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P., AND RIDDELL, A. 2017. Stan: A probabilistic programming language. *Journal of statistical software 76*, 1–32.

CAYTON-HODGES, G. A., FENG, G., AND PAN, X. 2015. Tablet-based math assessment: What can we learn from math apps? *Journal of Educational Technology & Society 18,* 2, 3–20.

CHIU, L.-H. AND HENRY, L. L. 1990. Development and validation of the mathematics anxiety scale for children. *Measurement and evaluation in counseling and development 23,* 3, 121–127.

CORBETT, A. T. AND ANDERSON, J. R. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4,* 4 (Dec), 253–278.

DANG, S. AND KOEDINGER, K. 2019. Exploring the link between motivations and gaming. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, C. F. Lynch, A. Merceron, M. Desmarais, and N. R., Eds. International Educational Data Mining Society, 276–281.

DECKER-WOODROW, L. E., MASON, C. A., LEE, J.-E., CHAN, J. Y.-C., SALES, A., LIU, A., AND TU, S. 2023. The impacts of three educational technologies on algebraic understanding in the context of covid-19. *AERA Open 9*, 23328584231165919.

DIETER, K. C., STUDWELL, J., AND VANACORE, K. P. 2020. Differential responses to personalized learning recommendations revealed by event-related analysis. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, Eds. Vol. 13. International Educational Data Mining Society, Online, 736–742.

DIHOFF, R. E., BROSVIC, G. M., AND EPSTEIN, M. L. 2003. The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record 53,* 4, 533–548.

DOLONEN, J. A. AND KLUGE, A. 2015. Algebra learning through digital gaming in school. In *Exploring the Material Conditions of Learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015*. Vol. 1. International Society of the Learning Sciences, Inc. [ISLS]., 252–259.

FENG, M. AND HEFFERNAN, N. T. 2006. Informing teachers live about student learning: Reporting in the assistments system. *Technology Instruction Cognition and Learning 3*, 1–14.

FRANGAKIS, C. E. AND RUBIN, D. B. 2002. Principal stratification in causal inference. *Biometrics 58,* 1, 21–29.

GARRIS, R., AHLERS, R., AND DRISKELL, J. E. 2002. Games, motivation, and learning: A research and practice model. *Simulation& Gaming 33,* 4 (Dec.), 441–467.

GASTON, J. AND COOPER, S. 2017. To three or not to three: Improving human computation game onboarding with a three-star system. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI 2017)*, G. Mark and P. Fussell, Eds. ACM, Denver Colorado USA, 5034–5039.

GEE, J. P. 2005. Learning by design: Good video games as learning machines. *E-Learning and Digital Media 2,* 1.

GOLDSTONE, R. L., LANDY, D. H., AND SON, J. Y. 2010. The education of perception. *Topics in Cognitive Science 2,* 2, 265–284.

GOODRICH, B., GABRY, J., ALI, I., AND BRILLEMAN, S. 2020. rstanarm: Bayesian applied regression modeling via Stan. R Package Version 2.21.1.

GURUNG, A., BARAL, S., LEE, M. P., SALES, A. C., HAIM, A., VANACORE, K. P., MCREYNOLDS, A. A., KREISBERG, H., HEFFERNAN, C., AND HEFFERNAN, N. T. 2023. How common are common wrong answers? crowdsourcing remediation at scale. In *Proceedings of the Tenth ACM Conference on Learning@ Scale (L@S2023)*, D. Spikol, A. P. Viberg, A. Martínez-Monés, and P. Guo, Eds. Association for Computing Machinery, 70–80.

GURUNG, A., BARAL, S., VANACORE, K. P., MCREYNOLDS, A. A., KREISBERG, H., BOTELHO, A. F., SHAW, S. T., AND HEFFERNA, N. T. 2023. Identification, exploration, and remediation: Can teachers predict common wrong answers? In *LAK23: 13th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery, 399–410.

HEFFERNAN, N. T. AND HEFFERNAN, C. L. 2014. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education 24,* 4 (Oct), 470–497.

JIN, H. AND RUBIN, D. B. 2008. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association 103,* 481, 101–111.

JUUL, G. J. 2009. Routledge, Chapter Fear of Failing? The Many Meanings of Difficulty in Video, 237–252.

KARAGIORGAS, D. N. AND NIEMANN, S. 2017. Gamification and game-based learning. *Journal of Educational Technology Systems 45,* 4 (June), 499–519.

LANDERS, R. N. 2014. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation& Gaming 45,* 6 (Dec.), 752–768.

LEE, M., CROTEAU, E., GURUNG, A., BOTELHO, A., AND HEFFERNAN, N. 2023. Knowledge tracing over time: A longitudinal analysis. In *The Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023).*, M. Feng, T. Kaser, and P. Talukdar, Eds. International Educational Data Mining Society, 296–301.

LEVIN, N., BAKER, R., NASIAR, N., STEPHEN, F., AND HUTT, S. 2022. Evaluating gaming detector model robustness over time. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, 398–405.

LU, X., SALES, A., AND HEFFERNAN, N. T. 2021. Immediate versus delayed feedback on learning: Do people's instincts really conflict with reality? *Journal of Higher Education Theory and Practice 21,* 16 (Dec.).

LU, X., WANG, W., MOTZ, B. A., YE, W., AND HEFFERNAN, N. T. 2023. Immediate text-based feedback timing on foreign language online assignments: How immediate should immediate feedback be? *Computers and Education Open 5*, 1–12.

MALKIEWICH, L. J., LEE, A., SLATER, S., XING, C., AND CHASE, C. C. 2016. No lives left: How common game features could undermine persistence, challenge-seeking and learning to program. In *Proceedings of The International Conference of the Learning Sciences (ICLS) 2016*. International Society of the Learning Sciences, 186–193.

MCKERNAN, B., MARTEY, R. M., STROMER-GALLEY, J., KENSKI, K., CLEGG, B. A., FOLKESTAD, J. E., RHODES, M. G., SHAW, A., SAULNIER, E. T., AND STRZALKOWSKI, T. 2015. We don't need no stinkin' badges: The impact of reward features and feeling rewarded in educational games. *Computers in Human Behavior 45*, 299–306.

MCLAREN, B. M., VAN GOG, T., GANOE, C., KARABINOS, M., AND YARON, D. 2016. The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computers in Human Behavior 55*, 87–99.

MIDGLEY, C., MAEHR, M. L., HRUDA, L. Z., ANDERMAN, E., ANDERMAN, L., FREEMAN, K. E., URDAN, T., ET AL. 2000. *Manual for the patterns of adaptive learning scales*. University of Michigan.

MIHAELA, C. AND HERSHKOVITZ, A. 2009. The impact of off-task and gaming behaviors on learning: Immediate or aggregate? In *Frontiers in Artificial Intelligence and Applications: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, V. Dimitrova, R. Mizoguchi, B. du Boulay, and A. Graesser, Eds. Vol. 200. Ios Press, 507–514.

MURRAY, R. C. AND VANLEHN, K. 2005. Effects of dissuading unnecessary help requests while providing proactive help. In *Artificial Intelligence in Education: 12th International Conference (AIED 2005)*, C.-K. Looi, G. McCalla, B. Bredeweg, and J. Breuker, Eds. Springer International Publishing, 887–889.

OTTMAR, E., LEE, J.-E., VANACORE, K., PRADHAN, S., DECKER-WOODROW, L., AND MASON, C. A. 2023. Data from the efficacy study of from here to there! a dynamic technology for improving algebraic understanding. *Journal of Open Psychology Data 11,* 1 (Apr), 1–15.

PAGE, L. C. 2012. Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness 5,* 3, 215–244.

PAGE, L. C., FELLER, A., GRINDAL, T., MIRATRIX, L., AND SOMERS, M.-A. 2015. Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation 36,* 4 (Dec), 514–531.

PAQUETTE, L. AND BAKER, R. S. 2017. Variations of gaming behaviors across populations of students and across learning environments. In *Artificial Intelligence in Education: 17th International Conference (AIED 2017)*, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Springer International Publishing, Cham, 274–286.

PAQUETTE, L. AND BAKER, R. S. 2019. Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. *Interactive Learning Environments 27,* 5–6 (Aug), 585–597.

PAQUETTE, L., BAKER, R. S., DE CARVALHO, A., AND OCUMPAUGH, J. 2015. Cross-system transfer of machine learned and knowledge engineered models of gaming the system. In *User Modeling, Adaptation and Personalization*, F. Ricci, K. Bontcheva, O. Conlan, and S. Lawless, Eds. Lecture Notes in Computer Science. Springer International Publishing, Cham, 183–194.

PAQUETTE, L., DE CARVALHO, AND A. M. J. A., & BAKER, R. S. 2014. Towards understanding expert coding of student disengagement in online learning. In *Proceedings of the 36th Annual Cognitive Science Conference*, P. Bello, M. Guarini, M. McShane, and B. Scassellati, Eds. CogSci, 1–6.

PARDOS, Z. A., BAKER, R. S. J. D., SAN PEDRO, M. O. C. Z., GOWDA, S. M., AND GOWDA, S. M. 2014. Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics 1,* 1, 107–128.

PATIKORN, T. AND HEFFERNAN, N. T. 2020. Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale*. Association for Computing Machinery, 115–124.

PHYE, G. D. AND ANDRE, T. 1989. Delayed retention effect: Attention, perseveration, or both? *Contemporary Educational Psychology 14,* 2 (Apr), 173–185.

PRIHAR, E., PATIKORN, T., BOTELHO, A., SALES, A., AND HEFFERNAN, N. 2021. Toward personalizing students' education with crowdsourced tutoring. In *L@S 2021 - Proceedings of the 8th ACM Conference on Learning @ Scale*. Association for Computing Machinery, 37–45.

R CORE TEAM. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAZZAQ, L., HEFFERNAN, N. T., AND LINDEMAN, R. W. 2007. What level of tutor interaction is best? *Frontiers in Artificial Intelligence and Applications 158*, 222–229.

RICHEY, J. E., ZHANG, J., DAS, R., ANDRES-BRAY, J. M., SCRUGGS, R., MOGESSIE, M., BAKER, R. S., AND MCLAREN, B. M. 2021. Gaming and confrustion explain learning advantages for a math digital learning game. In *Artificial Intelligence in Education: 22nd International Conference*, I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, Eds. Lecture Notes in Computer Science. Springer International Publishing, Utrecht, The Netherlands, 342–355.

RODRIGO, M. M. T., BAKER, R. S. J. D., D'MELLO, S., GONZALEZ, M. C. T., LAGUD, M. C. V., LIM, S. A. L., MACAPANPAN, A. F., PASCUA, S. A. M. S., SANTILLANO, J. Q., SUGAY, J. O., TEP, S., AND VIEHLAND, N. J. B. 2008. Comparing learners' affect while using an intelligent tutoring system and a simulation problem solving game. In *Intelligent Tutoring Systems*, B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, Eds. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 40–49.

ROSCHELLE, J., FENG, M., MURPHY, R. F., AND MASON, C. A. 2016. Online mathematics homework increases student achievement. *AERA Open 2,* 4 (Oct.), 2332858416673968.

SALES, A. C. AND PANE, J. F. 2019. The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics 13,* 1 (Mar), 420–443.

SHUTE, V. J. 2008. Focus on formative feedback. *Review of Educational Research 78,* 1 (Mar), 153–189.

SIEW, N. M., GEOFREY, J., AND LEE, B. N. 2016. Students' algebraic thinking and attitudes towards algebra: The effects of game-based learning using dragonbox 12 + app. *The Research Journal of Mathematics and Technology 5,* 1, 66–79.

STAR, J. R., POLLACK, C., DURKIN, K., RITTLE-JOHNSON, B., LYNCH, K., NEWTON, K., AND GOGOLEN, C. 2015. Learning from comparison in algebra. *Contemporary Educational Psychology 40*, 41–54.

STEKHOVEN, D. J. AND BUEHLMANN, P. 2012. Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics 28,* 1, 112–118.

TORRES, R., TOUPS, Z. O., WIBURG, K., CHAMBERLIN, B., GOMEZ, C., AND OZER, M. A. 2016. Initial design implications for early algebra games. In *Proceedings of the 2016 Annual Symposium on*

Computer-Human Interaction in Play Companion Extended Abstracts. CHI PLAY Companion '16. Association for Computing Machinery, New York, NY, USA, 325–333.

VANACORE, K., GURUNG, A., SALES, A., AND HEFFERNAN, N. T. 2024. The effect of assistance on gamers: Assessing the impact of on-demand hints & feedback availability on learning for students who game the system. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. Association for Computing Machinery, 462–472.

VANACORE, K., OTTMAR, E., LIU, A., AND SALES, A. 2024. Remote monitoring of implementation fidelity using log-file data from multiple online learning platforms. *Journal of Research on Technology in Education*, 1–21.

VANACORE, K., SALES, A., LIU, A., AND OTTMAR, E. 2023a. Benefit of gamification for persistent learners: Propensity to replay problems moderates algebra-game effectiveness. In *Tenth ACM Conference on Learning @ Scale (L@S '23)*, D. Spikol, O. Viberg, A. Martínez-Mones, and P. Guo, Eds. ACM, Copenhagen, Denmark, 164–173.

VANACORE, K., SALES, A., LIU, A., AND OTTMAR, E. 2023b. Heterogeneous effects of game-based failure on student persistence in an online algebra game. In *Society for Research Educational Effectiveness Conference (SREE 2023)*. SREE, 1–4.

VANACORE, K., SALES, A. C., HANSEN, B., LIU, A., AND OTTMAR, E. 2024. Effect of game-based failure on productive persistence: an application of regression discontinuity design for evaluating the impact of program features on learning-related behaviors. *Available at Social Science Research Network 4789291.*

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B., AND BÜRKNER, P.-C. 2021. Rank-normalization, folding, and localization: An improved r hat for assessing convergence of mcmc (with discussion). *Bayesian analysis 16,* 2, 667–718.

WALONOSKI, J. A. AND HEFFERNAN, N. T. 2006. Prevention of off-task gaming behavior in intelligent tutoring systems. In *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings 8*, T.-W. C. Mitsuru Ikeda, Kevin D. Ashley, Ed. Lecture Notes in Computer Science. Springer, 722–724.

WILLIAMS, J. J., KIM, J., RAFFERTY, A., MALDONADO, S., GAJOS, K. Z., LASECKI, W. S., AND HEFFERNAN, N. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*. L@S '16. Association for Computing Machinery, New York, NY, USA, 379–388.

XIA, M., ASANO, Y., WILLIAMS, J. J., QU, H., AND MA, X. 2020. Using information visualization to promote students' reflection on "gaming the system" in online learning. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*. L@S '20. Association for Computing Machinery, New York, NY, USA, 37–49.

Chapter 3

The Effect of Game-Based Failure on Productive Persistence

3.1 Effect of Game-Based Failure on Productive Persistence*

See manuscript below.

Effect of Game-Based Failure on Productive Persistence:
An application of regression discontinuity design for evaluating the impact of program features on learning-related behaviors

Kirk Vanacore, Worcester Polytechnic Institute, kpvanacore@wpi.edu, 0000-0003-0673-5721
Adam C. Sales, Worcester Polytechnic Institute, asales@wpi.edu, 0000-0003-0416-0610
Ben B. Hansen, Univerity Of Michigan, bbh@umich.edu, 0000-0002-0478-4776
Allison Lui, Worcester Polytechnic Institute, aliu2@wpi.edu, 0000-0003-1075-2575
Erin Ottmar, Worcester Polytechnic Institute, eottmar@wpi.edu, 0000-0002-9487-7967

Abstract

Failure is an integral part of the learning process, which students often view as a necessary evil. They tend to experience failure as evaluative and demotivating in educational settings. However, in games, failure is essential in making the gaming experience enjoyable and even motivating. In the current study, we evaluated the impact of game-based failure on students' persistence behavior using a regression discontinuity design. We found that game-based failure increases the likelihood of students engaging in productive persistence as they play an online gamified algebra program. This finding suggests that gamification features in learning programs help sidestep the negative aspects of failure and leverage those failure experiences for learning. This work also illustrates the usefulness of regression discontinuity designs in evaluating the impact of features in online learning games to provide insight into causal mechanisms through which program features influence students' learning processes.

1. Introduction

Learning typically requires struggle; evidence suggests that instructional content should be presented in ways that induce difficulty to improve the longer-term retention of information (Bjork & Bjork, 2020). Difficult learning experiences include failure, such as answering questions incorrectly, taking inefficient routes to solutions, or displaying suboptimal performance on a task (Kapur, 2012; VanLehn, 1988). To learn, students must exhibit productive persistence – persistence that leads to positive learning outcomes – by continuing to employ effort during experiences of failure.

However, during many learning experiences, students are taught that failure during performance (e.g., submitting an incorrect answer to a problem) indicates failure to learn. For many students, high-stakes environments, such as standardized tests, have made educational institutions feel like evaluative environments that focus heavily on response accuracy, which may influence students' beliefs that accuracy is the goal of learning and inaccuracy is always negative (Jones, 2007; Jones & Egley, 2004; Klinger & Luce-Kapler, 2008). At best, students may feel that getting questions wrong is an opportunity to learn; at worst, they perceive these “failures” as indications of inherent academic inadequacies. This perspective may demotivate students from exerting the requisite effort for productive persistence.

Alternatively, failure can be viewed as part of the learning process. For example, forgetting during memory tasks is helpful in fortifying memories for long-term retention (Bjork & Bjork, 2020; Bjork, Robert A., 2014; McGeoch, 1932). Viewing forgetting as a necessary step towards retention may mitigate its adverse effects on students' self-perception and help them retain their motivation to persist during the learning task. In some cases, gamification can help students persevere through failures by lowering the stakes of these experiences.

In the following paper, we explore how instructional programs can alleviate the demotivating nature of failure by presenting failure in the context of a mathematics game. Using data from the algebra game *From Here to There,* we find evidence that game-based failure positively affects students' persistence behavior of replaying problems after suboptimal performance. To estimate this effect, we use a regression discontinuity design (RDD), which allows for causal effects estimation for decisions made on cut points. This work contributes to the field of research on digital education by documenting a causal relation between a feature of gamification and productive persistence. Furthermore, the work serves as an example of how RDD can be employed to evaluate causal mechanisms of features in education programs.

2. Background

2.1 Failure and Learning

There are a variety of theories addressing how failure experiences contribute to learning. According to Piaget's Cognitive Development Theory (1977), when a new experience contradicts a learner's existing schemas, the learner enters a state of disequilibrium. The learner must adjust schemas to accommodate this contradictory information and restore equilibrium. Inaccuracy may induce such states of disequilibrium, motivating the learner to generate new understandings and then consolidate them into their preexisting schemas. Similarly, VanLehn’s (1988) impasse-driven learning theory posits that for learning to occur, students must overcome impasses in their work caused by misconceptions or skill deficiency. Other theories view failures in learning as a warning that learners must modify their cognitive model (Schank, 1999) or an indication that one must build new memory networks that address the discontinuity between prior knowledge states and the knowledge necessary to provide the correct solution (Gartmeier et al., 2008).

Tawfik and colleagues (2015) consolidated these perspectives into one theory of failure-based learning. They posit that after experiencing failures, learners evaluate their failure, which they call "failure-based problem solving." This evaluation is an interactive process that includes challenging existing mental models, identifying potential reasons for failure, and testing new solutions. Once the learners find the correct solution, they expand their mental model to accommodate what they have learned and reflect on how they may apply this new knowledge in the future.

There is ample evidence that failures during the learning process can lead to better outcomes. Studies have found that problem-solving exercises prior to instruction are beneficial to learning, even when learners respond incorrectly during/to these pre-instruction exercises (Kapur, 2012; Kapur & Bielaczyc, 2011; Loibl et al., 2017; Loibl & Rummel, 2014). This phenomenon is known as productive failure as the failure experience benefits the learner. Notably, Hartmann and colleagues (2021) found that even examining others' failed responses before instruction had learning benefits. Similarly, Loibl & Leuders (2019) found that examining incorrect responses and comparing them to correct ones is more effective than just learning correct responses (Loibl & Leuders, 2019). Overall, a meta-analysis on learning from failure showed that productive failure and failure-driven memory conditions had positive effects on learning in experimental studies (Darabi et al., 2018).

2.2 Persistence in Learning

Tawfik and colleagues (2015) note that all failure-based learning theories posit that failures are indications learners need to add an additional inquiry step into their learning process, which would not be induced by an unsuccessful response. Therefore, for a failure experience to be productive, the learner must persist through the failure to progress in the educational task. Persistence is generally defined as an aspect of conscientiousness in which a person is driven to complete challenging tasks (McClelland, 1961). It is considered essential for learning as persisting during difficult learning tasks allows students to work in the upper end of their zones of proximal development, where most learning occurs (Ventura, 2013; Vygotsky, 1978). A recent analysis of data from intelligent tutoring systems found consistency in student learning rates, suggesting that students do not vary in their ability to learn but in the number of opportunities required to learn (Koedinger et al., 2023). This suggests that all students master each knowledge component by exerting the requisite persistence.

Despite the centrality of persistence in learning, computer-based learning platforms, such as massive open online courses (MOOC) and intelligent tutors, have struggled to encourage students to persist in their educational tasks. Kizilcec and colleagues (2020) found few positive effects of behavioral interventions in MOOCs in encouraging students to persist by completing their courses. Janelli and Lipnevick (2021) found that pretests in a MOOC negatively impacted persistence while positively impacting learning for those completing the course. In intelligent tutors, Botelho and colleagues (2019) found evidence of students showing low persistence through a behavior they called "stopout," in which students refuse to exert effort in learning tasks by simply quitting the task often after submitting an incorrect response. Furthermore, an analysis of seven different behavioral interventions in an intelligent tutor found that most interventions failed to encourage students to master the knowledge component (Vanacore et al., 2023). Thus, there is a lack of understanding of exactly how to foster persistence in digital learning programs.

Although we generally desire students to exhibit persistence, not all persistence yields positive learning outcomes. Researchers distinguish between productive persistence, in which students adjust their strategies and behaviors in response to the failure to meet their learning goals

(Kai et al., 2018), versus unproductive persistence, in which students spend time struggling through failure without adjusting and achieving learning mastery (Baker et al., 2013, p. 200; Beck & Gong, 2013). Productive persistence has been associated with better long-term academic outcomes, whereas unproductive persistence is associated with poorer learning outcomes (Adjei et al., 2021; Beck & Gong, 2013). Therefore, it is important that students engage in productive behaviors after experiencing failures in order to experience the benefits of persistence.

2.2 Persistence Through Game-based Failure in Learning Environments

2.2.1 Gamified Instruction and Learning Behaviors. Gamification in learning environments may be one way to encourage productive persistence. In theory, gamified learning systems influence students' behaviors as they learn, which impacts their learning outcomes (Landers, 2014). However, studies on gamification have shown mixed effects on students' behaviors in learning contexts. Some work suggests that adding rewards and scorekeeping to learning programs has a positive impact on learning-related behaviors (Garris et al., 2002; Gaston & Cooper, 2017), while others have failed to find effects (Malkiewich et al., 2016; McKernan et al., 2015). For example, Garris and colleagues (2002) found that reward-based systems caused students to take more time and replay problems more frequently. Alternatively, Malkiewich and colleagues (2015) found that game-based features such as feedback and narrative elements did not positively impact students' persistence within the game. This variance in results suggests that the mechanisms that connect different gamification elements in learning programs to positive learning behaviors must be better understood to ensure that learning games are developed effectively to support these outcomes.

2.2.2 Gamification and Persistence. One potential feature of gamification that may promote persistence behaviors is problem replayability. In many traditional learning programs, such as mastery learning intelligent tutors, students attempt problems by submitting either a correct or incorrect response. While they may submit another answer after an incorrect response, once they have either submitted the correct answer or been given it by the system, the problem is complete. It makes little sense to reattempt the problem as they have already been given the solution. Alternatively, some gamified learning systems construct problems with multiple paths to a solution (Iseli et al., 2021; Z. Liu et al., 2017; Resnick et al., 2009; Woo & Falloon, 2022). In these systems, students can replay a problem directly after submitting a correct response to find another path to the solution.

Several studies illustrate the benefit of using these persistence opportunities in learning (Clark et al., 2011; Lee et al., 2022; Liu et al., 2017; Vanacore et al., 2023). In one online math game, a cluster of students who spent a large proportion of game time replaying completed problems showed the largest learning gains in algebraic understanding (J.-E. Lee et al., 2022). Furthermore, replaying problems or levels is associated with better learning outcomes (Clark et al., 2011; Li et al., 2010). Liu and colleagues (2017) found replaying puzzles in a gamified math program (ST Math) was associated with positive learning outcomes, but only if the students had already completed the level with the replayed puzzles. Lavoué and colleagues (2021) also found heterogeneity associated with replay behavior in a gamified math program: students' motivation varied based on their orientation towards replaying problems. Students with achievement-oriented engagement, in which students tried to attain the best possible performance on their first problem attempts, experienced greater intrinsic motivation than those with a perfection-oriented engagement, in which students replayed problems to try to improve their performance. Overall, replaying problems seems to be a productive persistence behavior in some contexts but not in others.

2.2.3 Game-Based Failure. One way that gamification may help learners persist through failure is by lowering the stakes of the failure. In traditional learning platforms, students may be more likely to view their performance as evaluative if feedback is presented as problem accuracy. Furthermore, presenting students' performance data to teachers through dashboards – as is common in many computer-based instruction programs (e.g., Heffernan & Heffernan, 2014; Molenaar & Knoop-van Campen, 2017; Vanacore et al., 2021) – may exacerbate students' perceptions of inaccuracy as inadequacy. The evaluative nature of this feedback could lead students to adopt a performance orientation, thus negatively affecting their intrinsic motivation and potentially influencing their persistence behaviors (Ryan & Deci, 2000; Lui, under review). Alternatively, games require difficulties that induce occasional failure to be engaging.

In video game theory, although a player's goal is to win, failure adds to the gaming experience by serving "as a contrast to winning" as well as adding content that makes the player "see the nuances of the game" (Juul, 2008). Similarly, in learning games, if failures are presented as a part of gameplay, they add to the learner's experience by providing a challenge while cueing them to attend to the nuances of content. Furthermore, students may perceive failure in a game as less evaluative and, therefore, less demotivating (Williams-Pierce, 2019). Gee (Gee, 2005) notes that learning games can make failure "pleasantly frustrating" by indicating to them "how and if they are making progress" and allowing the learner to fail without concluding, "I am a failure." Thus, learning games can present failure in ways that are salient to the learner while retaining an engaging playfulness.

3.1 Current Study

Although there is evidence that gamification can help students persist (Clark et al., 2011; Lee et al., 2022; Liu et al., 2017; Vanacore et al., 2023), and it is theoretically plausible that this enhanced persistence is due in part to the engaging nature of game-based failure (Juul, 2008; Williams-Pierce, 2019), there is a lack of causal research linking game-based failure to the persistence of educational games. In the current study, we address this gap by exploring the effect of receiving failure feedback on learners' likelihood of replaying problems after they exhibit suboptimal performance. To do so, we use a regression discontinuity design (RDD) within the treatment arm of a randomized controlled trial testing the efficiency of a gamified algebra program, *From Here To There!* (FH2T; Ottmar et al., 2015). An RDD is a quasi-experimental design that enables researchers to estimate the causal effects of a condition administered at a cut point along a scale. The primary purpose of this paper is to evaluate the impact of game-based failure on students' persistence behavior.

As a secondary goal, we seek to illustrate the usefulness of RDD in understanding digital learning environments. This design is commonly employed in economics, public policy, and education research (Lee & Lemieux, 2009; Ludwig & Miller, 2007), but it is underutilized in human-computer interaction research on educational platforms. Yet, features in digital learning environments are often good candidates for RDD, because many decisions in these programs are made based on a cut point. Some examples include the administering of rewards (Liu et al., 2022), prescribing usage recommendations (Dieter et al., 2020), and determining whether students have mastered a knowledge component (Bloom, 1968; Kelly et al., 2015; Yudelson et al., 2013). These mechanisms, which determine how students experience learning programs, provide opportunities for understanding the impacts of these associated features. We use our estimation of the effect of game-based failure on persistence to demonstrate how RDD can be leveraged for casual research to help understand how digital learning impacts students' learning processes and outcomes.

3.1 Gamified Algebra Instruction in FH2T

FH2T takes a nontraditional approach to algebra instruction by applying aspects of perceptual learning (Goldstone et al., 2010) and embodied cognition (Abrahamson et al., 2020). Instead of presenting students with typical algebra equations and expressions that students solve or simplify, FH2T gives students access to a starting expression (start state) that they must dynamically transform into a mathematically equivalent expression (goal state). Students can manipulate the expression by dragging numbers and symbols from one position to another on the screen or using a keypad when expanding terms. Only mathematically valid manipulations are accepted. Each valid manipulation counts as a step, which is logged and used to evaluate how efficiently the student transforms the expression from the start to the goal state. Figure 1 displays an example problem. FH2T has 252 problems that are presented sequentially by mathematical content and complexity. Students must complete one problem in the sequence before advancing to the next problem.

Several randomized control trials with middle school students consistently showed that students in the FH2T condition outperformed on algebra assessments compared to students who completed traditional online problem sets, including multiple choice and fill-in-the-blank problem sets with automated hints and feedback (Chan et al., 2022; Decker-Woodrow et al., 2023; Hulse et al., 2019). While these studies cannot determine which specific mechanisms led to improved mathematics outcomes in students, there is reason to believe that the gamification in the program played a role in its effectiveness, as the main difference between comparison conditions was the inclusion of gamification.

FH2T has some notable elements of gamification. The ability to dynamically manipulate expressions into a goal makes the problem seem like a puzzle to be solved instead of a math question to be answered. The mathematical rules of algebra govern the world and system, and, as with many games that impose parameters on the users, the rules of math cannot be broken. Finally, the evaluation mechanism provides characteristics of game-based failure, as described in the next section.

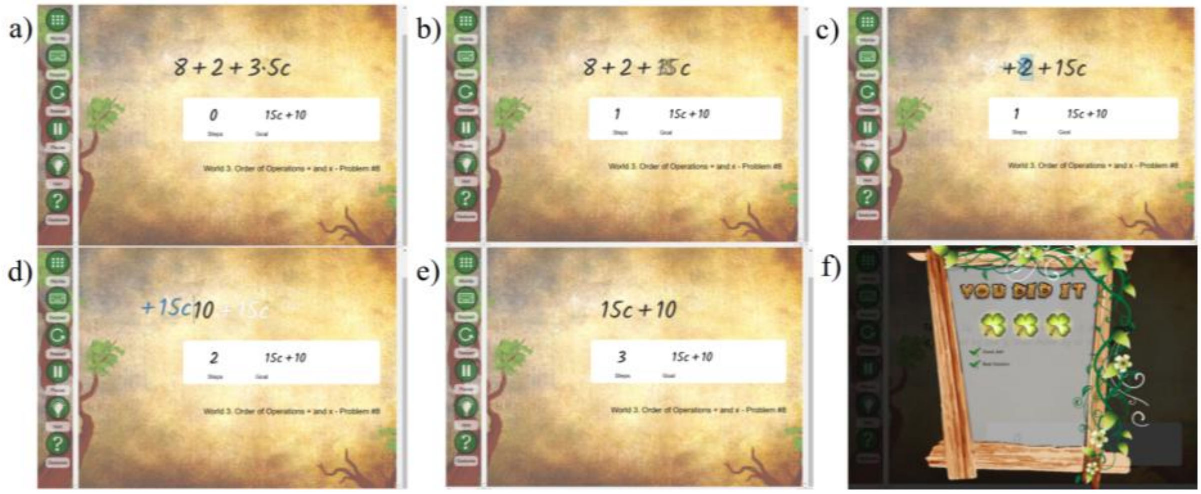


Figure 1: A sample problem in the gamified condition.

From the initial expression (a), the student taps on 3 5c to automatically multiply the 3 and 5 together (b). The student then drags 8 on top of the 2 to automatically add 8 and 2 together (c). Finally, the student drags the 15c to the other side of the 10 (d) to reorder the terms and reach the goal state (e). Finally, the student earned the maximum reward - three clovers - for completing the problem in the minimum number of steps (f).

3.3 Game-Based Failure in FH2T

In FH2T, each problem has an optimal solution (i.e., a minimum number of steps to complete the problem), and students receive feedback in the form of clovers based on how close they are to the optimal solution (Figure 2). The clovers act as performance contingent rewards: three clovers for optimal performance, two clovers if the student's solution took one or two steps more than the optimal solution, and one clover if their solution took more than two steps. When students are close to the optimal solution and receive two clovers, they are still congratulated by being told "Good Job" despite the crossed-out "Best Solution." However, when they only earn one clover, both "Good Job" and "Best Solution" are crossed out, communicating to the students that their performance needs improvement. Thus, we propose that *receiving one clover is an instance of game-based failure*, communicating that although the student completed the task, they did not do a good job. Notably, this information is only available to the student themselves and not their instructors. Therefore, it is still a low-stake failure that exists only in the game.

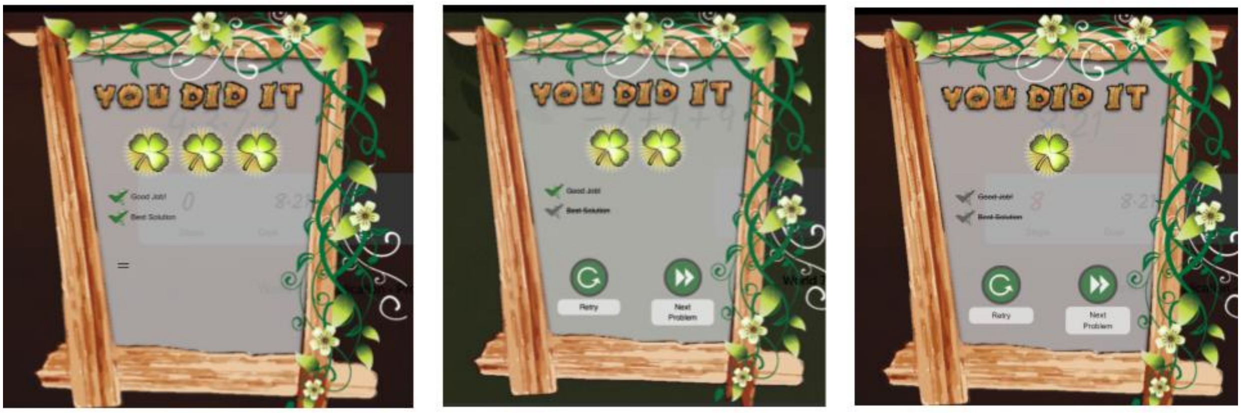


Figure 2: Performance-based feedback in FH2T

The clover reward screen shown after each problem. Students receive three clovers for an optimal solution, two clovers if they are within two steps of optimal, and one clover if they are over optimal. Students are given the option to either retry the problem again (a replay attempt) or to move on to the next problem if they earn one or two clovers and are automatically progressed to the next problem if they earn three clovers.

3.2 Replaying problems in FH2T as productive persistence

When students do not perform optimally, earning fewer than three clovers, they are encouraged to replay the problem. This encouragement is communicated by a replay button on the reward screen (Figure 2) displayed when students receive one or two clovers. Previous research found that students who spend more time replaying problems experience greater growth in algebraic knowledge (J.-E. Lee et al., 2022). Furthermore, we found that upon replaying the problems, students are likely to improve their performance on the replayed problem and are more likely to achieve optimal performance on subsequent problems than those who did not replay (Liu et al., under review). In the same study, a qualitative evaluation of students' replayed solutions revealed not only improved efficiency but also that students had adjusted their strategies, indicating improved algebraic understanding. Finally, in a recent study, we found that students with a higher propensity for replaying problems experience a greater impact of FH2T on their algebraic knowledge than those with a lower propensity for replaying problems (Vanacore et al., 2023)**.** This evidence suggests that replaying problems in FH2T constitutes a form of productive persistence because it is preceded by difficulty in the form of suboptimal performance and because it is associated with improvements in performance within the program and enhanced efficacy of the program on learning.

3.3 Research Questions

In the following study, we evaluate the impact of game-based failure on the productive persistence behavior of replaying problems. Furthermore, we examine whether the effect varies based on the problem's complexity and the student's prior knowledge to evaluate any heterogeneity in the effect.

This study has three research questions:

RQ1. Does experiencing game-based failure impact students' productive persistence?

RQ2. Does this effect vary based on the complexity of the problem?

RQ3. Does this effect differ for students with higher or lower prior knowledge?

4. Method

As mentioned above, this analysis involves a variation of RDD, Limitless Regression Discontinuity (LDR; Sales & Hansen, 2020). RDD allows for the estimation of causal effects when a cut point on a scale determines treatment assignment. In this case, the treatment of game-based failure (receiving one clover in FH2T) is entirely determined by students' performance on each problem (steps over optimal). Therefore, though the treatment of game-based failure is not randomized, treatment assignment is determined by one confounder, which is known. We use LDR because the method is amenable to binary outcomes and discrete scales. This section describes the data used in the LDR (Section 4.1) and explains the research design, assumptions, and procedure (Section 4.2).

4.1 Data

The current study uses secondary data from a larger efficacy study, which took place during the 2021-22 school year (Decker-Woodrow et al., 2023). A total of 52 seventh-grade mathematics teachers and their students from 11 middle schools were recruited from a large, suburban district in the Southeastern United States. The efficacy study included four conditions, but we only focused on the gamified condition described above for this analysis. Data for the entire study are open-sourced and available through the *Open Science Framework* (OSF; https://osf.io/r3nf2/). A full explanation of the data is available in Ottmar et al. (2023).

A total of 1,430 students were randomized into the FH2T condition, 1130 of whom completed at least one problem in the game. During the study, these 1130 students attempted 110,523 problems. For the current analyses, we are exclusively concerned with the instances in which students were encouraged to display productive persistence by replaying the problem after the suboptimal performance (i.e., only instances when students received one or two clovers). This encouragement came in the form of the "Retry" button being available, as described in section 3.3. There were 26,941 instances in which students displayed this suboptimal performance; this sample includes data 1,009 unique students and from 245 problems. Specific data sets for this analysis are also posted on OSF (link excluded for review). Table 1 presents the demographic data for both original and analytic samples.

Table 1: Student Demographics

	Original Sample **(*n* = 1,430)**	**Analytic Sample** **(n = 1,009)**
Gender		
Female	47.06% (673)	47.08% (476)
Male	52.94% (757)	52.82 (533)
Race/Ethnicity		
White	48.09% (678)	51.54% (520)
Asian	28.01% (395)	23.29% (235)
Hispanic	15.18% (214)	17.34% (175)
Black/African American	4.89% (69)	3.87% (39)
American Indian	0.64% (9)	0.69% (7)
Two or More Races	3.19% (45)	2.58% (26)
Missing Race/Ethnicity Data	0.01% (20)	0.69% (7)
Individual Education Plan (IEP)	8.74% (125)	11.79% (119)
English Speakers of Other Languages	9.67% (138)	11.30% (114)

4.1.1 Measures

Three measures are necessary for the regression discontinuity design (RDD; described in Section 4.2). RDDs require an outcome variable, a treatment indicator variable, and a running variable is used to determine the treatment. The outcome indicates whether the student displays productive persistence by replaying the problem. The treatment is the game-based failure of receiving one clover described in Section 3.3. The treatment indicator (*Z*) was coded such that Z_{ij} = 1 when students experienced the game-based failure of receiving one clover and Z_{ij} = 0 when students received two clovers. The running variable (*R*) is the steps over optimal students take on their first attempt at the problem. Z_{ij} is determined by the students taking more than two steps over optimal; therefore, the cut point, *c,* is two. We centered *R* around the cutoff *c*, such that the cut point is zero, one step above the cut point is .5, one point below is -.5, and so on.

We used two covariates to evaluate effect heterogeneity when addressing RQ2 and RQ3. The number of optimal steps for each problem measures the complexity of the problem. The greater the number of optimal steps, the more complex the problem. To assess students' prior knowledge, we used an algebraic knowledge pretest comprising ten multiple-choice items from a previously validated measure of algebraic understanding of equivalence (Star et al., 2015; α = .89). Table 2 presents descriptive statistics for these variables.

Table 2: Descriptive statistics of moderator variables

Variable	Mean	σ^2	Min	Max	Skew	Kurtosis
Optimal Steps	3.78	2.20	1	12	1.08	0.47
Prior Algebraic Knowledge	5.28	2.69	0	10	0.11	-1.07

In order to ensure the robustness of the RDD, we ran placebo tests - described in detail in Section 4.2.1 - which require placebo outcome variables. For these, we used four behaviors measured on the problems prior to the problem included in the RDD. These measures are used as placebo outcomes because the game-based failure on one problem cannot possibly have a causal effect on students' behaviors on previous problems. These variables included whether the student accessed the hint problem on the prior problem, attempted to execute a step that constituted a mathematical error on the prior problem, achieved the optimal solution on a prior problem, and replayed the previous problem.

4.2 Research Design

Since our research questions concern whether game-based failure causes students to replay problems, we employ a causal method of analysis: regression discontinuity design (RDD). Following the Rubin Causal Model (Rubin, 1974), in the case of binary treatment ($Z = 1$; $Z = 0$), let $Y_i(1)$ be the potential outcome associated with individual *i* were they to undergo the treatment condition (Z_i=1), and $\mathrm{Y_i}(0)$ be the potential outcome for individual *i* were they to undergo control condition (Z_i=0). Let the treatment effect for individual *i* be $\tau_i = Y_i(1) - Y_i(0)$. The fundamental problem of causal inference is that only one potential outcome is observed ($Y(1)_i$ or $Y(0)_i$). Yet, when the potential outcomes of Y_i are independent of Z_i, as is the case in a completely randomized controlled trial, the difference in the average effects across treatment and control provides an estimate of the treatment effect as shown in Equation 1.

$$\hat{\tau} = \sum \frac{Y_i(Z_i)}{\sum Z_i} - \frac{Y_i(1 - Z_i)}{\sum 1 - Z_i} \tag{1}$$

In RDD, treatment assignment is not randomized. Instead, it is determined by the value of a numeric "running variable" of *R* relative to a threshold, *c* (Imbens & Lemieux, 2008). In our current study, R_{ij} represents the steps over optimal student *i* employed when working on problem *j*, and Z_{ij} is completely determined by whether $R_{ij} > c=2$. Standard RDD methods (e.g. Imbens and Kalyanaraman, 2012, Calonico, et al. 2014) use linear regression models in an attempt to estimate the local average treatment effect, the limit of effect of treatment on the outcome as *R* approaches the cutpoint *c* from either side. However, since it relies on linear regression and limits, the standard approach is often applied when both the outcome and running variable are continuous.

In contrast, the outcome variable in the current study is binary (whether or not the students replay the problem) and the running variable is discrete (students' steps over optimal); therefore, we employ the method established by Sales & Hansen (2020) of LRD, which is more amenable to non-continuous variables. The discrete running variable poses challenges to standard RDD analysis because the LATE refers to a limit that does not exist, since *R* cannot truly approach the cutpoint. LRD does not require a continuous running variable because the effect estimate is averaged across all units within the bandwidth (Limitless ATE), not just locally at the cutpoint. Furthermore, binary outcomes present difficulties for the standard approach because methods based on linear models forgo the substantial efficiency gains from models that are designed for binary outcomes, such as logistic regression (Cox and Snell, 1989), while coefficients from logistic regressions may not interpretable as average causal effects (Freedman, 2008). LRD can accommodate logistic regression modeling without relying on the logistic regression coefficients for causal effects. Finally, LRD has the added benefit of using bounded-influence regression estimators (Maronna et al. 2019); these models are more robust than the typical local linear models to slight model misspecification or overly optimistic bandwidth choices.

In an RDD, the running variable *R* may confound the relationship between treatment *Z* and the potential outcomes for *Y*. On the other hand, because *R* completely determines *Z*, it is the only possible confounder (in the framework of Pearl, 2009 *R* blocks all backdoor paths between *Z* and *Y*). Sales and Hansen (2020) argue that variation in *Y(0)* could be decomposed into a confounded component driven by *R*, and to an unconfounded component that is independent of *R*. To estimate the Limitless ATE, researchers may model the relationship between *Y(0)* and *R*, and then subtract fitted values from *Y*–a process called "detrending," leaving residuals, which will be unconfounded. Formally, they posit residual ignorability: let $e_\theta(y_i|r_i)$ be the residuals of a function predicting Y_i using R_i, referred to as the detrending procedure. Residual ignorability states that if R_i is within a specified window of analysis *W*–typically, within a bandwidth *b* of *c*–the residuals produced when predicting the potential outcomeY(0)$_i$ are mean-independent of Z,

$$E\left[e_\theta\left(Y(0), R\right)|R \in W, Z\right] = E\left[e_\theta\left(Y(0), R\right)|R \in W\right]. \quad (2)$$

For RQ1 we used LRD to estimate the Limitless ATE, which requires four steps: (1) selecting the bandwidth for R_i for which the Limitless ATE can be applied; (2) fitting of the function used in the detrending procedure; (3) estimating the effect of treatment; (4) robustness checks and post-fit diagnostics. The full procedure is outlined in Sales and Hansen (2020) and summarized below. To address RQ2 and RQ3, we run separate Limitless RDDs for each quartile of the variables of interest – problem complexity and prior knowledge – and evaluate the differences in effect sizes between the quartiles.

For the detrending procedure, we use the robust MM-estimation through the *glmrob* function in R (Maechler et al., 2023). To estimate the treatment effect, we use the *lmitt* function in the *propertee* package. The R script for the entire analysis is posted on GitHub (link excluded for review).

4.2.1 Bandwidth Selection

To determine what bandwidth of *R* to include in the analysis, we estimated the effect of the treatment on the placebo outcomes described in Section 4.1.1. Since placebo outcomes were drawn from prior problems, they could not be affected by *Z*, so a significant effect estimate would indicate a violation of residual ignorability (Equation 2), possibly due to an overly wide choice of *b*. To estimate these effects, we follow the following procedures for fitting a detrending model (Section 4.2.2) and effects models (Section 4.2.3) for each placebo outcome at each possible bandwidth, moving sequentially from the broadest bandwidth to the narrowest. We selected the broadest bandwidth for which the treatment did not have a significant estimated effect on the placebo outcomes for all broader bandwidths.

4.2.2 Robust Fitter for the Detrending Procure

The detrending procedure requires a model predicting the likelihood that students, if in the control condition, would replay problems $Y(0)_i$ as a function of the running variable (R_{ij}). Since we only have access to observed outcomes *Y* and not potential outcomes *Y(0)*, we include *Z* in the model. Because each cut-point decision is made at the problem level, we include fixed effects for each problem (π_j), which were one-hot encoded excluding one problem.

$$logit(Y_{ij}) = \gamma_0 + \gamma_1 Z_{ij} + \gamma_2 R_{ij} + \sum \gamma_j \pi_j \tag{3}$$

The partial residuals from Model 4, $e_\theta(Y_{ij}|R_{ij}) = Y_{ij} - logit^{-1}(\hat{\gamma}_0 + \hat{\gamma}_2 R_{ij} + \hat{\gamma}_j)$, where $logit^{-1}(\cdot)$ is the inverse logit function, are used to estimate the main effect of the game-based failure (Z_{ij} = 1) on the probability that students will replay the problem (Y_{ij}) as described in the next section.

4.2.3 Treatment Effect Estimation

Finally, we estimate the differences in the residuals of the control group using a linear regression

$$e_\theta(Y|R) = \eta_0 + \tau Z_{ij} \tag{5}$$

Let $e_\theta(Y|R)$ be the function producing the residuals of the control potential outcomes. Let η_0 be the intercept and τ be the treatment effect of Z_{ij}. Thus, τ is the effect of game-based failure on students' persistence behavior, which addresses RQ1. The estimate of τ from model (5) is equal to the difference in means estimate (1), with residuals $e_\theta(Y|R)$ replacing raw outcomes *Y*. We refer to this treatment effect as the Limitless average treatment effect (Limitless ATE).

4.2.4 Robustness Checks and Post-fitting Diagnostics

As a robustness check, we estimated the main treatment effect across a range of bandwidths to ensure that the reported effect is not simply a product of the selected bandwidth. As further robustness checks, we then estimate the determining and treatment models using ML estimators and including random intercepts for problems and students. This ensures that the estimated effects are not simply a product of our chosen estimator. Finally, we evaluated plots of the residuals and the expected values of the detrending model, which allowed us to assess model misspecification.

5. Results

5.1 Bandwidth Selection

To select the appropriate bandwidth for a limitless average treatment effect, we estimated models for four placebo covariates at each possible bandwidth. The p-values for the treatment effect coefficient of each model are presented in Figure 3. There are no significant placebo effects when the $|R_{ij}| < 7$. The effective bandwidth used in the estimation is $R_{ij} \subset [-1.5, 6.5]$ or one step over optimal through seven steps over optimal. We only include problems with in which at least one student replayed and at least one did not, observations with both outcomes, dropping 28 problems from a total of 251 to a sample of 223 problems. These restrictions retain the overwhelming majority (99.54%) of the data for the original sample (*n* =26,566) as the sample was heavily skewed towards lower steps over optimal. All of the students in the analytic sample described in Table 1 are represented in the post-bandwidth-restricted data.

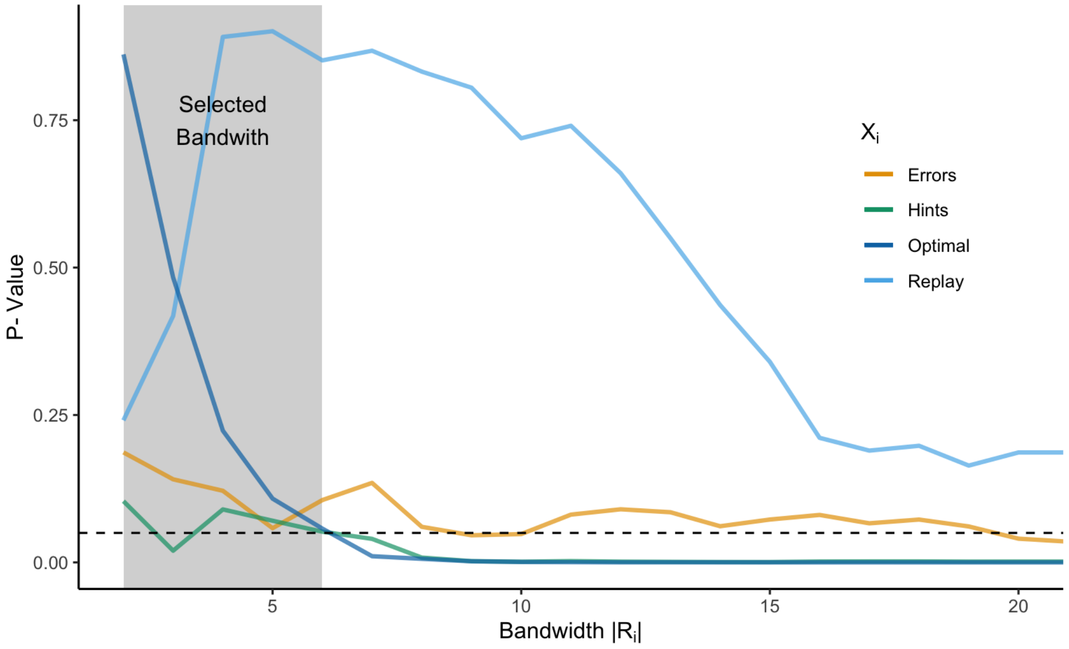


Figure 3: P-Values of Placebo Effect at Different Bandwidths

5.2 Fitter for Detrending Procedure

Table 3 presents the detrending models predicting the outcome, as described in Section 4.2.2. Figure 4 displays the regression discontinuity scatterplot produced by the trending model for RQ1. The coefficient for *Z* indicates the difference in log odds that students will replay a problem when experiencing game-based failure. Steps over optimal (R) is negatively associated with replay (γ_2 = -0.28, SE = 0.02, p < 0.001); thus, the further the students are away from optimal performance, the less likely they are to replay the problem. However, at cut point *c,* where game-based failure (Z = 1), the likelihood that students will persist by replaying the problem shifts discontinuously (γ_1 = 0.38, SE = 0.07, p < 0.001). This will be explored further in the section on causal effects estimates (Section 5.3).

Table 3: Limitless RDD Models

	ML-Estimator		
	Detrending Model		
	Log Odds	**SE**	***p***
Intercept	-.77***	.162	< 0.001
Z	0.38***	0.07	< 0.001
R	-0.28***	0.02	< 0.001
σ Random Intercepts			
Problems			
Students			
	Limitless ATE		
	Estimate	**SE**	***p***
τ	0.06***	0.02	< 0.001

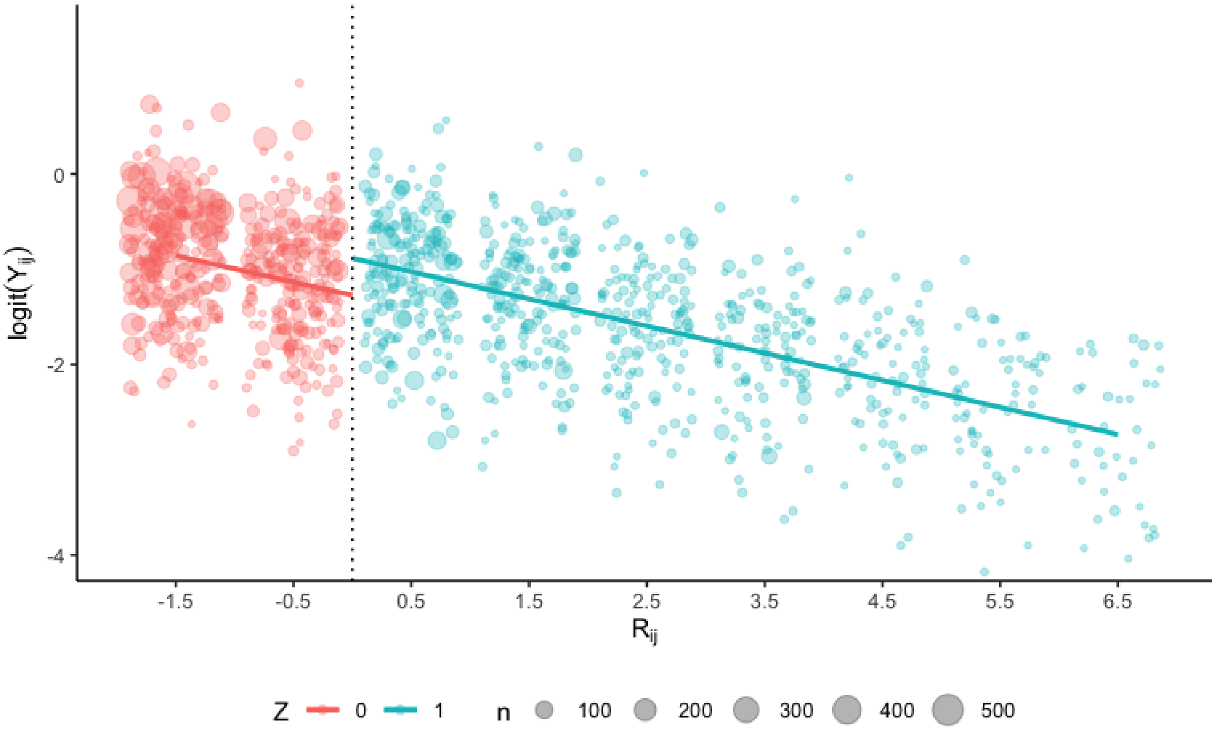


Figure 4: Predicted Outcomes from Detrending Model for RQ1

Jitter plot of predicted probabilities of replay persistence behavior by steps over optimal centered at the treatment cut point. Each point represents average log odds of replay for students within a problem at the particular R-value.

5.3 Limitless ATE Estimation

The Limitless ATE of game-based failure is a 0.06 increase in the probability of replaying (τ = 0.06, SE = 0.02, $p < 0.001$). This addresses RQ1, providing evidence that game-based failure positively impacts the likelihood that students will persist by replaying the problem.

To address RQ2 and RQ3, we reran the Limitless RDD procedure for each quartile of both problem complexity and students' prior knowledge. The limitless ATEs produced by these procedures are presented in Figure 5. (Tables of the estimates and statistical tests of the comparison are included in Section 9.1 of the Appendix). Overall, these analyses suggest the possibility but do not provide definitive proof of effect heterogeneity.

There is some evidence that students are less or not responsive to game-based failure on the most complex problems (Q4 of problem complexity; τ_{Q4} = 0.03, SE = 0.03, *p* = 0.23), but this evidence is inconclusive as there isn't a significant difference between the effects for Q4 and the less difficult problems (Q1, Q2, and Q3). Similarly, we have some evidence that students with low prior knowledge (Q1) may not experience an effect, as the point estimate is close to zero and the confidence intervals span zero (τ_{Q1} = 0.02, SE = 0.03, *p* = 0.65). Furthermore, the estimate for Q1 was significantly lower than the Q3 estimate (δ_{τ} = .11, SE = 0.04, *p* = 0.04), but no other comparisons had significant differences after correcting for multiple comparisons using the Holm procedure (Holm, 1979). Finally, the estimate for high-knowledge (Q4) students is not significant ($\tau_{Q4} =$ 0.07, SE = 0.04, *p* = 0.06). This may be an issue of a smaller sample of problems for this group because the number of instances in which high-knowledge students had suboptimal performance was lower than other students, causing a lower statistical power.

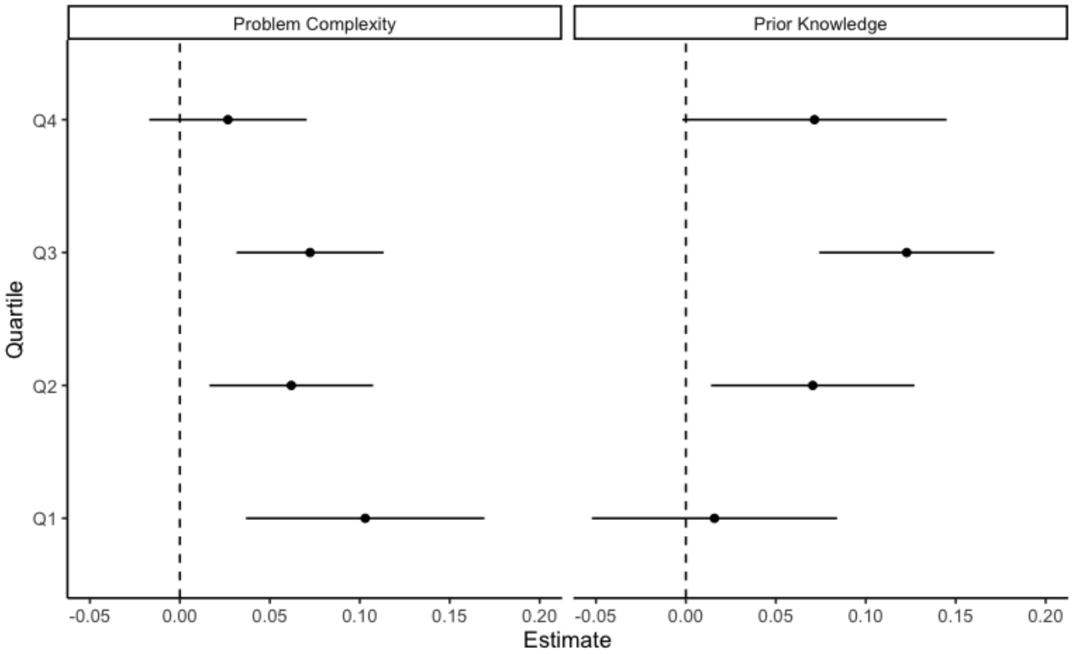


***Figure 5:* Effect Heterogeneity Across Problem Complexity and Prior Knowledge**

5.4 Robustness Checks

After estimating the Limitless ATE using the models presented above, we evaluate the robustness of these effects to ensure that they were not simply spurious artifacts of specific modeling decisions. We assessed this robustness in two ways: by estimating the effects at different bandwidths and by using different modeling techniques (i.e. ML-estimators and Multi-Level Models). All models showed a consistent positive effect of the game-based failure on the probability that students would replay problems. The effect estimates across bandwidths are discussed in this section. The other robustness check models are presented in the Appendix, along with model fit checks.

Figure 6 displays both the Limitless ATE calculated using data from multiple bandwidths. Although the effects decrease as the bandwidth increases, they remain significant across the bandwidths. Furthermore, the decrease in the magnitude of the effects stabilizes as the bandwidth increases such that the effect reported above is not significantly different from any of the smaller effects at larger bandwidths. Thus, the statistical significance of the effect estimate reported above is not a product of a specific bandwidth, and the magnitude would not significantly increase or decrease if a different bandwidth had been selected.

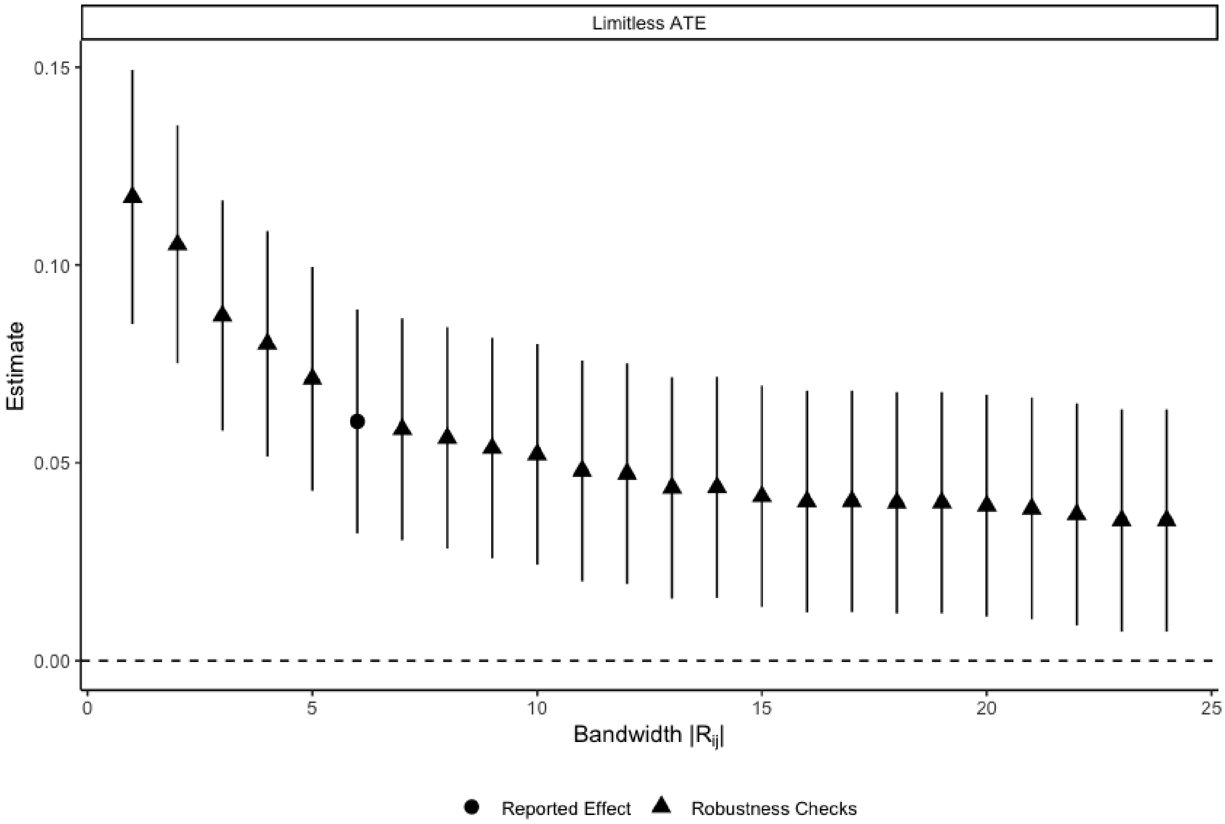


Figure 6: Limitless Average Treatment Effects across multiple bandwidths.

6. Discussion

6.1 Game-Based Failure and Productive Persistence

Overall, the experience of game-based failure had a positive effect on the likelihood that students would engage in productive persistence. According to Lander's *Theory of Gamified Learning* (2014), gamification impacts student outcomes by affecting their behaviors. This finding provides insight into one mechanism by which gamification influences students' behaviors. The results suggest that failure, which is generally demotivating, can positively affect student behavior when framed in the game context.

Notably, the only change the students experienced in the program was that they received a smaller nominal reward (the image of one clover as opposed to two) and that the words "good job" were crossed out (see Figure 2), thus communicating to the student's performance was inadequate. Generally, if we explicitly communicate to the students' performance that their performance is poor, we can expect students to be less motivated to continue and show lower persistence (Fong et al., 2019). Consider that students often stopout – quit a mastery learning task early – after getting a problem wrong (Botelho et al., 2019). Furthermore, low performance on high-stakes tests is also associated with decreased motivation (Amrein & Berliner, 2003; Roderick & Engel, 2001). Yet, in the context of a game, communicating that the student failed to do a good job had a positive impact on their persistence behavior, suggesting that students were motivated by the failure to engage in productive persistence.

This finding supports the theory that game-based failure can be "pleasantly frustrating" and, therefore, less demotivating than getting a problem wrong in another educational context (Gee, 2005). Though the goals of a traditional assignment and a game-based learning activity are ostensibly the same – perform well by presenting the correct or best solution – in the game context, feedback indicating suboptimal performance is likely perceived differently than just being provided with an indication that a response was incorrect. Rather than point towards the inadequacies of the students, in-game feedback points the students towards the "nuances of the game" (Juul, 2008), which in this case happen to coincide with the nuances of algebraic knowledge and understanding of equivalency. Gamified communication of failure makes the math problem feel like an activity to be won, as opposed to a perfunctory task that must be completed.

6.2 Learning from Game-Based Failure

As we discussed previously (Section 3.2), research on replaying problems within FH2T is associated with improved performance within the program (Liu et al., under review), greater growth in algebraic knowledge (Lee et al., 2022), and enhanced impact of the gamified program

(Vanacore et al., 2023)**.** Although the average effect of game-based failure on productive persistence was small in magnitude – a 0.06 increase in the probability that students would replay the problem – that does not mean the effect was insubstantial. Economies of scale are important here. Within the study itself, which only consisted of one school district, the difference in probability is associated with over a thousand problems replayed that would likely not have been reattempted otherwise. Considering the potential to scale the program far beyond the study, there is a large prospective impact on students' learning behaviors.

It is also notable that the program's efficacy in improving students' algebraic knowledge is associated with marginal differences in students' replay behavior. In a previous study, we found that students' propensity to replay problems moderated the impact of the gamified condition on students' algebraic knowledge, such that those with a higher propensity to replay experienced a greater positive effect from the program (Vanacore et al., 2023). Substantial differences in the effects were associated with small differences in the students' propensities to replay problems. Students with a low propensity to replay, with a probability of replaying the average problem of 0.08, experienced about half the effect of students with a high propensity to replay of 0.20. Furthermore, the main effect for students with a low propensity to replay was not discernibly different from zero. (The analysis was Bayesian, and the credible intervals spanned zero.) Thus, the small increases in students' propensity to replay problems caused by game-based failure may be crucial to the program's effectiveness.

The results also have implications for how we understand failure in the context of educational games. The connection between game-based failure and replaying problems after suboptimal performance is important because it allows access to the learning benefits of failure that have been theorized and studied previously (Kapur, 2008; Piaget, 1977; VanLehn, 1988). Yet, there are a couple of peculiarities in the failure experiences explored in this paper that should be delineated. There is a key distinction between the gamified experience of failure studied in this paper and impasse-driven learning. In impasse-driven learning, which is often relevant in mastery learning scenarios, failure induces a state of cognitive disequilibrium that prohibits progress (VanLehn, 1988). Alternatively, the game-based failure experience in our study allows the student to progress, but encourages them to learn from their mistakes. Students do not experience an impasse in their learning, although they may still need to work through a state of cognitive disequilibrium to achieve the optimal solution. Yet, the game-based failure is more subtle, which may contribute to its motivational nature.

Productive failure provides another perspective on failure and learning. Studies related to productive failure suggest that when students are induced to fail before instruction they experience learning benefits compared with students not primed by failure experiences (Kapur, 2008, 2012; Kapur & Bielaczyc, 2011). Although the students’ experience in the gamified learning program does not mirror Kapur and colleagues’ experiments, which often occurred in small groups with teachers guiding students through understanding their failures, the mechanisms may be similar. The system indicates that the student's struggle to produce an answer was suboptimal, and the student has to learn by reattempting the problem. Typically, students need instructors to guide the experience of productive failure as managing students' “frustration thresholds” can be difficult (Kapur & Bielaczyc, 2012). Alternatively, our finding that game-based failure increases productive persistence suggests that game-based failure may remove frustrating elements from the failure experience, allowing the students to benefit from, rather than be discouraged by, their failures.

6.2 Leveraging Regression Discontinuity Designs for Research in Computer in Education

The secondary goal was to illustrate the utility of RDD in the context of human-computer interaction research on educational programs. As noted above, the prevalence of decisions made based on cut points in computer-based learning platforms provides opportunities for causal insights into the impact of programmatic features on students' learning behaviors and outcomes. We have demonstrated one example of this, which shows the impact of game-based failure on student behavior and, more broadly, on student learning in digital environments. RDD can be leveraged to understand the impact of administering rewards (Liu et al., 2022), prescription of usage recommendations (Dieter et al., 2020), and to determine whether students are allowed to progress in mastery learning activities (Bloom, 1968; Kelly et al., 2015; Yudelson et al., 2013), as well as other decisions automatically made within computer-based learning platforms. This is especially useful when experimental designs, typically employed to estimate causal effects, are impractical or infeasible. Furthermore, we have demonstrated how limitless regression discontinuity design can allow for flexibility even when outcomes and running variables of the RDD are not continuous.

7. Limitations and Future Directions

There are limitations to this work that should be addressed in future research. First, the gamified condition has some singularities that make the generalizability of the findings contingent upon replication in other contexts. In most learning environments, reattempting a problem after completing it with an acceptable response would be impractical, especially in math problems with one correct solution. Therefore, the productive persistence outcome in our study is peculiar to the gamified condition here. Future work should consider how game-based failures may affect other manifestations of productive persistence. Yet, it is notable that the game-based failure experience of receiving a low-performance feedback reward and indicating that the student performance had suboptimal performance, while not commonly indicated by a crossed-out "good job," is typical in other gamified computer-based learning platforms. Future work should test the impact of this or similar forms of game-based failure on productive persistence in other environments.

The RDD in this study included two features, each of which complicates the usual account of RDDs: a binary outcome and a discrete running variable. Fortunately, a suite of alternative estimation methods and bandwidth choices all lead to estimates that are broadly consistent with our overall conclusions. If, instead, the results had (counterfactually) depended heavily on a particular model or bandwidth choice, the appropriate conclusions would have been much less clear. In other words, more methodological research is necessary to fully understand the strengths, weaknesses, and operating characteristics of our analytical approach, especially under circumstances as challenging as ours.

Conclusion

Overall, this research suggests that gamification may be used to mitigate the demotivating aspects of failure that can prevent students from experiencing the benefits of failure. In this case, something that may be perceived as punitive – an indication that students did not do a "good job" – is counterintuitively motivating to students regardless of the complexity of the problem or their prior abilities. This research adds to an increasingly nuanced understanding of how students respond to feedback and how gamification may influence the relationship between failure and persistence.

8. References

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The Future of Embodied Design for Mathematics Teaching and Learning. *Frontiers in Education*, *5*. https://www.frontiersin.org/articles/10.3389/feduc.2020.00147

Adjei, S. A., Baker, R. S., & Bahel, V. (2021). Seven-year longitudinal implications of wheel spinning and productive persistence. *22nd International Conference, AIED 2021*, 16–28. https://doi.org/10.1007/978-3-030-78292-4_2

Amrein, A. L., & Berliner, D. C. (2003). The Effects of High-Stakes Testing on Student Motivation and Learning. *Educational Leadership*, *60*(5), 32–38.

Baker, R. S. J. d., Corbett, A. T., Roll, I., Koedinger, K. R., Aleven, V., Cocea, M., Hershkovitz, A., de Caravalho, A. M. J. B., Mitrovic, A., & Mathews, M. (2013). Modeling and Studying Gaming the System with Educational Data Mining. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 97–115). Springer. https://doi.org/10.1007/978-1-4419-5546-3_7

Beck, J. E., & Gong, Y. (2013). Wheel-Spinning: Students Who Fail to Master a Skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 431–440). Springer. https://doi.org/10.1007/978-3-642-39112-5_44

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. https://doi.org/10.1016/j.jarmac.2020.09.003

Bjork, Robert A. (2014). Forgetting as a friend of learning. In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger (Eds.), *Remembering: Attributions, Processes, and Control in Human Memory* (pp. 15–28). Psychology Press.

Bloom, B. S. (1968). Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Evaluation Comment*, *1*(2). https://eric.ed.gov/?id=ED053419

Botelho, A. F., Van Inwegen, E. G., Varatharaj, A., & Heffernan, N. T. (2019). Refusing to try: Characterizing early stopout on student assignments. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, 391–400. https://doi.org/10.1145/3303772.3303806

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression‐discontinuity designs. Econometrica, 82(6), 2295-2326.

Chan, J. Y.-C., Lee, J.-E., Mason, C. A., Sawrey, K., & Ottmar, E. (2022). From Here to There! A dynamic algebraic notation system improves understanding of equivalence in middle-school students. *Journal of Educational Psychology*, *114*(1), 56. https://doi.org/10.1037/edu0000596

Clark, D. B., Nelson, B. C., Chang, H.-Y., Martinez-Garza, M., Slack, K., & D'Angelo, C. M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers & Education*, *57*(3), 2178–2195. https://doi.org/10.1016/j.compedu.2011.05.007

Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development*, *66*(5), 1101–1118. https://doi.org/10.1007/s11423-018-9579-9

Decker-Woodrow, L. E., Mason, C. A., Lee, J.-E., Chan, J. Y.-C., Sales, A., Liu, A., & Tu, S. (2023). The Impacts of Three Educational Technologies on Algebraic Understanding in the Context of COVID-19. *AERA Open*, *9*, 23328584231165919. https://doi.org/10.1177/23328584231165919

Dieter, K. C., Studwell, J., & Vanacore, K. P. (2020). Differential Responses to Personalized Learning Recommendations Revealed by Event-Related Analysis. *International Conference on Educational Data Mining (EDM)*, *13*. https://eric.ed.gov/?id=ED607826

Fong, C. J., Patall, E. A., Vasquez, A. C., & Stautberg, S. (2019). A Meta-Analysis of Negative Feedback on Intrinsic Motivation. *Educational Psychology Review*, *31*(1), 121–162. https://doi.org/10.1007/s10648-018-9446-6

Freedman, D. A. (2008). Randomization Does Not Justify Logistic Regression. *Statistical Science*, *23*(2). https://doi.org/10.1214/08-STS262

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, Motivation, and Learning: A Research and Practice Model. *Simulation & Gaming*, *33*(4), 441–467. https://doi.org/10.1177/1046878102238607

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative Knowledge: Understanding Professional Learning and Expertise. *Vocations and Learning*, *1*(2), 87–103. https://doi.org/10.1007/s12186-008-9006-1

Gaston, J., & Cooper, S. (2017). To Three or not to Three: Improving Human Computation Game Onboarding with a Three-Star System. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5034–5039. https://doi.org/10.1145/3025453.3025997

Gee, J. P. (2005). Learning by Design: Good Video Games as Learning Machines. *E-Learning and Digital Media*, *2*(1). https://doi.org/10.2304/elea.2005.2.1.5

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The Education of Perception. *Topics in Cognitive Science*, *2*(2), 265–284. https://doi.org/10.1111/j.1756-8765.2009.01055.x

Hartmann, C., van Gog, T., & Rummel, N. (2021). Preparatory effects of problem solving versus studying examples prior to instruction. *Instructional Science*, *49*(1), 1–21. https://doi.org/10.1007/s11251-020-09528-z

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hulse, T., Daigle, M., Manzo, D., Braith, L., Harrison, A., & Ottmar, E. (2019). From here to there! Elementary: A game-based approach to developing number sense and early algebraic understanding. *Educational Technology Research and Development*, *67*(2), 423–441. https://doi.org/10.1007/s11423-019-09653-8

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933-959.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Iseli, M., Feng, T., Chung, G., Ruan, Z., Shochet, J., & Strachman, A. (2021, July 26). *Using Visualizations* *of Students' Coding Processes to Detect Patterns Related to Computational Thinking*. 2021 ASEE Virtual Annual Conference Content Access. https://peer.asee.org/using-visualizations-of-students-coding-processes-to-detect-patterns-related-to-computational-thinking

Janelli, M., & Lipnevich, A. A. (2021). Effects of pre-tests and feedback on performance outcomes and persistence in Massive Open Online Courses. *Computers & Education*, *161*, 104076. https://doi.org/10.1016/j.compedu.2020.104076

Jones, B. D. (2007). The Unintended Outcomes of High-Stakes Testing. *Journal of Applied School Psychology*, *23*(2), 65–86. https://doi.org/10.1300/J370v23n02_05

Jones, B. D., & Egley, R. J. (2004). Voices from the Frontlines: Teachers' Perceptions of High-Stakes Testing. *Education Policy Analysis Archives*, *12*(39). https://eric.ed.gov/?id=EJ853506

Juul, G. J. (2008). Fear of Failing? The Many Meanings of Difficulty in Video. In *The Video Game Theory Reader 2*. Routledge.

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision Tree Modeling of Wheel-Spinning and Productive Persistence in Skill Builders. *Journal of Educational Data Mining*, *10*(1), Article 1. https://doi.org/10.5281/zenodo.3344810

Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, *26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*(4), 651–672. https://doi.org/10.1007/s11251-012-9209-6

Kapur, M., & Bielaczyc, K. (2011). Classroom-based Experiments in Productive Failure. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.

Kapur, M., & Bielaczyc, K. (2012). Designing for Productive Failure. *Journal of the Learning Sciences*, *21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Kelly, K., Wang, Y., Tamisha, T., & Neil, H. (2015). Defining Mastery: Knowledge Tracing Versus NConsecutive Correct Responses. *In the Proceedings of the 8th International Conference on*

Educational Data Mining. Educational Data Mining.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26). https://doi.org/10.1073/pnas.1921417117/-/DCSupplemental

Klinger, D. A., & Luce-Kapler, R. (2008). Walking in Their Shoes: Students' Perceptions of Large-Scale High-Stakes Testing. *Canadian Journal of Program Evaluation*, *22*(3), 29–52. https://doi.org/10.3138/cjpe.0022.004

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, *120*(13), e2221311120.

Landers, R. N. (2014). Developing a Theory of Gamified Learning: Linking Serious Games and Gamification of Learning. *Simulation & Gaming*, *45*(6), 752–768. https://doi.org/10.1177/1046878114563660

Lavoué, É., Ju, Q., Hallifax, S., & Serna, A. (2021). Analyzing the relationships between learners' motivation and observable engaged behaviors in a gamified learning environment. *International Journal of Human-Computer Studies*, *154*, 102670. https://doi.org/10.1016/j.ijhcs.2021.102670

Lee, D. S., & Lemieux, T. (2009). *REGRESSION DISCONTINUITY DESIGNS IN ECONOMICS*. NATIONAL BUREAU OF ECONOMIC RESEARCH.

Lee, J.-E., Chan, J. Y.-C., Botelho, A., & Ottmar, E. (2022). Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational Technology Research and Development*, *70*(5), 1575–1599. https://doi.org/10.1007/s11423-022-10138-4

Li, X., Ye, Y., Li, M. J., & Ng, M. K. (2010). On cluster tree for nested and multi-density data clustering. *Pattern Recognition*, *43*(9), 3130–3143. https://doi.org/10.1016/j.patcog.2010.03.020

Liu, A., Vanacore, K., & Ottmar, E. (2022). *Reward-based feedback systems create micro-failures that support persistence-related learning behaviors [Manuscript Under Review]*.

Liu, Z., Cody, C., & Barnes, T. (2017). The Antecedents of and Associations with Elective Replay in an

Educational Game: Is Replay Worth It? *Proceedings of the 10th International Conference on Educational Data Mining*. Educational Data Mining.

Loibl, K., & Leuders, T. (2019). How to make failure productive: Fostering learning from errors through elaboration prompts. *Learning and Instruction*, *62*, 1–10. https://doi.org/10.1016/j.learninstruc.2019.03.002

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a Theory of When and How Problem Solving Followed by Instruction Supports Learning. *Educational Psychology Review*, *29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Loibl, K., & Rummel, N. (2014). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science*, *42*(3), 305–326. https://doi.org/10.1007/s11251-013-9282-5

Ludwig, J., & Miller, D. L. (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly Journal of Economics*, *122*(1), 159–208. https://doi.org/10.1162/qjec.122.1.159

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & di Palma, M. A. (2023). *robustbase: Basic Robust Statistics* (0.95-1) [Computer software]. https://cran.r-project.org/web/packages/robustbase/index.html

Malkiewich, L. J., Lee, A., Slater, S., Xing, C., & Chase, C. C. (2016). *No Lives Left: How Common Game Features Could Undermine Persistence, Challenge-Seeking and Learning to Program*. https://repository.isls.org//handle/1/115

McClelland, D. C. (1961). *Achieving Society*. Simon and Schuster.

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*(4), 352–370. https://doi.org/10.1037/h0069819

McKernan, B., Martey, R. M., Stromer-Galley, J., Kenski, K., Clegg, B. A., Folkestad, J. E., Rhodes, M. G., Shaw, A., Saulnier, E. T., & Strzalkowski, T. (2015). We don’t need no stinkin’ badges: The impact of reward features and feeling rewarded in educational games. *Computers in Human Behavior*, *45*, 299–306. https://doi.org/10.1016/j.chb.2014.12.028

Ottmar, E., Lee, J.-E., Vanacore, K., Pradhan, S., Decker-Woodrow, L., & Mason, C. A. (2023). Data from the Efficacy Study of From Here to There! A Dynamic Technology for Improving Algebraic Understanding. *Journal of Open Psychology Data*, *11*(1), 5. https://doi.org/10.5334/jopd.87

Ottmar, E. R., Landy, D., & Weitnauer, E. (2015). Getting from here to there: Testing the effectiveness of an interactive mathematics intervention embedding perceptual learning. *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society.*, 1793–1798.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-SS057

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures. (Trans A. Rosin)* (pp. viii, 213). Viking.

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., & Kafai, Y. (2009). Scratch: Programming for all. *Communications of the ACM*, *52*(11), 60–67. https://doi.org/10.1145/1592761.1592779

Roderick, M., & Engel, M. (2001). The Grasshopper and the Ant: Motivational Responses of Low-Achieving Students to High-Stakes Testing. *Educational Evaluation and Policy Analysis*, *23*(3), 197–227. https://doi.org/10.3102/01623737023003197

Schank, R. C. (1999). *Dynamic Memory Revisited*. Cambridge University Press. https://vdoc.mx/documents/dynamic-memory-revisited-28oe04llmni0

Tawfik, A. A., Rong, H., & Choi, I. (2015). Failing to learn: Towards a unified design approach for failure-based learning. *Educational Technology Research and Development*, *63*(6), 975–994. https://doi.org/10.1007/s11423-015-9399-0

Vanacore, K. P., Gurung, A., McReynolds, A. A., Liu, A., Shaw, S. T., & Heffernan, N. T. (2023). Impact of Non-Cognitive Interventions on Student Learning Behaviors and Outcomes: An analysis of seven large-scale experimental inventions. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 10. https://doi.org/10.1145/3576050.3576073

Vanacore, K., Sales, A., Liu, A., & Ottmar, E. (2023). Benefit of Gamification for Persistent Learners: Propensity to Replay Problems Moderates Algebra-Game Effectiveness. *Tenth ACM Conference on Learning @ Scale (L@S '23)*. Learning @ Scale, Copenhagen, Denmark. https://doi.org/10.1145/3573051.3593395

VanLehn, K. (1988). Toward a Theory of Impasse-Driven Learning. In H. Mandl & A. Lesgold (Eds.), *Learning Issues for Intelligent Tutoring Systems* (pp. 19–41). Springer US. https://doi.org/10.1007/978-1-4684-6350-7_2

Williams-Pierce, C. (2019). Designing for mathematical play: Failure and feedback. *Information and Learning Sciences*, *120*(9/10), 589–610. https://doi.org/10.1108/ILS-03-2019-0027

Woo, K., & Falloon, G. (2022). Problem solved, but how? An exploratory study into students' problem solving processes in creative coding tasks. *Thinking Skills and Creativity*, *46*, 101193. https://doi.org/10.1016/j.tsc.2022.101193

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian Knowledge Tracing Models. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 171–180). Springer. https://doi.org/10.1007/978-3-642-39112-5_18

9. Appendix

9.1 Effect heterogeneity estimates (RQ2 & RQ3)

Table 4 presents the effects for each quartile of problem complexity and prior knowledge. Table 5 presents the estimated difference between the quartiles and statistical tests of those differences. The results are discussed in Section 5.3.

Table 4: Limitless ATE Estimates for RQ2 and RQ3

Quartile	**Estimate**	**Std. Error**	**P-value**
	Problem Complexity		
Q1	0.10	0.03	0.002
Q2	0.06	0.02	0.008
Q3	0.07	0.02	0.001
Q4	0.03	0.02	0.232
	Prior Knowledge		
Q1	0.02	0.03	0.648
Q2	0.07	0.03	0.015
Q3	0.12	0.02	0.000
Q4	0.07	0.04	0.057

Table 5: Comparisons of differences in effect sizes between

Comparison	**Estimated Difference**	**SE**	***p*-value**	**Adjusted *p*-values***
		Problem Complexity		
Q1_Q2	0.04	0.04	0.84	0.99
Q1_Q3	0.03	0.04	0.78	0.99
Q1_Q4	0.08	0.04	0.97	0.99
Q2_Q3	-0.01	0.03	0.37	0.99
Q2_Q4	0.04	0.03	0.86	0.99
Q3_Q4	0.05	0.03	0.93	0.99
		Prior Knowledge		
Q1_Q2	-0.06	0.05	0.11	0.45
Q1_Q3	**-0.11**	**0.04**	**0.01**	**0.04**
Q1_Q4	-0.06	0.05	0.14	0.83
Q2_Q3	-0.05	0.04	0.09	0.51
Q2_Q4	-0.00	0.05	0.49	0.99
Q3_Q4	0.05	0.05	0.87	0.99

*adjusted using the Holm–Bonferroni method

9.2 Model Fit Diagnostics

To evaluate potential model misspecification in the detrending and effect estimation models, we plotted residuals against the predicted values using binned residual plots presented in Figures 5 and 6. Neither plot revealed evidence of misspecification.

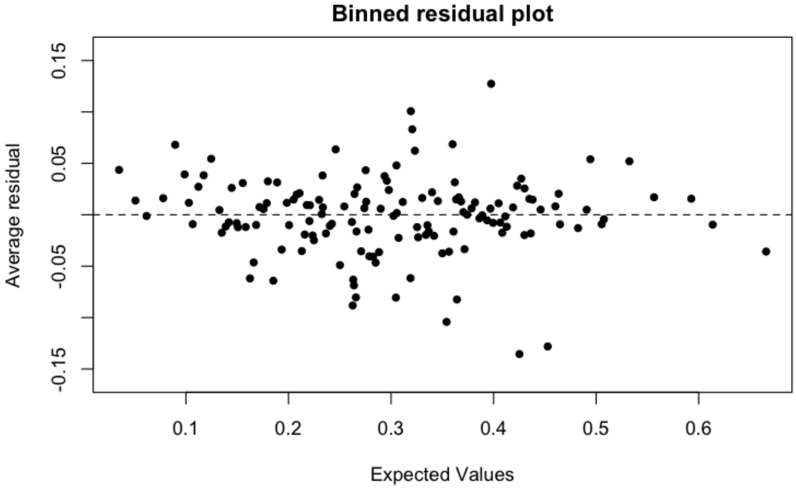


Figure 5: Binned Residual Plot for Detrending Model

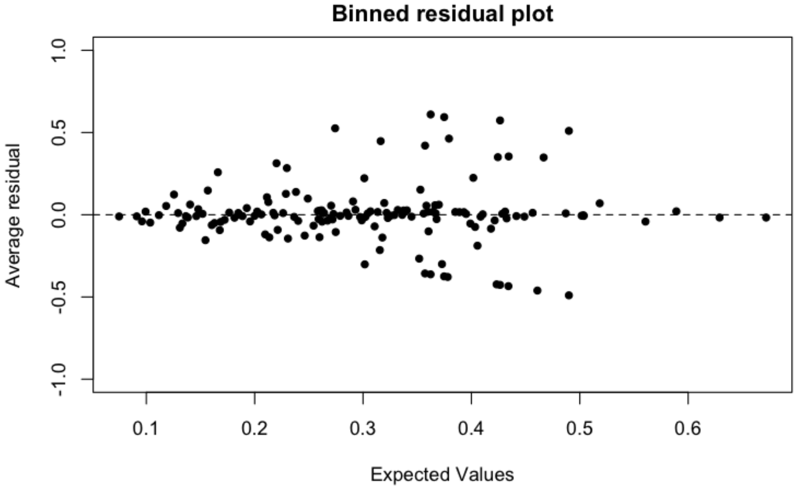


Figure 6: Binned ResidualPlot for Limitless ATE Estimation model

9.3 Robustness checks

Table 6 presents two alternative estimation methods for running models. Unlike the MM-estimator from our main analysis, the ML-estimator does not downweight outliers during the model estimation process. The MLM model, which includes random effects, was estimated to ensure that by the data's levels of nesting – students who attempted multiple problems and problems attempted by multiple students – did not confound the effects estimations. Estimates produced by both modeling methods show consistent findings with the estimates presented in Sections 5.2 and 5.3. This is further evidence of the robustness of the causal effect between game-based failure and replay behavior.

Table 6: Models for Effect Estimate Robustness Checks

	ML-Estimator			**MLM**		
	Detrending Model					
	Log Odds	**SE**	***p***	**Log Odds**	**SE**	***p***
Intercept	-0.73	0.044	< 0.001	-2.23	0.139	< 0.001
Z	0.33	0.063	< 0.001	0.60	0.087	< 0.001
R	-0.24	0.020	< 0.001	-1.00	0.027	< 0.001
σ Random Intercepts						
Problems				2.46		
Students				1.49		
	Limitless ATE					
	Estimate	**SE**	***p***			
τ	0.05	0.007	< 0.001	0.06	0.005	< 0.001

Chapter 4

The how Game-Based Failure can Productive Persistence

4.1 Benefit of Gamification for Persistent Learners**

See manuscript below.

Benefit of Gamification for Persistent Learners: Propensity to Replay Problems Moderates Algebra-Game Effectiveness

Kirk Vanacore
kpvanacore@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, United States of America

Adam Sales
asales@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, United States of America

Allison Liu
aliu2@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, United States of America

Erin Ottmar
eottmar@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, United States of America

ABSTRACT

Computer-assisted learning platforms (CALPS) increasingly include gamified elements to improve student outcomes by enhancing their engagement with content. Although evidence exists that gamified programs increase engagement and learning outcomes, there is little causal research on what programmatic mechanisms drive the effect between engagement and learning. In the following paper, we explore this relationship through a method of causal moderation known as fully latent principal stratification. Using data from a large-scale randomized control trial assessing gamified and traditional CALP systems' effects on algebraic knowledge, we estimate the impact of using the gamified CALP on students who engage with one of its key gamification elements—replaying a problem after a suboptimal attempt. The gamified CALP asks students to manipulate algebraic expressions from start to goal states and provides feedback based on the efficiency of these manipulations, allowing students to replay the problems when their efficiency can be improved. We find that the effect of gamification is greater for students with a higher propensity to replay problems. This finding suggests that gamification elements that provide students with opportunities to retry problems are driving the game's efficacy and provide evidence for a scalable mechanism of gamification that can improve students' learning.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**; • **Human-centered computing** → *Empirical studies in interaction design.*

KEYWORDS

Productive Persistence; Gamification; Computer-Assisted Learning Platforms; Causal Inference

ACM Reference Format:
Kirk Vanacore, Adam Sales, Allison Liu, and Erin Ottmar. 2023. Benefit of Gamification for Persistent Learners: Propensity to Replay Problems Moderates Algebra-Game Effectiveness. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3573051.3593395

L@S '23, July 20–22, 2023, Copenhagen, Denmark
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0025-5/23/07.
https://doi.org/10.1145/3573051.3593395

1 INTRODUCTION

Gamified computer-assisted learning programs (CALPs) can increase students' engagement and motivation as well as problem-solving abilities and learning outcomes [13, 19, 27, 52]. CALPs, such as ASSISTments [26] and MATHia [51], provide students with problems, instruction, and feedback through a digital interface. Some CALPs are gamified, defined as the infusion of "game design into non-game contexts," resulting in a playful user experience [8]. This process includes introducing game-like features — such as point systems, narrative elements, awards and badges, competitions, and features that provide a mix of successes and failures – into the learner's experience [30, 36]. In theory, gamification impacts learning by changing students' behaviors, mediating and moderating the relations between instructional content and outcomes [36]. Although many studies have evaluated the impact of gamification on students' learning and behaviors [6, 13, 33, 43, 49, 60], few trace a causal path from gamified programs through behaviors to learning outcomes [34].

In the current paper, we begin to address this gap by assessing one theoretical relation between gamification and learning through a learning related behavior. More specifically, we examine whether the improved learning outcomes associated with the algebra-focused learning game *From Here to There!* (FH2T) [45] depends on students' likelihood to persist when confronted with a gamed-based failure. We do this by using a novel technique of fully latent principal stratification (FLPS) [39] to evaluate the impact of gamification on students who exhibit a game-related behavior — replaying problems to achieve optimal performance — on their algebraic knowledge. In doing so, we provide evidence for the causal link between gamification, behavior change, and learning outcomes while displaying the usefulness of FLPS in exploring causal relations using data from large-scale CALPs.

2 BACKGROUND

2.1 Theories of Gamification

Pedagogical systems are more effective when they require learners to participate, engage in problem-solving, and provide immediate feedback [9, 25, 57]. Gamified CALPs are particularly well suited to meet these needs, as they can engage learners in active problem-solving while providing automated and targeted feedback in playful, low-stakes environments [9, 13, 14].

There are many theories for how to gamify CALPs and how gamified elements can impact learning effectively. One review of 32 papers related to learning-based games identified 118 theories used to explain the connection between gamification and learning [34]. This review found that most of the theories did not focus on the connection between gamification and learning; instead, they were more general theories applied to learning games (e.g., self-determination theory, flow theory, experiential learning theory, etc.). Alternatively, Landers [36] proposed that gamification indirectly influences learning outcomes. Specifically, gamification is thought to affect learning-related behaviors or attitudes, and these behaviors/attitudes influence learning by 1) moderating the relationship between instructional design quality and academic outcomes and 2) mediating learning outcomes directly. Notably, Landers's theory is not exclusive: it provides a framework through which other theories can be used to influence learners' behaviors and attitudes, which subsequently impact learning.

2.2 Impact of Gamification

Numerous studies have shown that gamification can positively influence students. These positive effects generally fall into three categories: learning achievement, motivation and engagement, and interaction and social connection [40, 65]. For example, Jaguvst and colleagues[30] found that digital math lessons led to increased or sustained performance over time in primary school students when those lessons included game elements, compared to a non-gamified version which produced decreased performance over time. Furthermore, different combinations of individual game elements also had differential effects on mathematics performance, with the largest performance increase shown for a gamified condition that included competition, narratives, and adaptive difficulty. Similarly, Wijers and colleagues [64] found that adolescent students reported high levels of motivation and interest when playing a mobile mathematics game and believed they had learned the game's material well.

However, gamification's effects are not always consistent. For example, adolescent students showed increased performance after using mobile mathematics game, but no change in their intrinsic motivation [24], and undergraduate students enrolled in an online learning course with badges showed decreased intrinsic motivation over time [35]. Indeed, different game elements or combinations of elements can result in significantly different effects [53]. Further, the same game elements can also have different effects depending on the student, even when these students are in the same class [16].

Because of these complex interactions, most studies fail to provide theoretical explanations on how gamification relates to learning, engagement, and motivation [65]. Thus, questions remain regarding when and how gamification can successfully impact academic outcomes.

2.3 Gamification and Persistence Behaviors

Gamification may influence students' behavior by encouraging them to persist by re-attempting problems when they have not achieved optimal performance [21, 38, 40]. Persistence is considered an aspect of conscientiousness in which a person is driven to complete challenging tasks [42]. Persisting after experiencing difficulty allows students to work in the upper ends of their zones of proximal development, where most learning occurs [62, 63]. However, not all types of persistence are equal. Researchers distinguish between productive persistence, in which students adjust their strategies and behaviors in response to failure to better meet their learning goals [32], versus unproductive persistence, in which students spend time struggling through failure without adjusting and achieving learning mastery [5, 7]. Productive persistence has been associated with better long-term academic outcomes [2], whereas unproductive persistence is associated with poorer learning outcomes [2, 7].

Generally, persisting through failure appears to be easier in game contexts [22], suggesting that gamified CALPs can effectively support students during challenging learning activities. Indeed, gamification can have positive effects on student persistence [21, 44]. For example, O'Rourke and colleagues [44] implemented a point system in a fraction game that increased children's playtime and persistence through game levels compared to children who played the same game without points. Gaston and Cooper [21] similarly implemented a reward system in a game about protein structures and found that players were more likely to use fewer moves, take more time per move, and replay levels more often. However, as with gamification's effects on learning outcomes, effects on persistence are also inconsistent across gamification elements, as the inclusion of some game elements can instead hinder persistence behaviors [11, 41].

Gamified CALPs can provide opportunities to continuously and easily engage in persistence behaviors, which is not always possible in traditional instructional tasks. For example, some gamified CALPs allow students to immediately reset or retry problems an unlimited number of times to fix errors, try alternative solution paths, and/or improve their performance [40]. Several studies illustrate the benefit of using these persistence opportunities in gamified CALPs. In one online math game, a small group of students who spent a large proportion of game time replaying completed problems showed the largest learning gains in algebraic understanding [38]. Lavoue and colleagues [37] found two patterns of student engagement behaviors within a digital math learning environment: achievement-oriented engagement, in which students tried to attain the best possible performance on their first problem attempts; and perfection-oriented engagement, in which students replayed problems to try to improve their performance. The former engagement pattern led to a decrease in intrinsic motivation, whereas the latter replay-focused engagement pattern led to increases in extrinsic and identified motivation (a form of self-determined motivation in which one's actions are driven by one's attitudes, values, and needs). This evidence suggests that students who take advantage of gamified opportunities for productive persistence may show better academic outcomes compared to those who avoid these opportunities.

3 CURRENT STUDY

Although there is evidence that gamification can impact learning and influence persistence behaviors, there is a lack of empirical evidence of a causal link between the behaviors and the outcomes, as theorized by Landers [36]. The current study seeks to fill this gap by conducting a causal moderation analysis on data from an efficacy study assessing the impact of FH2T [45], which consists of a game-based educational technology focused on improving algebraic knowledge. The study compares this gamified condition (e.g FH2T) to active control of traditional algebra problems administered through ASSISTments [26].

Data for our analysis were generated during a larger efficacy study comparing FH2T to three other conditions (results reported in [15] and the full study data can be found in [46]). In the current study, we focus on two conditions: the gamified FH2T condition and the active control. The study was conducted in ten schools from a district in the Southern US during the 2021-22 school year. Randomization occurred at the student level, so teachers had students of different conditions in the same class. The study consisted of nine half-hour sessions conducted fortnightly throughout the school year.

The FH2T is meant to improve students' understanding of algebra using perceptual learning [23] and embodied cognition [1]. The game consists of worlds (problem sets) that teach different mathematical concepts (from Addition to Inverse Operations), and each world contains 18 problems for a total of 252 problems. All students started at World 1 (Addition) and worked through progressive worlds in a set order. Each problem has a unique starting expression and goal state. Several studies have demonstrated FH2T's effectiveness at improving elementary and middle school students' mathematical understanding of equivalence [10, 15, 28].

3.1 Problems in Gamified and Active Control Condition

The two conditions represent different approaches to teaching algebra. The gamified condition asks students to transform a starting algebraic expression into a mathematically-equivalent goal state. An example problem is shown in Figure 1. Students reach the goal state by manipulating expressions by tapping or dragging numerals or variables using a mouse or touch screen. The system responds by providing a fluid visualization showing how these actions transform the expressions. Students learn new gesture actions through video demonstrations as they progress through the game, and they must use and combine these learned gesture actions to successfully transform expressions from one state to another as the game progresses. For each problem, students can make an infinite number of expression transformations, but there is always an optimal number of steps.

Alternatively, in active control, students complete the problem sets by submitting answers through multiple choice or filling in answers. The condition included 218 problems, divided into nine problem sets corresponding to the study's nine sessions. An example of the active control problem can be found in Figure 2. Students in this condition do not receive feedback until after completing each problem set.

3.2 Replaying Problems in Gamfied condition

In the gamified condition, students are encouraged to replay problems when they do complete them in the optimal number of steps. Unlike many traditional math problems where there is typically only one correct answer, replaying in the gamified condition allows students to explore multiple possible solutions. Previous studies have shown that students who replay problems in the gamified CALP become more efficient in achieving those solutions and are more likely to reach the optimal solution on the first attempt of the next problem [29]. In the gamified condition, students receive feedback for an ordinal "reward" represented by clovers (see Figure 3). They can receive between one and three clovers depending on their efficiency: three clovers if the students took the optimal number of steps to reach the goal state, two clovers if they were within two steps of optimal, and one clover for all other completed attempts. When students receive three clovers, they are automatically moved to the next problem. However, when students earn one or two clovers, they are prompted to either replay the problem or move on to the subsequent problem. Previous analyses suggest that the number of clovers received has a causal effect on whether students replay the problem [61]. These findings suggest that students are responding to the gamified nature of the program by displaying persistent learning behaviors when taking the opportunity to replay problems to improve their skills and understanding.

Replaying the problem is a product of students' own volition; there were no external ramifications of receiving fewer than three clovers as teachers did not have access to student performance data. Therefore the ability to replay in the gamified condition is one of the key differentiators of the CALP, as traditional problem sets are not optimized for students to learn from multiple attempts of the same problem.

3.3 Research Question

As explained above, gamification theoretically impacts learning in part because it creates opportunities for students to exhibit beneficial learning behaviors, such as persisting, which enhances the effect of the instruction. In the current study, we hypothesize that the opportunity to replay in the gamified condition allows students to change behavior, which improves learning. If this is true, then the students who replay in the gamified condition should experience a larger effect than those who do not replay. We seek to test this hypothesis by answering the question: Does the impact of the gamified condition vary based on whether students engage with the replay element gamification?

4 METHOD

4.1 Fully Latent Principal Stratification

The fundamental question of this paper—whether replaying the problem moderates the impact of the gamified program on the student's algebraic knowledge— poses an interesting methodological complication as students in the control condition did not have the opportunity to replay problems. One way to approach this complication is to view students' replay behavior as a function of a latent characteristic: every student's propensity for replaying *if*

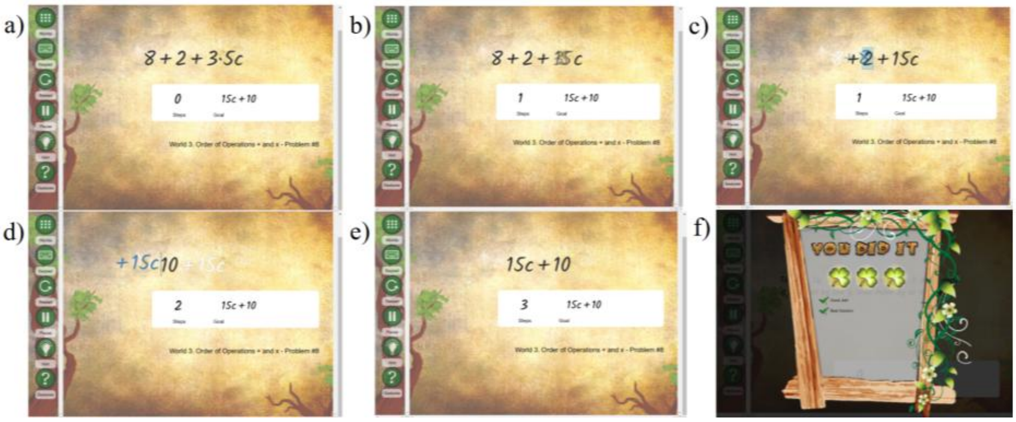
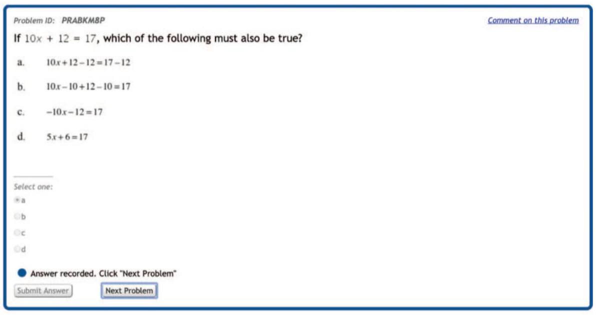


Figure 1: Gameified Condition Example Problem

A sample problem in the gamified condition. From the initial expression (a), the student taps on 3 5c to automatically multiply the 3 and 5 together (b). The student then drags 8 on top of the 2 to automatically add 8 and 2 together (c). Finally, the student drags the 15c to the other side of the 10 (d) to reorder the terms and reach the goal state (e). Finally, the student earned the maximum three clovers for completing the problem in the minimum number of steps (f).



Problem ID: PRABKM8P

Comment on this problem

If 10x + 12 = 17, which of the following must also be true?

a. $10x+12-12=17-12$

b. $10x-10+12-10=17$

c. $-10x-12=17$

d. $5x+6=17$

Select one:

● a

○ b

○ c

○ d

● Answer recorded. Click "Next Problem"

Submit Answer Next Problem

Figure 2: Control Example Problem

A sample problem from the control condition. The students select a response and then submit it. They only receive feedback after each problem set is completed.

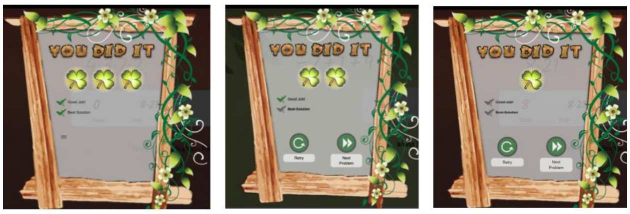


Figure 3: Replay encouraging feedback

The clover reward screen that is shown after each problem (with one, two, or three clovers shown depending on the number of steps students use to solve the problem). Students are given the option to either retry the problem again (a replay attempt) or to move on to the next problem if they earn one or two clovers, and are automatically progressed to the next problem if they earn three clovers.

they had been given the opportunity (i.e. if assigned to the treatment condition). This latent characteristic is present at baseline (though not observed) and defined for the entire sample, so it can be considered a baseline covariate or potential moderator, similar to students' pretest knowledge. Therefore, it is independent of the random treatment assignment. However, unlike pretest knowledge, only students in the treatment condition were given the opportunity to display the behavior; therefore, its value in the control condition is unknown. To estimate the treatment effect for the students with a high propensity to replay, we use fully latent principal stratification (FLPS) [39, 56].

Principal stratification is a causal inference method used in randomized control trials for estimating the intervention effects on subgroups determined after the treatment has begun [20, 48]. Traditional estimation of effects for subgroups requires that they be defined before intervention and be independent of any treatment. For instance, in the case of pretest knowledge, simply interacting the treatment with the pretest knowledge score provides information about how the treatment effect varies across subgroups of students with similar prior knowledge. However, subgroups defined based on students' potential program implementation cannot be observed at baseline, and are never observed for students randomized to the control condition. When program implementation consists of a complex sequence of users' behaviors or choices, and is defined based on, say, a cluster analysis or a latent variable model, then it is not directly observed for students randomized to the treatment condition either, but must be estimated. This is particularly true in CALPs, where students may display behaviors during the program such as meeting usage targets [17], "gaming the system" [4], or mastering knowledge components in mastery learning [54]. These behaviors can all be viewed as indicators of membership in subgroups (i.e., high fidelity users, gamers, mastery users) that are unknown for the control groups who do not have the opportunity to use, game, or master as they did not interact with the CALP. For instance, Sales & Pane [54] demonstrated the use of principal stratification with a latent variable to determine whether the effect of Cognitive Tutor Algebra I on students varies based on whether students were likely to master knowledge components.

An alternative to FLPS is to use the control group only and create an observational study comparing students who replay to those who do not replay within the gamified condition. Yet, to use this method for estimating causal effects, we must assume strong ignorability — that there are no omitted confounders that would bias the results. This may be a dubious assumption as students may be more or less likely to replay problems due to various unobserved factors (e.g., familiarity with technology, motivation, goal orientation, time management, etc.). When comparing observational methods to principal stratification, Sales & Pane [55] note that the observational methods rely heavily on the strong ignorability assumption, and they suggest using principal stratification when this assumption is not met. The disadvantage of principal stratification is that it focuses on the mechanism indirectly, in that the effect estimate is for students who are likely to use the mechanism (replay), not the effect of the mechanism (replay) itself. This will be discussed further in the Limitations section.

Let τ_i be subject i's individual treatment effect, i.e., the difference between what i's posttest score would be if i were randomized to treatment and what it would be were they randomized to control. Let $\mathcal{T}$ and C be the samples of students randomized to treatment and control, respectively. Let α_{ti} represent student i's propensity to replay a problem if randomized to the treatment condition (the t subscript indicates that this variable refers to potential behavior if assigned to treatment). Although α_{ti} exists for C, as students in the control still have an unspecified propensity to replay problems even if they do not have the opportunity, α_{ti} can only be estimated from replay behavior for $\mathcal{T}$. The principal effect is the treatment effect for the subgroup of students with a particular value for α_t:

$$\tau(a) = E[\tau|\alpha_t = a] \tag{1}$$

To estimate the function $\tau(a)$, we (1) estimate α_t for $\mathcal{T}$, (2) model α_t as a function of covariates observed in both groups, (3) use that model to impute α_t for C, (4) estimate $\tau(a)$ by including a treatment interaction in a linear regression model. In practice, we iterate through these steps many times in a Bayesian principal stratification model with a continuous variable, consisting of measurement and outcome submodels, as outlined by [31] and [47].

4.1.1 Modeling Replay Behavior. First, we estimate α_t by running a multilevel logistic submodel predicting whether students in the treatment (gamified) condition replay on each problem they attempt as delimited in the equation 2. Let R_{ji} be a binary indicator of whether student i replayed problem j. Let P_{ki} be covariate predictor k of K student-level predictors, which are measured at baseline for both $\mathcal{T}$ and C (described in section 4.2.3). Let E_j be covariate predictor j of J problems. Let the random intercepts be μ_i for students, μ_t for teachers, and μ_s schools, each modeled as independent and normal with mean 0 and a standard deviation estimated from the data.

$$logit(R_{ji}) = \gamma_0 + \sum_{k=1}^{K} \gamma_k \mathrm{P}_{ki} + \sum_{j=1}^{J} \delta_j \mathrm{E}_j + \mu_i + \mu_t + \mu_s \tag{2}$$

Using the parameters from equation 2, students' propensity to replay is defined as

$$\alpha_{ti} = \sum_{k=1}^{K} \gamma_k \mathrm{P}_{ki} + \mu_i + \mu_t + \mu_s \tag{3}$$

We impute α_{ti} for C with random draws from a normal distribution with mean $\sum_k \gamma_k \mathrm{P}_{ki} + \mu_{t[i]} + \mu_{s[i]}$, where $\mu_{t[i]}$ and $\mu_{s[i]}$ are the random intercepts for student i's teacher and school, respectively, and standard deviation equal to the the estimated standard deviation of μ_i. Note that randomization occurred at the student level (i.e. teachers had students in the treatment and control in their classes). Therefore we are able to include the random intercepts for schools and teachers from submodel 2 in valuation for α_{ti} for students in the control. However, μ_i is unknown for C, but we can assume its distribution is the same in the two conditions because of the randomization.

4.1.2 Modeling Posttests $\tau(a)$. To estimate the treatment effect for students with varying propensities to replay, we run a multilevel linear regression predicting student's post-test algebraic knowledge (Y_i). The submodel includes interaction between α_{ti}—estimated for $\mathcal{T}$ and randomly imputed for C—and Z_i, an indicator of being in the treatment.

$$Y_i = \beta_0 + \beta_1 \mathrm{Z}_i + + \beta_2 \alpha_{ti} + \beta_3 \alpha_{ti} \mathrm{Z}_i + \sum_{k=1}^{K} \lambda_k \mathrm{P}_{ki} + \nu_t + \nu_s \tag{4}$$

Using the parameters from the submodel 4, the treatment effect for students with a particular propensity replay is modeled as

$$\tau(a) = \beta_1 + \beta_3 a \tag{5}$$

Submodels (2) and (4) together formed a Bayesian FLPS model, which we fit using the Stan Markov Chain Monte Carlo software [3].

4.2 Data and Variables

The data for this study exists at two different levels. There are student-level variables, including the learning outcome, demographics, pretest math anxiety, and pretest state test scores. Replay behavior is a problem-level variable, which exists for each problem that students in the gamified condition attempt. All data from the study are available through OSF and a full explanation of the data can be found at [46].

The sample consists of 1209 students: 824 in the gamified condition and 385 in the active control (sampling was intentionally weighted towards the gamified condition). The students where taught by 42 teachers, in 138 classes. In the treatment condition, students completed 88,949 problems. Missing data were imputed using using singly-impution with the Random Forest routine implemented by the missForest package in R [50, 59].

4.2.1 Learning Outcome. The learning outcome for the study is students' algebraic knowledge, which was assessed using ten multiple-choice items from a previously validated measure of algebra understanding (Cronbach's α = .89; see the items on OSF[1]) [58]. Within the ten items, four focused on conceptual understanding of algebraic equation-solving (e.g., the meaning of an equal sign), three focused on procedural skills of equation-solving (e.g., solving for a variable), and three focused on flexibility of equation-solving strategies (e.g., evaluating different equation-solving strategies). These 10 items together assessed a range of students' knowledge in algebraic equation-solving, which is the ultimate goal of each of the education technologies tested. The assessment was taken before and after the intervention.

4.2.2 Replay Behavior. Students in the gamified condition had the opportunity to replay the problem as described in Section 3.2. As the replay count per problem is highly skewed (skew = 7.62), we treat it as a binary outcome: whether or not the student replayed a problem at least once. Out of the 88,949 problems which students played, students replayed 12.85% of the time. Students only replayed problems more than once 3.18% of the time.

4.2.3 Predictors. Students' demographic data—race/ethnicity, individualized education plan status (IEP), and English as a second or foreign language (ESOL) status—were provided by the students' school district along with their most recent standardized state test scores in math. Race ethnicity was dummy coded, with white students as the reference category. Three pretest scores were collected by the original studies' researchers: prior algebraic knowledge, math anxiety, and perceptual processing skills. Pretest algebraic knowledge was a variant of the learning outcome. The math anxiety assessment was adapted form from the Math Anxiety Scale for Young Children-Revised [12], which assessed negative reactions towards math, numerical inconfidence, and math-related worrying (Cronbach's α=.87; see the items on OSF[1]). The perceptual processing assessment evaluates to students' ability to detect mathematically equivalent and nonequivalent expressions as quickly as possible (see the items on OSF[1]) [18]. Log forms of assessment times of the pretest alphabetic knowledge were also included in the models. Polynomials and splines of the pretests were included when they improved model fit.

5 RESULTS

The results include two models, the measurement model (2), which predicts replay behavior, and the outcome model (4), which predicts algebraic knowledge. The FLPS models were run using Bayesian estimation through Markov chain Monte Carlo chains calling Stan through rstan [3] in R [50]; code is posted GitHub[2]. Convergence was evaluated using trace plots and $\hat{R}$. The maximum $\hat{R}$ for estimated parameters was 1.04.

5.1 Measurement Submodel of Replay Behavior

To ensure that the measurement submodel would provide accurate predictions of replay behavior, we first estimated the submodel in R using the glmer function from the lme4 package. This allowed us to select the transformations of variables that produced the most accurate models. The predictors used in the FLPS measurement submodel produced a model with an AUC of 0.85 with a sensitivity of 0.79 and a specificity of 0.74. This model was then estimated in Stan and presented in Table 1. The table provides the mean estimates form the posterior distributions as well as standard errors (SE) and credible intervals (CI). Notably, only two of the coefficients had credible intervals that did not span zero. Students were more likely to replay when they had scores pretest scores of 6.5 out of 10 or above. Students with higher math anxiety were also more likely to replay problems.

The propensity to replay any given problem in log odds—α_{ti}—was estimated using the model parameters for each student, and the distribution of α_{ti} is displayed in Figure 4. α_t has a mean of -0.53 and a standard deviation of 1.20. Converted to probability, on average, students have a 0.30 probability of replaying the average problem (SD = 0.21). Figure 4 also shows the distribution of the fixed effects for the problems (γ_e), which were not displayed in the Table 1.

5.2 Outcomes Submodel

The outcomes submodel is also displayed in Table 1. The credible intervals for the main treatment effect—which is the effect for students for whom α_{ti} is estimated to be zero, suggesting that there may be no treatment effect for students who had log odds of replaying any given problem of zero. By sampling from the posterior distribution, we calculate the probability that the main treatment effect is greater than zero is 0.83. Although the credible intervals for the α_{ti} coefficient also spanned zero, the probability that the coefficient is less than zero is 0.96. This suggests that, for students in the control condition, a higher likelihood of replaying is associated with lower post-test algebraic knowledge scores. This finding is expected as replaying is associated with having sub-optimal performance on the first attempt. Therefore a high likelihood of replaying without the opportunity to replay should be associated with low performance.

The creditable intervals for interaction between α_{ti} and the treatment did not span zero, and the probability that the coefficient was greater than zero was 0.98. This suggests that the treatment effect is greater for students with a higher propensity to replay problems. Figure 5 displays the interaction using a random sample of α_{ti} drawn from the posterior distribution.

Using the outcomes submodel parameters, we estimate the treatment effects for students with low and high propensity to replay. The estimated effect of the treatment for students with $\alpha_{ti} = 0$ is an increase of 0.639 points on the algebraic knowledge post-test, whereas the estimated effect for students with $\alpha_{ti} = 1$ (those with a probability of replaying a problem on the average problem of 0.20) is an increase of 1.27 points on the algebraic knowledge post-test. Notably, only 10% of the students in the sample have such a high propensity for replaying. The estimated effect for students in the upper quartile of replay behavior ($\alpha_{ti} \geq 0.23$; i.e., a probability of replay on the average problem ≥ of .10) is at least 0.78 points. Conversely, the effect for students with $\alpha_{ti} = -1$ (those with a probability of replaying a problem on the average problem of 0.03) was

[1] https://osf.io/r3nf2/
2

[2] https://github.com/kirkvanacore/FH2T_FLPS_replay

Table 1: Parameters from the measurement and outcome models

	Measurement Submodel			**Outcomes Submodel**		
Predictors	*Estimate*	*SE*	*CI*	*Estimate*	*SE*	*CI*
Intercept	-0.544	1.271	-3.26 - 1.9	3.839	0.667	2.46 - 5.11
Gamified Condition (Z_i)				0.639	0.696	-0.71 - 2.08
Propensity to Replay (α_{ti})				-0.564	0.311	-1.13 - 0.05
Z_i X α_{ti}				**0.632**	0.310	0.02 - 1.19
Virtual Instruction	-0.09	0.122	-0.33 - 0.15	0.095	0.267	-0.48 - 0.62
Hispanic	-0.02	0.066	-0.15 - 0.11	**0.163**	0.074	0.02 - 0.31
Asian & Pacific Islander	-0.015	0.079	-0.17 - 0.14	**0.359**	0.085	0.19 - 0.52
Black	0.07	0.057	-0.04 - 0.18	0.04	0.061	-0.08 - 0.16
Individualized education Plan	-0.009	0.059	-0.13 - 0.11	**0.128**	0.062	0.01 - 0.25
English as a Second Language	-0.072	0.069	-0.21 - 0.07	-0.024	0.074	-0.17 - 0.12
Pretest Algebraic Knowledge (Spline, x < 6.5, 1)	0.097	0.069	-0.04 - 0.23	**0.428**	0.075	0.28 - 0.57
Pretest Algebraic Knowledge (Spline, x > 6.5, 2)	**0.218**	0.078	0.06 - 0.37	**0.668**	0.091	0.49 - 0.84
Pretest Time on Task (Log)	0.025	0.057	-0.09 - 0.14	0.002	0.062	-0.12 - 0.12
Pretest Math Anxiety	**0.133**	0.054	0.03 - 0.24	0.014	0.061	-0.11 - 0.13
Pretest Math Anxiety (Squared)	-0.05	0.072	-0.19 - 0.09	0.120	0.072	-0.02 - 0.26
Perceptual Processing	-0.074	0.072	-0.22 - 0.07	0.011	0.075	-0.14 - 0.16
Perceptual Processing Response Time (Log)	0.115	0.080	-0.04 - 0.27	**0.353**	0.089	0.18 - 0.53
Prior Math State Test Scores	-0.055	0.085	-0.22 - 0.11	**0.847**	0.092	0.67 - 1.03
Missing Data Count	-0.033	0.069	-0.17 - 0.10	**-0.061**	0.067	-0.19 - 0.07
Random Effects						
σ^2	3.29					
$var(\mu_0)/var(v_0)$	1.465_i					
	0.182_t			0.445_t		
	0.147_s			0.453_s		
N	824_i					
	252_j					
	42_t			42_t		
	11_s			11_s		
Observations	88949			1209		

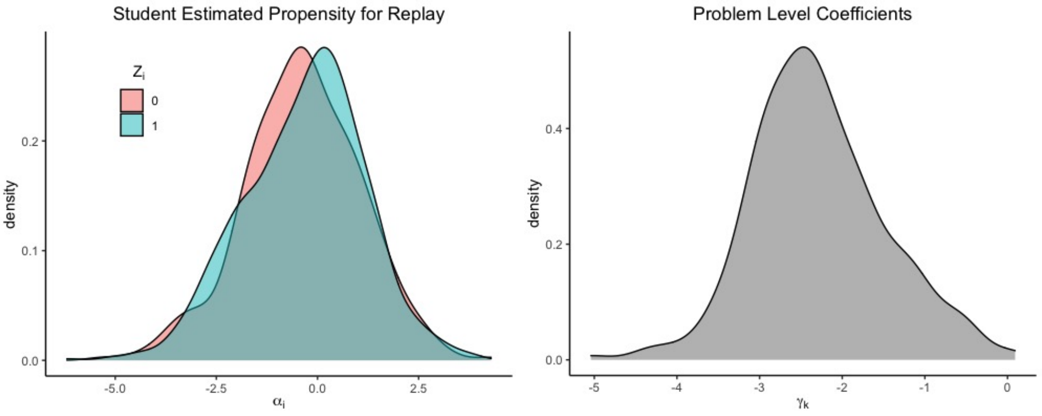


Figure 4: Distribution of the estimated parameters

α_{ti} are drawn form a random sample of posterior distribution. γ_e are the mean model parameters.

0.007. Furthermore, the effect on students in the bottom quartile of replay behavior($\alpha_{ti} \geq -1.26$; i.e., a probability of replay on the average problem ≥ of .02) was -0.16.

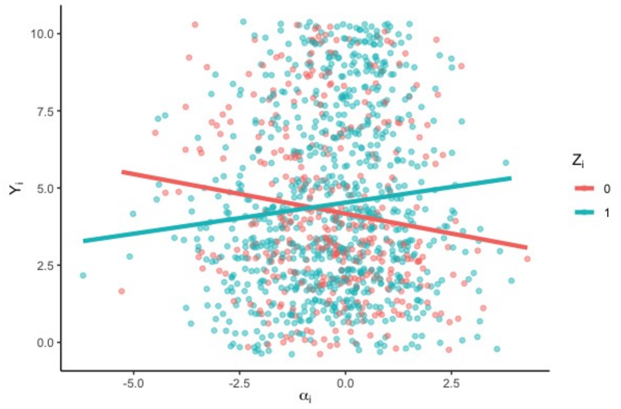


Figure 5: Interaction between α_i and Z_i from an random draw form the posterior distribution.

Notably, The credible intervals for students' race/ethnicity, whether they had an individualized education plan, their pretest algebra score, their response time on the perceptual processing task, their math state test scores, and the number of data points imputed all had coefficients that did not span zero. This suggests that these are all substantial predictors of students' post-test performance.

6 DISCUSSION

In this study, we find that students with a higher propensity for persisting by replaying problems when they achieve suboptimal performance benefit more from gamification than those with a lower propensity for persistence. Furthermore, it is possible that there is little to no effect for students with a low propensity for persistence. This suggests that the causal mechanism of gamification in impacting students learning is the opportunity that gamified CALP provides for replaying problems. Students who take advantage of this opportunity are more likely to benefit, and those who do not either experience a lower effect, or no effect at all. Regarding Lander's theory of gamified learning [36], the gamified condition influences students' persistence behavior by replaying problems, thus increasing the instruction's effectiveness.

This finding has two major implications. First, gamification's effects may depend on students' propensity towards certain behaviors. Effect heterogeneity has been documented in gamified CALPs [16]. What is notable about the present study findings is that we find this heterogeneity based on latent behavioral profiles that are only observed during the study. Past research has shown correlation evidence suggesting differences in program behaviors are related to differences in learning, such as the importance of delayed responses when answering questions [10, 38]. For non-gamified CALPs, there is also correlation evidence that in-program behaviors such as wheel spinning and productive presence are predictive of learning outcomes [2]. Our findings add causal evidence that the effects depend on how students use the program.

The second implication of this work is that students' utilization of a specific feature may be driving the overall effect of a program. This finding suggests that providing opportunities for students to display productive presence, like replaying problems, may increase CALPs' effectiveness. The ability to replay the same problems is in many ways peculiar to the gamified CALP in this study because of the nature of the problems, which focus on efficiency in manipulated equations as opposed to producing a correct response. Yet there are other instances in which students can display persistence, such as in mastery learning CALPs, which allow students to attempt similar problems until they have mastered a skill. Furthermore, increasing students' persistence behavior may also improve a program. Previous research has found that students are responsive to performance-based feedback, encouraging them to replay problems [40, 61]. This suggests that developers of CALPs may be able to influence students' persistence behaviors and, by proxy, enhance the program's efficacy.

7 LIMITATIONS AND FUTURE DIRECTIONS

Although our findings show that persistence behaviors moderate the relationship between the gamified program and learning, our work has some notable limitations. First, as mentioned above, the replay option is specific to the studied gamified CALP, and it may be difficult to implement such a feature in other programs. In principle, features in different gamified CALPs may provide options for persistence behaviors in other ways, yet future research must be conducted to confirm that the causal mechanism found here persists under different circumstances. Second, we focus solely on one behavior to be the causal mediator in this case, yet there are likely other behaviors that interact with the treatment effect (e.g. usage, wheel-spinning, gaming-the-system, response times). Future research should evaluate these potential mediators to provide a more robust understanding of how and for whom gamified CALPs are effective.

Furthermore, there are limitations to FLPS. Although FLPS allows for the estimation of heterogeneity in the causal effect associated with the latent characteristic of propensity to replay, we are not estimating the effect or replaying directly on the impact of the gamified condition [31]. It is only students who have a tendency to replay experience a greater effect. This tendency to replay may actually be associated with other student characteristics that may drive the heterogeneity of the effect. More work should be done to understand how the availability of a replay option impacts student learning to fully understand the relationship between this feature of the game and the game's overall effectiveness. Finally, effects estimated using FPLS are contingent on the underlining quality of the model itself, and it is unclear what error level in the model may bias the effect. More work must be done to understand how to evaluate these models to ensure they provide unbiased treatment effect estimates.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through an Efficacy and Replication Grant (R305A180401) and a Methodology Grant (R305D210036) to Worcester Polytechnic Institute. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

REFERENCES

[1] Dor Abrahamson, Mitchell J Nathan, Caro Williams-Pierce, Candace Walkington, Erin R Ottmar, Hortensia Soto, and Martha W Alibali. 2020. The future of embodied design for mathematics teaching and learning. *Frontiers in Education* 5 (2020), 147.

[2] Seth A Adjei, Ryan S Baker, and Vedant Bahel. 2021. Seven-year longitudinal implications of wheel spinning and productive persistence. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I.* Springer, Utrecht, The Netherlands, 16–28.

[3] Dorsa Mohammadi Arezooji. 2020. A Markov Chain Monte-Carlo Approach to Dose-Response Optimization Using Probabilistic Programming (RStan). (2020).

[4] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.

[5] Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. 2004. Detecting student misuse of intelligent tutoring systems. In *Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30-September 3, 2004. Proceedings 7.* Springer, Alagoas, Brazil, 531–540.

[6] Ivan L Beale, Pamela M Kato, Veronica M Marin-Bowling, Nicole Guthrie, and Steve W Cole. 2007. Improvement in cancer-related knowledge following use of a psychoeducational video game for adolescents and young adults with cancer. *Journal of Adolescent Health* 41, 3 (2007), 263–270.

[7] Joseph E Beck and Yue Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16.* Springer, Memphis, TN, USA, 431–440.

[8] L Benton, H Johnson, M Brosnan, E Ashwin, and B Grawemeyer. 2011. Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems.

[9] Elizabeth Boyle, Thomas M Connolly, and Thomas Hainey. 2011. The role of psychology in understanding the impact of computer games. *Entertainment computing* 2, 2 (2011), 69–74.

[10] Jenny Yun-Chen Chan, Ji-Eun Lee, Craig A Mason, Katharine Sawrey, and Erin Ottmar. 2022. From Here to There! A dynamic algebraic notation system improves understanding of equivalence in middle-school students. *Journal of Educational Psychology* 114, 1 (2022), 56.

[11] Catherine C Chase, Laura J Malkiewich, Alison Lee, Stefan Slater, Ahram Choi, and Chenmu Xing. 2021. Can typical game features have unintended consequences? A study of players' learning and reactions to challenge and failure in an educational programming game. *British Journal of Educational Technology* 52, 1 (2021), 57–74.

[12] Lian-Hwang Chiu and Loren L Henry. 1990. Development and validation of the Mathematics Anxiety Scale for Children. *Measurement and evaluation in counseling and development* 23, 3 (1990), 121–127.

[13] Thomas M Connolly, Elizabeth A Boyle, Ewan MacArthur, Thomas Hainey, and James M Boyle. 2012. A systematic literature review of empirical evidence on computer games and serious games. *Computers & education* 59, 2 (2012), 661–686.

[14] Sara De Freitas. 2006. Learning in immersive worlds. (2006), 3–71 pages.

[15] Lauren E Decker-Woodrow, Craig A Mason, Ji-Eun Lee, Jenny Yun-Chen Chan, Adam Sales, Allison Liu, and Shihfen Tu. 2023. The impacts of three educational technologies on algebraic understanding in the context of COVID-19. *AERA open* 9 (2023), 23328584231165919.

[16] Mouna Denden, Ahmed Tlili, Nian-Shing Chen, Mourad Abed, Mohamed Jemni, and Fathi Essalmi. 2022. The role of learners' characteristics in educational gamification systems: a systematic meta-review of the literature. *Interactive Learning Environments* (2022), 1–23.

[17] Kevin C Dieter, Jamie Studwell, and Kirk P Vanacore. 2020. Differential Responses to Personalized Learning Recommendations Revealed by Event-Related Analysis. *International Educational Data Mining Society.*

[18] Bye J. K.; Lee J. E.; Chan J. Y. C.; Closser A. H.; Shaw S. T.; Ottmar E. 2022. Toward Improving Effectiveness of Crowdsourced, On-Demand Assistance from Educators in Online Learning Platforms. In *Poster presented at the annual meeting of the American Educational Research Association (AERA).*

[19] Aroutis Foster. 2008. Games and motivation to learn science: Personal identity, applicability, relevance and meaningfulness. *Journal of interactive learning research* 19, 4 (2008), 597–614.

[20] Constantine E Frangakis and Donald B Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58, 1 (2002), 21–29.

[21] Jacqueline Gaston and Seth Cooper. 2017. To three or not to three: Improving human computation game onboarding with a three-star system. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems.* 5034–5039.

[22] James Paul Gee. 2003. What video games have to teach us about learning and literacy. *Computers in entertainment (CIE)* 1, 1 (2003), 20–20.

[23] Robert L Goldstone, Tyler Marghetis, Erik Weitnauer, Erin R Ottmar, and David Landy. 2017. Adapting perception, action, and technology for mathematical reasoning. *Current Directions in Psychological Science* 26, 5 (2017), 434–441.

[24] Iwan Gurjanow, Miguel Oliveira, Joerg Zender, Pedro A Santos, and Matthias Ludwig. 2019. Shallow and deep gamification in mathematics trails. In *Games and Learning Alliance: 7th International Conference, GALA 2018, Palermo, Italy, December 5–7, 2018, Proceedings 7.* Springer, 364–374.

[25] Aaron Haim, Ethan Prihar, and Neil T Heffernan. 2022. Toward Improving Effectiveness of Crowdsourced, On-Demand Assistance from Educators in Online Learning Platforms. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II.* 29–34.

[26] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24 (2014), 470–497.

[27] Rui Huang, Albert D Ritzhaupt, Max Sommer, Jiawen Zhu, Anita Stephen, Natercia Valle, John Hampton, and Jingwei Li. 2020. The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational Technology Research and Development* 68, 4 (2020), 1875–1901.

[28] Taylyn Hulse, Maria Daigle, Daniel Manzo, Lindsay Braith, Avery Harrison, and Erin Ottmar. 2019. From here to there! Elementary: A game-based approach to developing number sense and early algebraic understanding. *Educational Technology Research and Development* 67 (2019), 423–441.

[29] Daniel C. Hyde, Alison Liu, Francesco Sella, Jérôme Prado, and Kirk Vanacore. 2022. Developmental Perspectives: Digital Interventions and Mathematics Learning in Typical and Atypical Populations. In *Proceedings of 2022 International Mind, Brain and Educations Society Conference.* 19.

[30] Tomislav Jagušt, Ivica Botički, and Hyo-Jeong So. 2018. Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Computers & education* 125 (2018), 444–457.

[31] Hui Jin and Donald B Rubin. 2008. Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* 103, 481 (2008), 101–111.

[32] Shimin Kai, Ma Victoria Almeda, Ryan S Baker, Cristina Heffernan, Neil Heffernan, et al. 2018. Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining* 10, 1 (2018), 36–71.

[33] Turkan Karakus, Yavuz Inal, and Kursat Cagiltay. 2008. A descriptive study of Turkish high school students' game-playing characteristics and their considerations concerning the effects of games. *Computers in Human Behavior* 24, 6 (2008), 2520–2529.

[34] Jeanine Krath, Linda Schürmann, and Harald FO Von Korflesch. 2021. Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior* 125 (2021), 106963.

[35] Elias Kyewski and Nicole C Krämer. 2018. To gamify or not to gamify? An experimental field study of the influence of badges on motivation, activity, and performance in an online learning course. *Computers & Education* 118 (2018), 25–37.

[36] Richard N Landers. 2014. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming* 45, 6 (2014), 752–768.

[37] Elise Lavoué, Qinjie Ju, Stuart Hallifax, and Audrey Serna. 2021. Analyzing the relationships between learners' motivation and observable engaged behaviors in a gamified learning environment. *International Journal of Human-Computer Studies* 154 (2021), 102670.

[38] Ji-Eun Lee, Jenny Yun-Chen Chan, Anthony Botelho, and Erin Ottmar. 2022. Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational technology research and development* 70, 5 (2022), 1575–1599.

[39] Sooyong Lee, Adam Sales, Hyeon-Ah Kang, and Tiffany Whittaker. 2023. Fully Latent Principal Stratification: combining PS with model-based measurement models.. In *Quantitative Psychology: The 87th Annual Meeting of the Psychometric Society, 2022.*, M. Wiberg, D. Molenaar, González, J.-S. Kim, and H. Hwang (Eds.). Springer Cham.

[40] Allison S Liu, Kirk Vanacore, and Erin Ottmar. 2022. How Reward-And Error-Based Feedback Systems Create Micro-Failures to Support Learning Strategies. In *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 1633-1636.* International Society of the Learning Sciences.

[41] Laura J Malkiewich, Alison Lee, Stefan Slater, Chenmu Xing, and Catherine C Chase. 2016. No Lives Left: How Common Game Features Could Undermine Persistence, Challenge-Seeking and Learning to Program. Singapore: International Society of the Learning Sciences.

[42] David C McClelland. 1961. *Achieving society.* Vol. 92051. Simon and Schuster.

[43] Megan Miller and Volker Hegelheimer. 2006. The SIMs meet ESL Incorporating authentic computer simulation games into the language classroom. *Interactive technology and smart education* (2006).

[44] Eleanor O'Rourke, Erin Peach, Carol S Dweck, and Zoran Popovic. 2016. Brain points: A deeper look at a growth mindset incentive structure for an educational game. In *Proceedings of the third (2016) acm conference on learning@ scale.* 41–50.

[45] Erin Ottmar, David Landy, Robert L Goldstone, and Erik Weitnauer. 2015. Getting From Here to There!: Testing the Effectiveness of an Interactive Mathematics Intervention Embedding Perceptual Learning.. In *CogSci*.

[46] Erin Ottmar, Ji-Eun Lee, Kirk Vanacore, Siddhartha Pradhan, Lauren Decker-Woodrow, and Craig A Mason. 2023. Data from the Efficacy Study of From Here to There! A Dynamic Technology for Improving Algebraic Understanding. *Journal of Open Psychology Data* 11 (2023), 5.

[47] Lindsay C Page. 2012. Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness* 5, 3 (2012), 215–244.

[48] Lindsay C Page, Avi Feller, Todd Grindal, Luke Miratrix, and Marie-Andree Somers. 2015. Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation* 36, 4 (2015), 514–531.

[49] Marina Papastergiou. 2009. Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & education* 52, 1 (2009), 1–12.

[50] RDCT R Team et al. 2014. R: A language and environment for statistical computing. *Vienna, Austria: R Foundation for Statistical Computing* (2014).

[51] Steven Ritter and Stephen Fancsali. 2016. MATHia X: The Next Generation Cognitive Tutor.. In *EDM*. ERIC, 624–625.

[52] Teomara Rutherford, George Farkas, Greg Duncan, Margaret Burchinal, Melissa Kibrick, Jeneen Graham, Lindsey Richland, Natalie Tran, Stephanie Schneider, Lauren Duran, et al. 2014. A randomized trial of an elementary school mathematics software intervention: Spatial-temporal math. *Journal of Research on Educational Effectiveness* 7, 4 (2014), 358–383.

[53] Michael Sailer, Jan Ulrich Hense, Sarah Katharina Mayr, and Heinz Mandl. 2017. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in human behavior* 69 (2017), 371–380.

[54] Adam C Sales and John F Pane. 2019. The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics* 13, 1 (2019), 420–443.

[55] Adam C Sales and John F Pane. 2021. Student log-data from a randomized evaluation of educational technology: A causal case study. *Journal of Research on Educational Effectiveness* 14, 1 (2021), 241–269.

[56] Adam C Sales, Asa Wilks, and John F Pane. 2016. Student Usage Predicts Treatment Effect Heterogeneity in the Cognitive Tutor Algebra I Program. *International Educational Data Mining Society* (2016).

[57] Nicholas C Soderstrom and Robert A Bjork. 2015. Learning versus performance: An integrative review. *Perspectives on Psychological Science* 10, 2 (2015), 176–199.

[58] Jon R Star, Courtney Pollack, Kelley Durkin, Bethany Rittle-Johnson, Kathleen Lynch, Kristie Newton, and Claire Gogolen. 2015. Learning from comparison in algebra. *Contemporary Educational Psychology* 40 (2015), 41–54.

[59] Daniel J Stekhoven and Peter Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.

[60] Katherine A Sward, Stephanie Richardson, Jeremy Kendrick, and Chris Maloney. 2008. Use of a web-based game to teach pediatric content to medical students. *Ambulatory Pediatrics* 8, 6 (2008), 354–359.

[61] Kirk Vanacore, Alison Liu, Adam Sales, and Erin Ottmar. 2022. The Impact of reward-based feedback on persistence behavior in an education game.. In *Proceedings of 2022 International Mind, Brain and Educations Society Conference*.

[62] Matthew Ventura, Valerie Shute, and Weinan Zhao. 2013. The relationship between video game use and a performance-based measure of persistence. *Computers & Education* 60, 1 (2013), 52–58.

[63] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[64] Monica Wijers, Vincent Jonker, and Kristel Kerstens. 2008. MobileMath: the Phone, the Game and the Math. In *proceedings of the European conference on game based learning, Barcelona*. 507–516.

[65] Zamzami Zainuddin, Samuel Kai Wah Chu, Muhammad Shujahat, and Corinne Jacqueline Perera. 2020. The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review* 30 (2020), 100326.

4.2 How Behavioral Tendencies Towards Persistence Moderates the Effectiveness of Computer-Based Learning Platforms

See manuscript below.

Behavior-Moderated Effects: Effectiveness of Computer-Based Learning Platforms Varies Bases on Students Tendencies Toward Productive Persistence

Kirk Vanacore

Worcester Polytechnic Institute

Due Date

Abstract

This study investigates the role of productive persistence, specifically reattempting problems after sub-optimal performance, in the efficacy of computer-based learning platforms (CBLPs) designed for middle school algebra instruction. Utilizing data from an efficacy trial involving three CBLPs —– From Here to There (FH2T), DragonBox, and ASSISTments -— we employ a Fully Latent Principal Stratification (FLPS) model to evaluate how students' propensity to persist moderates program efficiency. Results indicate that reattempting problems is a productive persistence behavior, contributing positively to performance improvement. However, the moderating effect of this behavior varies across the CBLPs, suggesting that the peculiarities of instructional design may influence how students engage in persistent behavior and learn from the content. The study highlights the need to design educational technologies that foster persistence by encouraging students to attempt multiple problem-solving strategies. These findings underscore the importance of understanding behavior-moderated effects, which show how program efficacy varies based on students' engagement tendencies. Studying this heterogeneity provides insights into why programs' impact varies across students and what institutional design features may help all students benefit from educational innovations.

Behavior-Moderated Effects: Effectiveness of Computer-Based Learning Platforms Varies Bases on Students Tendencies Toward Productive Persistence

Introduction

Education researchers increasingly recognize effect heterogeneity as essential to efficacy research (Boedeker & Henson, 2020; Bryan et al., 2021; Fadda et al., 2022; Schudde, 2018). A program's impact may vary by implementation (Durlak & DuPre, 2008; Lendrum & Humphrey, 2012) or differ based on students' characteristics (Decker-Woodrow et al., 2023a; Pane et al., 2023). Kizilcec and Lee (2022) argue that educational innovations should not expand the gaps between advantaged and disadvantaged students. Evaluating heterogeneity can allow us to understand who benefits from innovative programs and guard against potential gap-increasing innovations.

Heterogeneity analyses typically evaluate effects by student's prior knowledge or demographic groups; however, students' specific tendencies towards interacting with the programs can also moderate the impact of those programs. This type of heterogeneity, which we call behavior-moderated effects, is largely unexplored. Typically, assessing students' specific interactions with program components may be difficult for education program implementations; however, computer-based learning platforms (CBLPs) allow researchers to log detailed data on student behaviors, thus providing an opportunity to understand how these behaviors relate to the platform's impact.

One particularly salient behavior is reattempting to solve problems after a poor performance. The opportunity to reattempt problems after receiving immediate feedback is a critical feature of many CBLPs. For example, mastery learning programs allow students to complete comparable problems until they have mastered the underlying knowledge component (Bloom, 1968; C.-L. Kulik et al., 1990). In some cases, programs encourage students to try similar problems after submitting an incorrect answer (Kuchipudi, 2024). Furthermore, some gamified platforms allow students to explore multiple routes to a solution (Pradhan et al., 2024).

Reattempting problems may indicate persistence in learning tasks, which is essential to learning difficult concepts and skills. Persisting during difficult academic tasks allows students to work in the upper ends of their zones of proximal development, where most learning occurs (Ventura et al., 2013; Vygotsky & Cole, 1978). Since productive persistence is essential to learning, the efficacy of programs that create opportunities for persistence may depend on whether students take advantage of those opportunities.

In this paper, we estimate this dependence. This study is a secondary data analysis from an efficacy study of three CBLPs used to teach algebra to middle school students. The two gamified CBLPs, From Here To There (FH2T) and DragonBox 12+ (DragonBox), allow students to dynamically manipulate expressions and equations to solve algebra problems. In both of these conditions, students can reattempt problems after using inefficient solution paths. Another CBLPs, ASSISTments, presents typical textbook-style problem sets with on-demand hints and immediate feedback. In this condition, students must submit the correct answer before progressing to the next problem. We compare these conditions to an Active Control, also administered through ASSISTments, of textbook-style problems on which students can make one attempt.

Overall, we find that students with high propensities for persistence benefit more from all the conditions in which they can reattempt problems (i.e., all conditions other than the Active Control). However, the differences between conditions' moderating slopes suggest that nuances in the presentation of persistent opportunities may influence students' learning. Furthermore, we find that students with low persistence propensities may experience adverse effects compared to the control condition. Overall, the study provides evidence that the efficacy of these CBLPs may be contingent upon students' behavioral tendencies.

Background

The following section provides an overview of this paper's context and theoretical underpinnings. First, we describe the nuances of CBLPs and provide evidence of their impact on learning. This section includes an overview of how gamification may influence students' behaviors in CBLPs. Next, we discuss how reattempting problems fits into the learning process, along with the potential benefits and drawbacks of the behavior. Finally, we examine how this behavior often manifests in various CBLPs.

Computer-Based Learning Platforms

In recent decades, educational intuitions have increasingly relied on online programs to administer homework, assessment, and instruction (Gallop, 2019; Gray & Lewis, 2021). Many of these technologies were created to function as individual tutors by responding to students as they learn skills and solve problems (Bloom, 1968; A. Corbett, 2001). Meta-analyses and reviews of CBLPs have mostly found modest positive effects, with evidence that efficacy varies by the type and implementation style of the program (Cheung & Slavin, 2011; J. A. Kulik & Fletcher, 2016). In the following section, we characterize two instructional designs common in CBLPs: traditional platforms that seek to emulate a tutoring environment and gamified platforms that incorporate game-like features into learning platforms to increase engagement and learning. Notably, these characteristics are not necessarily mutually exclusive – a traditional platform can be gamified, and gamified platforms often incorporate features associated with traditional platforms – but for this paper, we differentiate between program features related to each of these instructional designs.

Traditional CBLPs generally present students with problems similar to those found in a typical worksheet (e.g., equations to be solved, word problems, etc.) while often providing added support like hints, feedback, and scaffolding. Problems in traditional CBLPs are typically multiple-choice, fill-in, or open-response questions (Gurung et al., 2024). Features are incorporated into traditional CBLPs to varying degrees of adaptability and responsiveness to students; some CBLPs only provide simple features like on-demand hints and automated feedback, whereas others employ more suffocated features to guide students through learning activities. Often, these programs use a mastery learning method, giving students problems on the same knowledge component until they reach a mastery threshold. Statistically and machine learning algorithms are often employed to determine which problems a student should receive (Kozierkiewicz-Hetmańska & Nguyen, 2010), which support will be most helpful (Prihar et al., 2023), and whether the student has learned the content or mastered the skill (Abdelrahman et al., 2022; A. T. Corbett & Anderson, 1994; Yudelson et al., 2013). These features are generally meant to emulate high-quality human tutors who can guide students through a learning experience with appropriate challenges and support.

Studies have found positive effects of traditional CBLPs on student learning, such as Mathia (formally Cognitive Tutor) and ASSISTments (Feng et al., 2023; Pane et al., 2010, 2014; Roschelle et al., 2016). Feng et al. (2023) found that ASSISTments improved long-term learning and disproportionately helped students of color, providing evidence that ASSISTments is a gap-closing intervention; however, there is evidence that the effects of the programs may vary by how students interact with them. Sales and Pane (2021) found that students who used hints in Cognitive Tutor experience greater effects. Alternatively, in ASSIStments, the impact of hint access varies based on how students use those hints (Vanacore, Gurung, Sales, & Heffernan, 2024). Thus, the effect of CBLPs may depend on whether these systems can influence students' interactions with their content and features.

Gamified CBLPs

CBLPs may influence students' behavior through gamification, the method of incorporating game-like elements into non-game experiences, such as learning tasks. According to the *Theory of Gamified Learning*, gamification indirectly impacts learning outcomes by influencing students' attitudes and learning-related behaviors (Landers, 2014). Gamification can potentially improve students' motivation and engagement with learning content (A. Liu et al., 2022; Zainuddin et al., 2020).

Evaluations of gamified CBLPs have produced mixed results. A meta-analysis of 19 experiments on the cognitive effects of gamified CBLPs found modest but consistently positive effects of gamification (Sailer & Homner, 2020). Jagušt et al. (2018) found that including specific game elements – competition, narratives, and adaptive difficulty – in digital math lessons led to increased performance as students used the program, whereas non-gamified versions caused a decrease in performance. However, another study found that students benefited more from a non-gamified intelligent tutoring system than a gamified CBLP and that gamified CBLPs do not outperform their peers (Long & Aleven, 2017).

There is also mixed evidence that gamification can influence student engagement. Richey et al. (2021) found that the effect of gamification on learning was mediated by student gaming the system behaviors (i.e., their tendencies to exploit CBLPs' features to progress through content without learning). These results indicated that gamification caused a reduction in gaming the system behaviors, thus improving students' learning. Alternatively, another study found mixed results on gamification's ability to mitigate students' gaming the system tendencies (Vanacore, Gurung, Sales, & Heffernan, 2024). Similarly, performance-based rewards have had mixed effects on engagement (Hyde et al., 2022; Malkiewich et al., 2016; Vanacore, Sales, et al., 2024). Many of these differences may be due to how CBLPs are gamified, as different game elements or combinations of elements result in significantly different effects (Sailer et al., 2017). Further, the same game elements can have different effects on different students (Denden et al., 2022; Vanacore, Sales, et al., 2023). Because of these complex interactions, most studies fail to provide theoretical explanations of how gamification relates to learning, engagement, and motivation (Zainuddin et al., 2020). Thus, questions remain regarding when and how gamification features affect academic outcomes.

Persistence and Learning through CBLPs

Persistence is considered an aspect of conscientiousness in which a person is driven to complete challenging tasks (McClelland, 1961). Students must persist when struggling to learn difficult skills or understand challenging concepts. However, many students tend to disengage when struggling in CBLPs (Botelho et al., 2019). Thus, a substantial body of research investigates measuring and understanding persistence-related behaviors in CBLPs, including both productive and unproductive performance (Adjei et al., 2021; Kai et al., 2018; Owen et al., 2019; Park, 2023; Shute et al., 2015; Ventura et al., 2013).

Persistence in CBLP has been measured in a variety of ways. Researchers have used students' time spent on challenging and unsolved problems to measure persistence (Shute et al., 2015; Ventura et al., 2013). In mastery learning activities, productive presence is often defined as mastering a knowledge component after completing over ten problems within an activity, signifying that students needed to exert substantial effort to learn the skill (Adjei et al., 2021; Kai et al., 2018). Alternatively, unproductive persistence, also known as wheel spinning, is categorized by students exerting effort without learning (Owen et al., 2019; Park, 2023). Notably, productive persistence is positively associated with short and long-term academic outcomes, whereas unproductive persistence is not predictive of academic achievement (Adjei et al., 2021; Ventura et al., 2013).

Displaying Persistence by Reattempting Problems

Features allowing students to practice similar or reattempt problems are common in CBLPs. Programs that provide immediate feedback may allow students to resubmit answers after an incorrect response. Mastery learning platforms present students with multiple problems with similar skills, allowing students to reattempt the problems they have displayed sufficient mastery of the skill. Other similar programs allow students to review or replay problems they have previously attempted. Finally, more process-oriented programs present problems with multiple paths to solutions, allowing students to find new paths to a solution in multiple ways (Vanacore, Lee, et al., 2023). These programs often evaluate performance using the quality of solution paths and allow students to reattempt problems based on improving those paths (Decker-Woodrow et al., 2023b; Long & Aleven, 2017).

Reattempting problems after sub-optimal performance may indicate a student's persistence in the learning activity. In one online math game, students who spent a considerable proportion of game time reattempt completed problems showed the largest learning gains in algebraic understanding (Lee et al., 2022). Lavoué et al. (2021) found that students who reattempted problems to improve performance tended to have higher levels of self-determined motivation.

There is some evidence that gamification can increase students' tendencies towards reattempting problems. Performance-based feedback in game settings can encourage students to reattempt problems by including game-based failure, which, unlike typical failure experiences in educational settings, can motivate students to persist (Vanacore, Sales, et al., 2024). Other studies have found that performance-based feedback and reward systems can increase engagement, persistance, and reattempt behaviors (Gaston & Cooper, 2017; O'Rourke et al., 2016).

Current Study

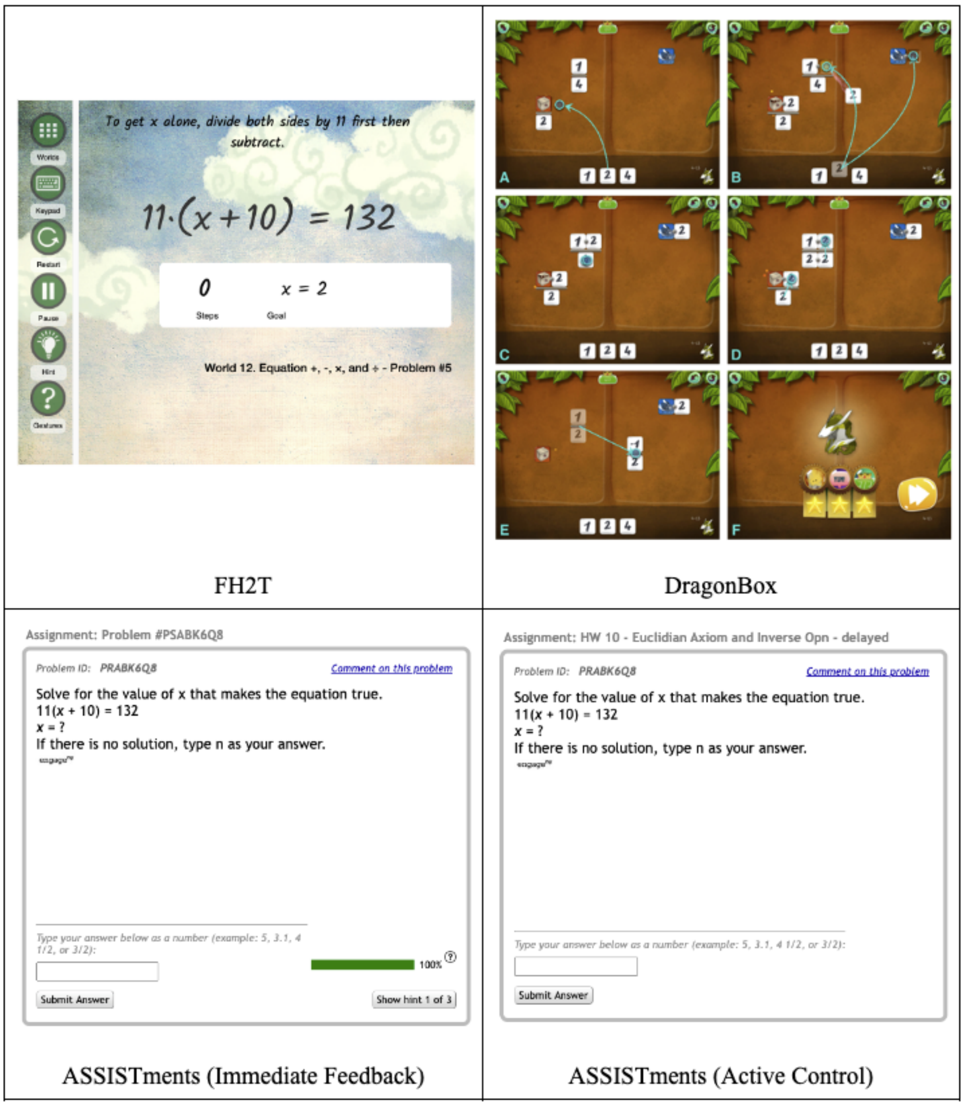
Study Design

This study uses open-sourced data from a randomized controlled trial conducted in a school district in the Southeastern United States, involving 52 teachers in 10 schools. The randomization procedure involved ranking students within classrooms based on their prior state mathematics assessment scores, blocking them into quintets, and randomly assigning them into either the From Here to There (40%), DragonBox-12 (20%), problem sets with on-demand hints and immediate Feedback administered through ASSISTments (20%), or an Active Control (20%). The study's original purpose was to evaluate and understand From Here to There; thus, the researchers disproportionately sample weighted towards that condition.

The study was implemented through nine weekly half-hour sessions in which the students worked in their assigned program. The sessions were intended to be part of students' in-school coursework. However, as the study was conducted during the pandemic, some students attended school remotely (39.54% started the year remotely, and 27.49% remained remote throughout the entire school year). Due to the pandemic, the study involved substantial attrition. Decker-Woodrow et al. (2023a) provides a complete description of the study design, attrition analysis, and the results. Vanacore, Ottmar, et al. (2024) presents an analysis of the study implementation fidelity, and Ottmar et al. (2023) contains a full description of open-sourced data.

Conditions

The study involves four conditions administered through two gamified CBLPs and one traditional CBLP, all focusing on teaching aspects of algebraic equation equivalency. Each condition teaches procedural ability, conceptual knowledge, and flexibility in solving algebraic expressions. Example problems from each condition are presented in Figure 1. Each condition is described in detail below.



To get x alone, divide both sides by 11 first then subtract.

$11 \cdot (x + 10) = 132$

World 12. Equation +, -, x, and ÷ - Problem #5

FH2T

DragonBox

Assignment: Problem #PSABK6Q8

Problem ID: PRABK6Q8

Solve for the value of x that makes the equation true.
11(x + 10) = 132
x = ?
If there is no solution, type n as your answer.

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

ASSISTments (Immediate Feedback)

Assignment: HW 10 - Euclidian Axiom and Inverse Opn - delayed

Problem ID: PRABK6Q8

Solve for the value of x that makes the equation true.
11(x + 10) = 132
x = ?
If there is no solution, type n as your answer.

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

ASSISTments (Active Control)

Figure 1

Example problems from each condition.

From Here to There! (FH2T)

FH2T[1] is an algebra game that applies aspects of perceptual learning (Goldstone et al., 2010, 2017) and embodied cognition (Abrahamson et al., 2020) to teach students fluency when using algebraic expressions. FH2T problem gives students a starting expression (start state) that they must dynamically transform into a mathematically equivalent expression (goal state). Students can manipulate the expression by dragging numbers and symbols from one position to another on the screen or using a keypad when expanding terms. The program prevents students from making mathematically invalid manipulations. FH2T evaluates performance based on the efficiency of the solution process – the number of valid manipulations (steps) they take to reach the goal state. FH2T includes 252 problems that are presented sequentially by mathematical content. The problems become more complex through the program. Students must complete each problem in the sequence before advancing.

DragonBox-12 (DragonBox)

DragonBox[2] is an educational game that provides instruction in algebraic concepts to secondary school students (ages 12-17). For each problem, students must isolate a box containing a dragon—equivalent to solving an equation for x. This design incorporates research-based instructional designs, including discovery-based learning, embedded gestures, diverse representations of concepts, immediate feedback, and adaptive difficulty (Cayton-Hodges et al., 2015; Torres et al., 2016). Students start by learning algebra rules without using or manipulating numbers or traditional algebraic symbols. Instead they engage with the algebraic concepts as if they are puzzles. The program introduces numbers and traditional algebraic symbols gradually after the student has learned the underlying concepts. Furthermore, DragonBox includes a narrative goal: students must release the dragon from the box by isolating it in the equation. Previous analyses found that DragonBox positively affects engagement and attitudes toward math, but findings are mixed regarding its efficacy in improving learning outcomes (Kluge &

[1] https://graspablemath.com/fh2t.html

[2] https://dragonbox.com/products/algebra-12

Dolonen, 2015; Z. Liu et al., 2017; Long & Aleven, 2017; Siew et al., 2016).

On Demand Hints and Immediate Feedback in ASSISTments (Immediate Condition)

Both Immediate Conditions and the Active Control (described below) were administered through ASSISTments[3] (Heffernan & Heffernan, 2014), an online homework system that helps students as they solve traditional problem sets. The problem sets in ASSISTments resemble problems students encounter in their textbooks and homework assignments. ASSISTments presents students with problems one at a time on their screen. Each condition included 218 problems selected from three curricula – *EngageNY*, *Utah Math*, and *Illustrative Math* – to address specific algebra skills. The problems were organized into nine problem sets, administered throughout nine half-hour sessions.

In the Immediate Condition, students could request hints while solving problems. After every submitted answer, they also received automatic feedback on whether their answer was correct or incorrect. Each problem contained a series of hints with a similar structure. The first hint gave the students the first step to answering the problem. The second hint gave the student a worked example of a similar problem. The final hint provided the student with the steps to complete the problem as well as the problem's solution. Students could submit as many answers as needed but could only move on once they had entered the correct answer. Multiple studies have found positive effects of this method when implemented as a substitute for paper homework (Feng et al., 2023; Roschelle et al., 2016).

Post-Assignment Hints and Feedback (Active Control)

The Active Control provided post-assignment assistance rather than on-demand hints and immediate feedback. In this condition, ASSISTment administered the problem sets in "test mode," so students did not receive any feedback or hints while submitting answers in each problem set. They submitted one answer for each problem and progressed through the problem set without receiving an indication of whether their answers were correct. At the end of each

[3] https://new.assistments.org/

problem set, students received a report with performance feedback. They also had the opportunity to review their responses, revisit problems, and request hints.

Reattempting Problems

Each treatment condition (i.e., FH2T, DragonBox, and Immediate Condition) allowed students to reattempt problems differently. In the Immediate Condition, students had to reattempt the problems until they submitted the correct answer. Although they could review their results, they could not reattempt the problems after submitting correct answers. Although ASSISTments has mastery-based learning activities and features that allow students to receive problems similar to those they struggled with, neither was active during the study.

In both gamified conditions, upon completing a problem, students could repeat problems. Unlike many traditional math problems, which evaluate students based on whether they provide the correct answers, FH2T and DragonBox focus on students' processes for finding solutions. The student's task for each problem is to manipulate the equation to a specific state in the optimal number of steps. This instructional design allowed students to explore multiple paths to a solution.

After completing each problem in FH2T, students receive feedback for a "reward" represented by clovers (see Figure 2). They are given between one and three clovers depending on their efficiency: three clovers if the students took the optimal (i.e., minimal) number of steps to reach the goal state, two clovers if they were within two steps of optimal, and one clover for all other completed attempts. When students receive three clovers, they are automatically moved to the next problem. However, when students earn one or two clovers, they are given the option to either reattempt the problem or move on to the subsequent problem. Previous analyses suggest that the number of clovers received has a causal effect on whether students reattempt the problem (Vanacore, Sales, et al., 2024). This findings suggest that students are responding to the gamified nature of the program by displaying persistent learning behaviors when taking the opportunity to reattempt problems to improve their skills and understanding.

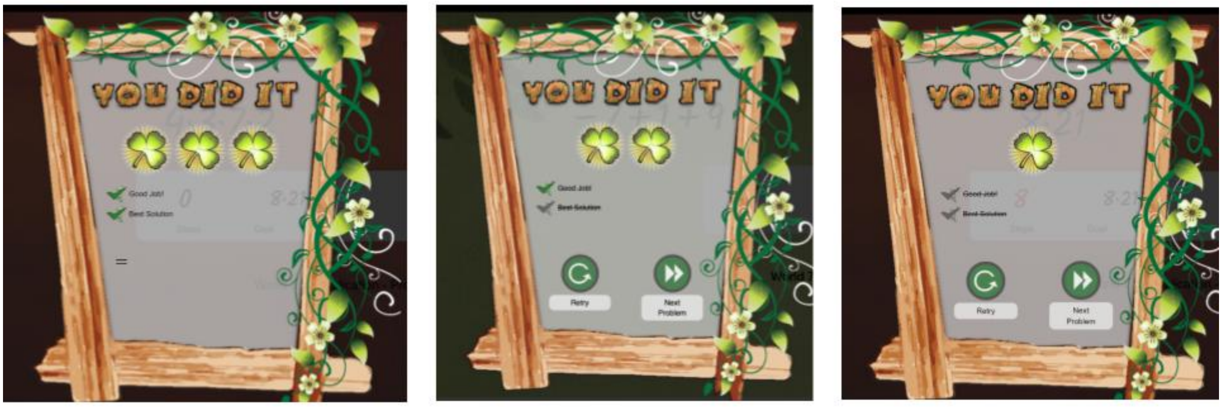
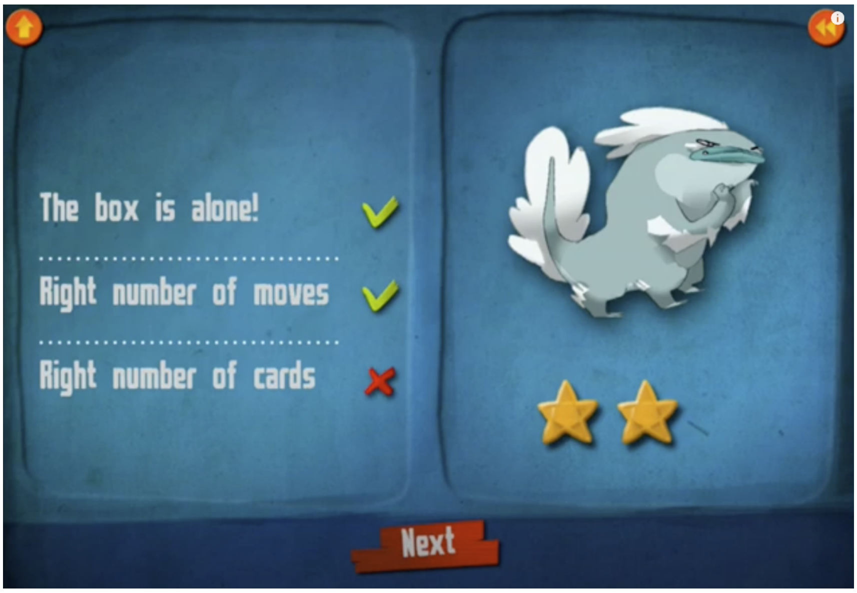


Figure 2

Performance based feedback in FH2T. Note that when students did not achieve the "Best Solution," the "Retry" option is available.

DragonBox operates similarly by asking students to isolate a variable and are evaluated based on whether they do it efficiently ("Right number of moves") and whether they have the least number of variables or numbers in the solution ("Right number of cards"). Students receive a star for completing the problem, a star for the correct number of moves, and a star for the right number of variables (see Figure 3). As in FH2T, students can reattempt problems when they have not earned all the stars.



The box is alone! ✓

Right number of moves ✓

Right number of cards ✗

Figure 3

Performance-based Feedback in DragonBox. Note that the button in the upper left corner allows students to reattempt problems.

There are two crucial distinctions between the designs of FH2T and DragonBox. First, in FH2T, students cannot complete the problems with extra variables; in other words, the student needs to completely achieve the goal state before completing the problem. This is not the case in DragonBox. Second, there is a subtle but important difference between how the conditions present the option of reattempting problems. FH2T presents a “retry" button in the center of the students’ screen when they use a sub-optimal solution. In Dragbox, to reattempt a problem, students select the back button in the screen’s upper right corner.

Notably, these rewards had no internal or external ramifications. In the FH2T and DragonBox, students could see how many rewards they earned but could not use them in the game. In the study, teachers did not have access to students’ performance data. Therefore, there was no external pressure on students to reattempt problems.

Research Questions

This study explores how differences in behavioral tendencies – specifically, whether students display productive persistence by reattempting problems after sub-optimal performance

– are associated with varying effects of CBLPs. The following questions are meant to evaluate whether reattempting problems is a productive persistence behavior (Research Question 1) and whether the effects of the treatment conditions vary based on this behavior (Research Questions 2 and 3).

Research Question 1 Does students' performance improve after they reattempt problems?

First, we evaluate whether reattempting problems is a productive persistence behavior by whether the behavior is associated with improved performance.

Research Question 2 Does the effect of each treatment condition vary based on students' propensity to reattempt problems after sub-optimal performance?

Next, we evaluated the effect heterogeneity of each treatment condition based on students' propensities to reattempt problems after sub-optimal performance. We hypothesized that students with higher propensities to reattempt problems would experience greater effects in all treatment conditions than those with lower propensities.

Research Question 3 Is this heterogeneity consistent across the conditions?

Finally, we evaluate whether the effect heterogeneity differs by treatment condition. We are specifically concerned with whether the pattern of effect heterogeneity is consistent across the gamified conditions. We hypothesized that the effects of the gamified conditions vary similarly by students' propensities to reattempt problems and the heterogeneity of both gamified conditions would be greater compared to the Immediate Condition.

Method

Participants

The original study sample included 3,271 middle school students. As stated above, the study was continued during the COVID-19 pandemic, and there was substantial attrition. The original efficacy study reported that the differential attrition was not statistically significant and met the What Works Clearing House Standards for tolerable threats of bias under optimistic assumptions (Decker-Woodrow et al., 2023b; WWC, 2020). The current study sample differs from the original efficacy study in that the efficacy dropped all students with missing pretest or posttest scores. The current study excludes all students without post-test scores, which are used as the outcome, but imputes all other missing data. Thus, our sample is slightly larger than that of the efficacy study.

The sample for the current study consists of 1976 students: 402 in the Immediate Condition, 835 in the Delayed Condition, 372 in the DragonBox, and 817 in FH2T. These students were taught by 36 teachers in 10 schools. Demographic data was provided by the district. The sample was balanced between male and female students (49.00% Female, 51.00% male). The district did not report any non-binary students. About half the student population was White (51.21%), 25.29% was Asian, 15.08% was Hispanic/Latino, 4.41% was Black/African American, 0.56% American Indian, 0.10% was Pacific Islander, 3.29% were two or more races and 0.06% did not have reported race/ethnicity data. English was not the first language for 10.26% of the students. Students who received accommodations (i.e., Individual Education Plans or 504 Plans) accounted for 16.11% of the student population, and 16.62% were labeled “gifted" by the district.

Data & Variables

Persistence Behavior

We used students’ log-file data from FH2T and created an indicator variable for when students reattempted a problem after performing sub-optimally (i.e., using an inefficient path to the solution and receiving fewer than three clovers). Although students could reattempt problems in DragonBox, we did not have access to their log-file data. Students in the Immediate Condition had to submit multiple responses after an incorrect answer before progressing; consequently, it is difficult to determine whether a student displayed persistence when submitting subsequent answers or was simply attempting to progress through the system. Because of these complications, persistence was estimated empirically for students in FH2T and then imputed for students in other conditions, as explained below.

Pretreatment Predictors

To estimate students' propensity to persist in FH2T, we use data from assessments administered before students used their assigned condition. The original studies' researchers administered the following pretests: algebraic knowledge, math anxiety, and perceptual processing skills. Pretest algebraic knowledge was a variant of the learning outcome described below. The math anxiety assessment was adapted form from the *Math Anxiety Scale for Young Children-Revised* (Chiu & Henry, 1990), which assessed negative reactions towards math, numerical inconfidence, and math-related worrying (Cronbach's α =.87; items accessible on OSF[4]). Five items adapted from the Academic Efficacy Subscale of the *Patterns of Adaptive Learning Scale* to assess math self-efficacy (Midgley et al., 2013; Cronbach's α = .82; items accessible on OSF [5]). The perceptual processing assessment evaluates students' ability to detect mathematically equivalent and nonequivalent expressions as quickly as possible (Bye et al., 2022; items accessible on OSF [6]). The district provided additional data on the students, including their demographics and most recent standardized state test scores in math (of Education, 2020). Demographic data included race/ethnicity, individualized education plan status (IEP), and English as a second or foreign language (ESOL) status. Race/ethnicity was dummy-coded, with white students as the reference category because they were the majority population.

Table 1 presents the correlations between the pretest scores. As expected, many of the scores are moderately or highly correlated. However, math anxiety and its related subscores are not highly correlated with the mathematical knowledge scores. Furthermore, math self-efficacy is not highly correlated with scores other than math anxiety. This suggests that while these measures assess related constructs, they still evaluate different aspects of students' mathematics knowledge and math identity, which may be leveraged to assess their propensity to reattempt mathematical tasks.

[4] https://osf.io/rq9d8

[5] https://osf.io/rq9d8

[6] https://osf.io/r47ev

Table 1

Correlations Between Pretest Covariates

	–	2	3	4	5	6	7	8	9	10
1. Algebraic Knowledge Total Score	–									
2. Procedural Sub-Score	0.77	–								
3. Conceptual Sub-Score	0.89	0.53	–							
4. Flexibility Sub-Score	0.75	0.39	0.50	–						
5. Math Anxiety Total Score	-0.25	-0.16	-0.25	-0.18	–					
6. Negative Reaction Math Sub-Score	-0.26	-0.16	-0.26	-0.19	0.80	–				
7. Numeric Inconfindence Sub-Score	-0.20	-0.12	-0.21	-0.13	0.78	0.39	–			
8. Math Self-Efficacy	0.33	0.21	0.33	0.25	-0.58	-0.50	-0.47	–		
9. Perceptual Sensitivity Score	0.70	0.53	0.67	0.47	-0.24	-0.25	-0.18	0.30	–	
10. Math State Test Score	0.63	0.43	0.63	0.44	-0.28	-0.27	-0.24	0.37	0.60	–

All correlations were significant at p < 0.001.

We standardized all continuous scores (z-score) to improve model convergence and ease interpretation. Assessment test times were log-transformed. We included polynomials of the pretest scores when they improved model fit. We imputed missing pretreatment data using single-imputation with the Random Forest routine implemented by the missForest package in R (R Core Team, 2016; Stekhoven & Buehlmann, 2012).

FH2T Problem Features

When estimating students' propensity to reattempt problems, we accounted for how complex the problem is and where it is situated in the sequence of each problem set. The minimum number of steps it would take to complete each problem served as a measure of problem complexity. The sequence within each problem set (referred to as ”Worlds" in FH2T) was also included in the model.

Learning Outcome

We evaluated learning using a standardized measure of students' algebraic knowledge, administered two weeks after students stopped using the programs. The assessment included ten multiple-choice items from a previously validated measure of algebra understanding (Star et al., 2015; Cronbach's α = .89; items accessible on OSF[7]). Four of the items evaluated conceptual understanding of algebraic equation-solving (e.g., the meaning of an equal sign), three evaluated procedural skills of equation-solving (e.g., solving for a variable), and three evaluated flexibility of equation-solving strategies (e.g., evaluating different equation-solving strategies). Together, these ten items assessed students' knowledge in algebraic equation-solving, the improvement of which was the goal of the interventions. The learning outcome was assessed before and after the interventions. Students' scores were standardized (z-score) to ease model fit and interpretation.

Analytic Approach

Reattempt Behavior Exploration

First, we conducted an exploratory data analysis that examined the reattempt behavior in FH2T, with the aim of understanding the context in which students tend to reattempt problems. To address Research Question 1, we use multilevel logistic regression to assess whether reattempting problems after sub-optimal performance is associated with better performance on subsequent problems. This indicates whether the behavior is likely productive or unproductive: if students improve their performance after reattempting a problem, the behavior may be productive.

Fully Latent Principal Stratification

Research Questions 2 and 3 pose methodological complications. First, our moderator variable of interest, reattempt behavior, does not manifest equally across conditions, and in the Active Control, students cannot reattempt problems. Additionally, students' tendencies toward reattempting problems may be confounded by conditions: reattempting in FH2T is not necessarily equivalent to reattempting in the Immediate Condition. Typically, to estimate effect heterogeneity,

[7] https://osf.io/uenvg

researchers must define subgroups before the intervention and be independent of any treatment. For example, when testing if treatment effects vary based on students' prior knowledge of the content, interacting with the treatment indicator with measures of test performance collected before the treatment provides an estimate of effect heterogeneity. Yet, our subgroups of interests – students with different reattempt behaviors – only become evident after an intervention begins.

Principal stratification is a method used in randomized controlled trials for estimating the intervention effects on subgroups that emerge after the treatment has begun (Frangakis & Rubin, 2002; Page et al., 2015). For example, the effectiveness of an educational program may vary based on the extent to which students engage with the program's components; however, subgroups identified based on students' interactions with a treatment program are unknown before the treatment begins. Furthermore, these subgroups may only be observed for students randomized into some conditions. The essential premise of principal stratification is that even subgroups defined by interactions with the treatment exist before treatment and equally across conditions due to randomization.

Fully Latent Principal Stratification (FLPS) models these moderator behaviors as latent characteristics (Sales & Pane, 2019). For participants who are randomized to the conditions in which the behavior of interest is expressed, the latent state can be estimated from their behavioral data. For participants who are randomized to other conditions, the latent state is essentially imputed using data collected before the randomization. The latent state variable can be interpreted as a propensity to display the behavior if randomized to the condition in which the behavior is expressed. This propensity exists before randomization and is, therefore, unconfounded by any condition. In a model predicting the outcome, the interaction of the treatment and the latent state variable estimates the heterogeneity associated with the behavior in question (i.e., behavior-moderated effects).

This method has particular benefits for education efficacy research as students' engagement within learning activities is often characterized by a complex series of behaviors. For example, in many CBLPs, students may display an array of behaviors during the program, such as meeting implementation goals (Dieter et al., 2020; Vanacore, Ottmar, et al., 2024), mastering knowledge components in mastery-learning activity (Sales & Pane, 2019), and gaming-the-system behaviors (Vanacore, Gurung, Sales, & Heffernan, 2024) that may be indicators of latent student characteristics (i.e., high fidelity users, mastery users, gamers). Understanding how specific behaviors may influence the CBLP's impact on students' learning is often difficult. FLPS provides an approach to addressing the relationships between in-program behaviors and program effects.

Since we are concerned with how the effects of each CBLP in this study vary based on their propensity to reattempt problems when using FH2T, regardless of the actual assignment, we define FH2T as the primary treatment and the Active Control as the control. The Immediate Condition and DragonBox will treated as separate treatments as delineated below.

Let τ_i be subject i's individual treatment effect: the difference between what i's posttest score would be if i were randomized to treatment and their score if randomized to control. Let $\mathscr{T}$ be an index of the treatment conditions where $FH2T$, DB, and IC signify FH2T, DragonBox, and the Immediate Condition, respectively. Let α_{ti} be i's be a measurement of a student's propensity to reattempt problems. α_{ti} is defined for students who were not randomized to FH2T, as each student had the potential to reattempt problems had they been randomized to FH2T, even if this potential was never realized.

The principal effect is the treatment effect for the subgroup of students with a particular value for α_t:

$$\tau^{\mathscr{T}}(\alpha) = E[\tau|\alpha_t] \tag{1}$$

The FLPS method allows us to estimate the function $\tau(\alpha)$ for each treatment condition. Intuitively, FLPS can be viewed as three steps: (1) estimate α_t using reattempt data from FH2T as a function of pre-treatment covariates observed in both groups, (2) use that model to impute α_t for students who were not randomized to FH2T, (3) estimate $\tau(\alpha)$ by interaction α_t and the treatment conditions in a linear region model. In reality, these steps are fit simultaneously using iterations through these steps in a Bayesian principal stratification model with a continuous variable consisting of measurement and outcome submodels, as outlined by Jin and Rubin (2008) and Page (2012). We fit the FLPS model using the Stan Markov Chain Monte Carlo (MCMC) software (Arezooji, 2020).

FLPS Implementation

Measurement Submodel

We estimate α_t by running a multilevel logistic submodel predicting whether the reattempted problems in FH2T using students' FH2T problem logs. Equation 2 models two components that may contribute to the behavior: the student and the problem. This model resembles a Rasch model, which has shown promise when evaluating students' behavioral tendencies in CBLPs (Huang et al., 2023).

Let R_{ji} be a binary indicator of whether student i reattempts problem j. As noted above, only problem attempts in which student i performed sub-optimally on their first attempt at problem j are included in the model. Let α_i be student i's contribution to the likelihood of the reattempt behavior and η_j be the baseline likelihood for all students on the problem j.

$$log(odds(R_{ji} = 1)) = \gamma_0 + \alpha_i + \eta_j \tag{2}$$

η_j is estimated using problem features (P_{fj}) and random intercepts for each problem (μ_j). Let P_{fj} be problem-level feature f for problem j.

$$\eta_j = \sum_{f=1}^{F} P_{fj}\nu_f + \mu_j \tag{3}$$

We model α using student-level variables. Let C_{ki} be covariate predictor k of K student-level predictors, which are measured at baseline for both $\mathscr{T}$ and $\mathscr{C}$. Let the random intercepts be π_i for students, η_t for teachers, and ρ_s schools, each modeled as independent and with normal distributions with means of zero and standard deviations estimated from the data.

$$\alpha_i = \sum_{k=1}^{K} \gamma_k \mathrm{C}_{ki} + \pi_i + \eta_t + \rho_s \tag{4}$$

The MCMC procedure imputes α_{ti} for all students not assigned to FH2T with random draws from a normal distribution with mean $\sum_k \gamma_k \mathrm{P}_{ki} + \eta_{t[i]} + \rho_{s[i]}$, where $\eta_{t[i]}$ and $\rho_{s[i]}$ are the random intercepts for student i's teacher and school, respectively, and standard deviation equal to the estimated standard deviation of π_i. The student level randomization (*i.e.* teachers had students in different conditions in their classes) allows us to include the random intercepts for schools and teachers from Equation 4. However, π_i is unknown for students not in FH2T, but we can assume its distribution is the same in all conditions because of the randomization.

Before fitting the full FLPS model, we used bidirectional stepwise regression production to select student and problem-level predictors. Next, we validated the measurement model using a ten-cross-fold validation procedure. Appendix A provides a full description of the cross-validation procedure. The area under the receiver operating characteristic curve (AUC) – which measures the balance between true positive and false positive predictions at different cut-point thresholds – was used to assess model performance. A minimum AUC of 0.65 is acceptable in FLPS measurement models based on prior simulations (Sales et al., 2022).

Outcomes Submodel

The outcome submodel interacts students' estimated propensities to reattempt problems (α_{ti}) with the condition indicators, thus estimating the association between α_{ti} and the effect of each condition. Let Y_i be the learning outcome for student i. Let λ_k be the fixed effects for covariates C_{ki}. Let the random effects for a teacher be ρ_t and for school be ψ_s. Let $Z_i^{\mathscr{T}}$ be dummy code variables for student i in condition $\mathscr{T}$. Note that the Active Control is the reference category.

$$Y_i = \beta_0 + \beta_1 \alpha_{ti} + \sum_{\mathscr{T} \in \{FH2T, DB, IC\}} Z_i^{\mathscr{T}} (\beta_2^{\mathscr{T}} + \beta_3^{\mathscr{T}} \alpha_{ti}) + \sum_{k=4}^{K} \beta_k \mathrm{C}_{ki} + \rho_t + \psi_s + \varepsilon_i \tag{5}$$

Together, submodels 2 and 5 formed an FLPS model, which we fit using the Stan Markov Chain Monte Carlo software (Arezooji, 2020). We evaluated convergence using trace plots and whether the $\hat{R}$, which measures convergence by comparing between- and within-chain estimates of each parameter, was below the recommended threshold of 1.05 (Vehtari et al., 2021). The maximum $\hat{R}$ for the model parameters was 1.03.

Descriptive statistics of the posterior distributions of β_0^Z and β_1^Z are used to evaluate the main effects of each condition and moderating effects of students' propensity to reattempt problems (Research Question 2). Comparisons of these distributions are used to address whether the moderator effects consist of gamified conditions (Research Question 3).

Results

Reattempt Behavior Exploratory Data Analysis

Table 2 presents frequencies of students' reattempt behavior. In FH2T, students reattempted 14,546 problems, averaging 13% of all the problems they completed. Students most commonly attempted problems after sub-optimal performance (n=8973; 61.69% of total reattempts). When students performed sub-optimally, they reattempted the problem just under one-third of the time (30.74%). A substantial minority of the reattempts occurred after students performed optimally (n=5573, 38.31% of total reattempts). Yet, this behavior occurred infrequently compared to the total number of optimal performances (6.67%). Since this paper concentrates on productive persistence, we focus only on instances of reattempt after sub-optimal performance.

Table 2

Reattempt Frequencies

Reattempt	n	Percent
All Problems	14546	13.00%
After sub-optimal Attempts	8973	30.74%
After Optimal Attempts	5573	6.74%

Figure 4 displays this relation between the total number of problems students completed and the percent of sub-optimal problems that they reattempted, along with a box plot of the reattempt behavior. On average, students tend to reattempt problems after they perform sub-optimally slightly less than a quarter of the time (median = 23.33). However, the distribution of reattempts is positively skewed, with just 25% of students attempting these problems between 56.25% and 100.00% of the time. Although it might seem logical that spending time attempting problems would lead to less content covered, there was a small positive but significant correlation (*r*= 0.09, *p* = 0.002) between the number of problems completed and the percentage of problems reattempted. Thus, if anything, attempting problems after sub-optimal performance is associated with covering more, not less, content.

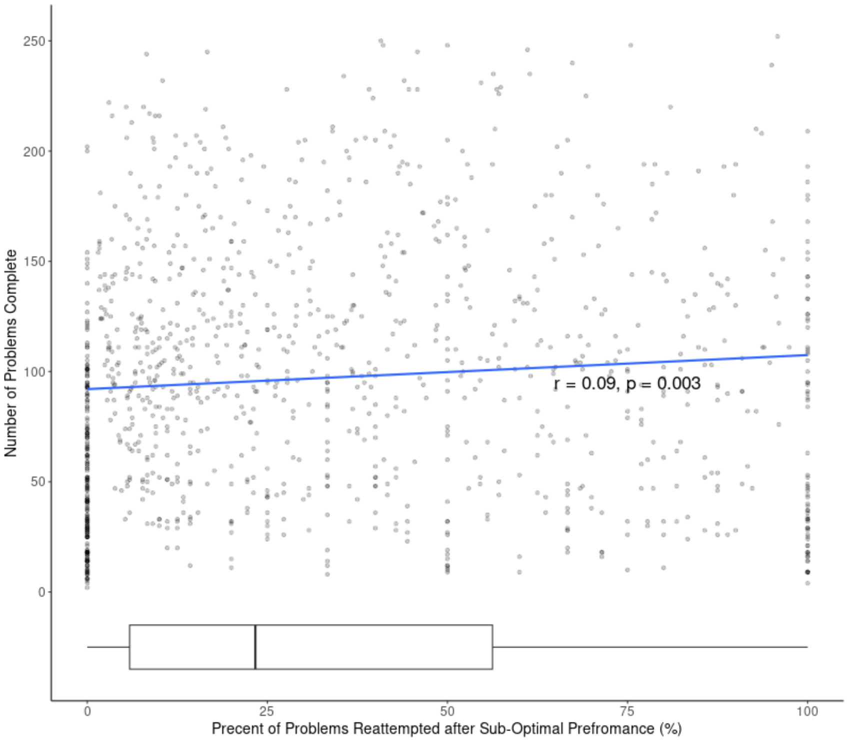


Figure 4

Scatter plot of the relation between the present of problems each student reattempted after sub-optimal performance and the total number of problems the students completed.

Associations with Performance (Research Question 1)

As an initial test of whether students may benefit from reattempting problems, we evaluated whether this behavior was associated with improved performance on subsequent problems. Table 3 presents frequencies of the next problem performance based on whether students reattempt problems. The table only includes instances in which students performed sub-optimally on the problem. 70.64% of students who reattempted problems after sub-optimal performance performed optimally on the next problem, compared with 58.24% of those who did not reattempt the problem. This difference was statistically significant (χ^2 398.97, DF = 1, $p <$ 0.001).

Table 3

Probability of next problem optimal by reattempt behavior after sub-optimal performance

		Next Problem	
	n	Sub-Optimal	Optimal
Did not reattempt	19756	41.76%	58.24%
Reattempt	8849	29.36%	70.64%

To ensure that this result was not explained by students' prior knowledge of the content or any predisposition toward math, we ran a logistic regression predicting whether students performed optimally on the next problem after optimal performance. We included students' pretest algebraic knowledge and math anxiety as well as random intercepts for both the student and the problem. Table 4 presents the model estimates. Students were significantly more likely to perform optimally on the next problem after a sub-optimal performance when reattempting the problem compared to when they did not reattempt it (β_1 = 0.46, SE = 0.05, p < 0.001), after accounting for student pretest scores and baseline probabilities for optimal performance for the student and the problem. Overall, this increased likelihood of performance improvement associated with reattempting problems suggests that the behavior is an example of predictive persistence.

Table 4

Predicting Next Problem Optimal after Sub-optimal Performance

	Estimate (Log Odds)	SE	Z-value	P-value
Intercept	0.62	0.14	4.38	< 0.001
reattempt	0.46	0.05	10.24	< 0.001
Pretest Algebraic Knowledge	0.22	0.03	8.11	< 0.001
Pretest Math Anxiety	-0.03	0.03	-1.22	0.222

Random Intercepts	SD	Variance
Student	0.18	0.42
Problem	4.25	2.06

Fully Latent Principal Stratification Model

Measurement Submodel

Prior to estimating the full FLPS model, we calculated the intra-class correlations (ICC) for different model components and conducted the cross-validation procedure described in the Measurement Model Section. Table 5 presents the variances and ICCs associated with the student (α) and the problem (η). These are estimated by running a null model with only random intercepts for the student and problem. Most of the variance in reattempt behavior is associated with the student (52%), while a substantial portion is associated with the problem (19%). This suggests that reattempting problems is largely driven by individual student differences.

Table 5

Measurement model intra-class correlations

	Variance	ICC
α (Student)	5.99	0.52
η (Problem)	2.19	0.19

The estimated model parameters are presented in Table 6. In Bayesian analyses, parameter estimates and their uncertainty are derived from the descriptive statistics of the posterior distribution. Thus, the mean of the posterior distribution is the point estimate, and the standard deviation (SD) functions similarly to a standard error in frequentist statistics. The credible interval (CI) is the range of the position distribution that falls within the 95% probability distribution and can be interpreted similarly to confidence intervals in frequentist statistics. One benefit of Bayesian methods is the ability to calculate the probability that a point estimate is different from zero, which is indicated in the table.

When interpreting the measurement model, we only consider parameters that have at least a 90% probability of being either greater or smaller than zero. Note that the covariates in the model were selected to create the best predictive model and not to explain the reattempting problem behavior. Many variables are highly correlated with one another. Since this collinearity may influence the parameter estimates and their uncertainty, our interpretations are made with caution.

Overall, higher-knowledge students were more likely to reattempt problems after sub-optimal performance. Of the student-level covariates, students' pretest algebraic flexibility (γ_3 = 0.23, CI = 0.06 –– 0.40) and math state test performance (γ_5 = 0.23, CI = 0.02 –– 0.46) had positive associations with reattempt behavior. This suggests that students' abilities to think flexibly about algebraic equations may be key in determining whether a student tries a different method by attempting a problem. The positive correlation between math state test performance and reattempt behaviors may indicate that general math knowledge may play a key role in

Table 6

Fully Latent Principal Stratification Model Parameter Estimates

	Measurement Submodel			Outcomes Submodel		
	Estimate	SD	CI	Estimate	SD	CI
Intercept	-2.13*	0.42	-2.82 – -1.43	-0.08	0.10	-0.24 – 0.09
α				-0.10*	0.05	-0.18 – -0.02
Immediate Condition				0.07°	0.05	-0.01 – 0.16
Immediate Condition*α				0.07°	0.05	-0.02 – 0.15
FH2T				0.08°	0.05	-0.01 – 0.15
FH2T*α				0.14*	0.05	0.07 – 0.22
DragonBox				0.17*	0.05	0.08 – 0.26
DragonBox*α				0.07°	0.05	-0.02 – 0.16
Algebraic Procedural Knowledge	0.11	0.11	-0.07 – 0.29	0.05*	0.02	0.02 – 0.09
Algebraic Conceptual Knowledge	0.17°	0.13	-0.04 – 0.39	0.24*	0.03	0.2 – 0.28
Algebraic Flexibility	0.23*	0.11	0.06 – 0.40	0.04*	0.02	0.01 – 0.08
Perceptual Sensitivity	0.40*	0.12	0.20 – 0.60	0.09*	0.03	0.05 – 0.13
State Math Test	0.23*	0.13	0.02 – 0.46	0.18*	0.03	0.14 – 0.23
State Math Test (Squared)	0.01	0.07	-0.10 – 0.12	0.00	0.01	-0.02 – 0.03
Pretest Time (Log)	-0.02	0.11	-0.20 – 0.16	0.02	0.02	-0.01 – 0.05
Pretest Items Complete	0.01	0.12	-0.20 – 0.21	0.02	0.02	-0.02 – 0.05
Math Anxiety	-0.13	0.15	-0.37 – 0.11	-0.02	0.03	-0.07 – 0.03
Math Anxiety (Squared)	0.12*	0.07	0.00 – 0.23	0.01	0.01	-0.01 – 0.03
Numeric Inconfidence	0.17°	0.14	-0.05 – 0.40	0.02	0.03	-0.02 – 0.06
Math Self Efficacy	-0.05	0.12	-0.25 – 0.14	0.05*	0.02	0.02 – 0.09
Gifted	0.14°	0.10	-0.02 – 0.31	0.08*	0.02	0.05 – 0.11
English Speakers of Other Languages	-0.16*	0.09	-0.31 – -0.01	0.01	0.02	-0.02 – 0.04
Virtual Instruction	-0.11	0.16	-0.37 – 0.15	0.11	0.10	-0.06 – 0.27
Problem Complexity	-0.34*	0.06	-0.43 – -0.25			
Problem Sequence within Problem Set	-0.39*	0.04	-0.46 – -0.32			

** 95% CI excludes 0*

° 90% CI excludes 0

students' decisions to attempt problems. Other assessments of students' prior knowledge were also positively correlated with the reattempt behavior, including algebraic conceptual knowledge (γ_2 = 0.17, CI = -0.04 — 0.39), perceptual sensitivity (γ_4 = 0.40, CI = 0.20 -– 0.60), and whether students were labeled “gifted" (γ_{13} = 0.14, CI = -0.02 — 0.31) by the district.

Although higher-performing students are more likely to reattempt problems, the profile of students with a high propensity for this behavior may be more nuanced. While math anxiety was not likely associated with reattempting problems, the quadratic transformation of math anxiety was positively associated with the behavior (γ_{10} = 0.12, CI = 0.00 – 0.23). This suggests that students with the highest levels of math anxiety are more likely to reattempt problems. Similarly, students with higher numerical inconfidence were also more likely to reattempt problems (γ_{11} = 0.17, CI = -0.05 – 0.40). Taken as a whole, the measurement model suggests that higher-performing students with higher math anxiety, specifically in numerical inconfidence, are most likely to reattempt problems.

The problem-level parameters give us some understanding of the context in which students reattempt problems. On average, students were less likely to reattempt more complex problems (ν_1 = -0.34, CI = -0.43 –– -0.25) and those that came later in the problem set (ν_2 = -0.39, CI = -0.46 –– -0.32).

Figure 5 presents the relationship between students' estimated propensity to reattempt problems after sub-optimal performance (α_{ti}) and the percentage of problems that they reattempt problems after sub-optimal performance. The plot shows a close but imperfect relationship between the two variables. The size of the points represents the number of problems in which the student performed sub-optimally. Notably, students with sparse data tend to have α_{ti} values closer to zero than those with more data and the same percentage of problems reattempted. This suggests the measurement model appropriately regularizes students' propensities based on the number of available observations.

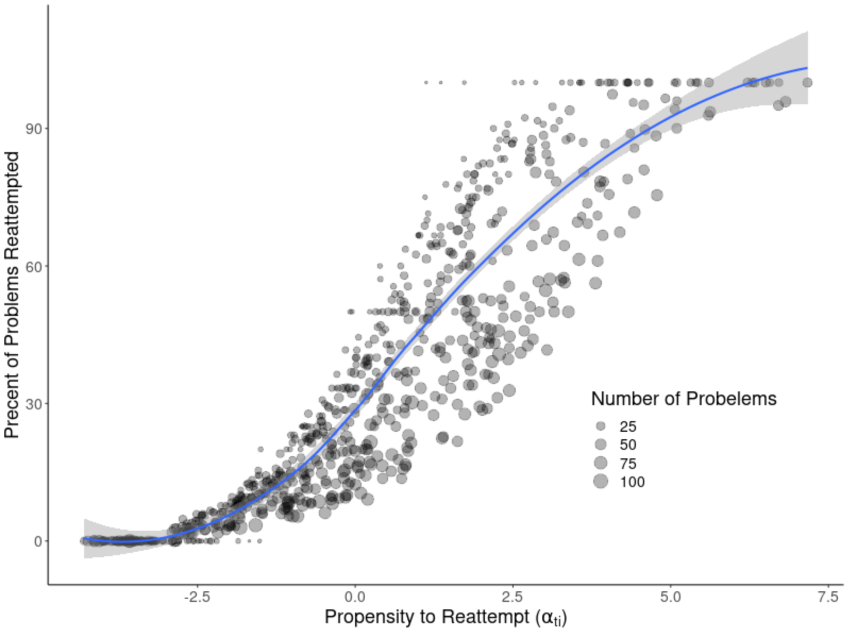


Figure 5

Relationship between students' estimated propensity to reattempt problems after sub-optimal performance (α_{ti}) and the percentage of problems that they reattempt problems after sub-optimal performance.

Outcomes Submodel

The parameters of the outcomes submodel are presented in the rightmost columns of Table 6. The association between students' propensity to reattempt problems and their post-test algebraic knowledge was likely negative (β_2 = -0.10, SD = 0.05). Thus, higher propensities to reattempt problems were not indicative of higher post-test performance for students who did not have access to features that allowed them to reattempt problems (i.e., delayed condition). This finding is particularly interesting because the behavior was positively associated with pre-test math performance measures.

The main effects of each condition (β_2^Z) are the expected effects for students with average reattempt behavior ($\alpha_t = 0$). All of the conditions had likely positive main effects. It is highly probable that DragonBox (β_2^{DB} = 0.17, CI = 0.08 –- 0.26) positively impacted algebraic knowledge for students with average propensities to reattempt problems. It is also probable that

FH2T (β_2^{FH2T} = 0.08, CI = -0.02 — 0.15) and Immediate Condition (β_2^{IC} = 0.07, CI = -0.01 — 0.16) had positive main effects.

The interaction effects indicate whether students' propensities to reattempt problems moderates the effect of CBLPs (Research Question 2). It is highly probable that interaction between FH2T and the propensity to reattempt problems was positive (β_3^{FH2T} = 0.14, CI = 0.07 — 0.22). The interaction effects between the propensity to reattempt and the Immediate Condition (β_3^{IC} = 0.07, CI = -0.02 — 0.15) and DragonBox (β_3^{DB} = 0.07, CI = 10.02 — 0.16) were likely positive, but smaller in magnitude – the estimates were half the size of the FH2T interaction. Figure 6 visualizes these slopes by presenting the effect sizes at different levels of α_t for each condition.

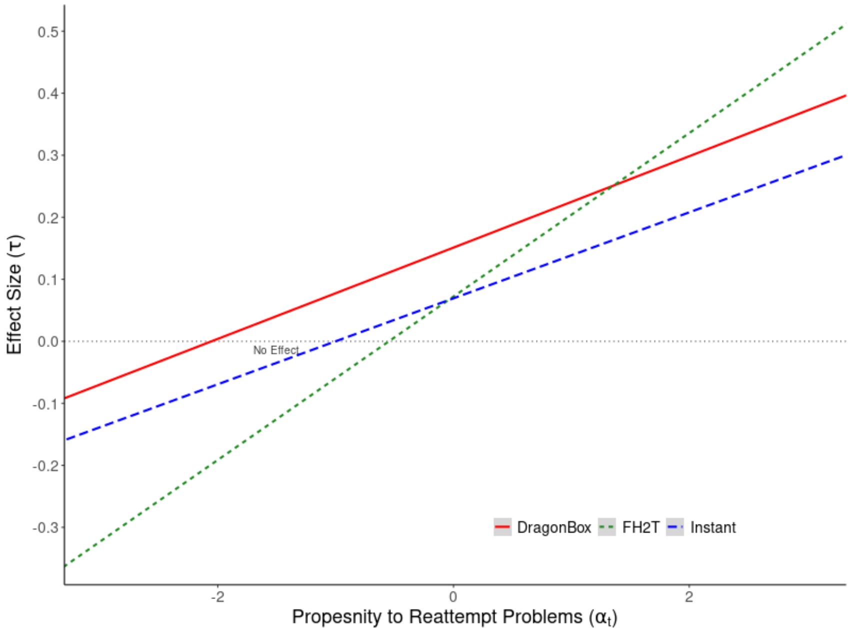


Figure 6

Plot of effect sizes (τ) by propensity to reattempt problems (α).

Table 7 presents the differences in main and interaction effects between conditions. Notably, DagonBox had greater main effects than both FH2T ($\Delta\tau$ = 0.10, $P(> 0)$ = 0.96) and the Immediate Condition ($\Delta\tau$ = 0.10, $P(> 0)$ = 0.95). To address whether the moderating association differs between gamified and non-gamified conditions (Research Question 3), we compare the posterior distributions of each interaction coefficient to one another. The effect of FH2T likely varied more based on students' propensities to reattempt problems compared with DragonBox ($\Delta\tau = 0.06$, $P(> 0) = 0.91$) and the Immediate Condition ($\Delta\tau = 0.07$, $P(> 0) = 0.93$). The FLPS model did not provide evidence that there was a substantial difference in the interactions between DragonBox and Immediate Condition ($\Delta\tau = 0.01$, $P(> 0) = 0.58$).

Table 7

Difference between Main and Interaction effects across conditions

	Main Effect			Interaction Effect		
Comparisons	$\Delta\tau$	$P(>0)$	$P(<0)$	$\Delta\tau$	$P(>0)$	$P(<0)$
FH2T vs DragonBox	-0.10	0.04	0.96	0.06	0.91	0.09
FH2T vs Immediate	< 0.01	0.51	0.49	0.07	0.93	0.07
DragonBox vs Immediate	0.10	0.95	0.05	0.01	0.58	0.42

Discussion

This study finds that students' tendency towards productive persistence may be essential to whether they reap the benefits of learning platforms; however, having a high propensity to persist may be insufficient for learning from CBLP. Differences in these learning environments that allow students to display productive persistence likely influence whether students learn from innovative programs in algebra education. Thus, it is paramount that when designing learning platforms, we consider the opportunities that these learning experiences provide for students to respond to their sub-optimal performance. Providing opportunities for multiple attempts at a problem and even multiple paths to a solution may allow students to struggle through misconceptions and learn from mistakes. More broadly, our analyses highlight the need to understand how students' behavioral tendencies can influence whether innovative educational technologies effectively teach students.

These analyses suggest that reattempt behavior is a particularly salient action for students' performance in the CBLPs and their benefit from the platforms. Taken as a whole, the finding that students are more likely to improve performance after attempting problems, along with the positive moderation effects with each of the treatments, indicate that reattempting problems is often an example of productive performance. This finding adds causal evidence to the previous studies showing that productive persistence behaviors in CBLPs are positively associated with short and long-term academic outcomes (Adjei et al., 2021; Ventura et al., 2013). However, the negative association between this propensity and post-test performance for students in the Active Control suggests that having the propensity to persist without access to the ability to reattempt problems is likely determinantal to learning; students needed the opportunity to persist through specific CBLP features to benefit from their latent tendencies.

The analysis also provides insights into which student characteristics are predictive of productive persistence and which perspectives and abilities may be barriers for students to display the behavior. Students' algebraic flexibility was particularly predictive of reattempting problems after sub-optimal performance. This association indicates that students who did not display the behavior may struggle with the mental flexibility required to produce multiple solutions. Attempts to foster flexibility by explicitly having students produce multiple solutions to algebra problems have produced promising results (Newton et al., 2010; Star & Seifert, 2006). Incorporating these practices into algebra CBLPs like FH2T and DragonBox might help programs increase students' algebraic flexibility, thus improving their chances of reattempting problems.

More general student characteristics might also contribute to this behavior. For example, the "need for cognition" – the intrinsic drive to rationally understand one's experiences and world – may drive some students to persist during learning tasks (Cacioppo & Petty, 1982). Similarly, epistemic curiosity, in which students find pleasure in finding solutions to academic problems, might explain persistence behaviors (Litman, 2008). However, the corrections between the reattempt behavior and students' math anxiety and lack of numeric inconfidence suggest that curiosity as a feeling of deprivation may be a greater driver of the reattempt behavior (Litman & Jimerson, 2004). At times, students experience curiosity as a frustrating feeling of a lack of understanding instead of the pleasurable pursuit of knowledge. In our case, students with heightened math anxiety and low confidence might respond with curiosity due to irritation at their inadequate response. Future work to tease apart the student's underlying motivations for this behavior might help produce more generalizable predictions of persistence behaviors.

It is also notable that as students progressed through the problem sets and the content became more complex, they were less likely to reattempt problems. This finding suggests that they may experience some fatigue in persistence. Future work should evaluate whether nudging students towards reattempting more complex problems is both effective at changing behavior and productive at improving learning outcomes. Alternatively, there could be an effect of diminished returns where reattempting many more problems does not contribute to the effect of the program. Thus, more research in this area is required.

Although all three treatment conditions likely moderated students' propensity to persist productively, the unequal moderating relationships across programs suggest that differences in the programs' instructional designs may cause various behavioral manifestations of the persistence propensity. In contrast to our hypothesis, the two gamified conditions had different magnitudes of effect heterogeneity. FH2T had a more substantial effect heterogeneity based on persistence than the other conditions. This finding suggests that specific program features, rather than general gamification, may drive efficacy patterns. Although propensity to persist had similar moderating effects on DragonBox and the Immediate Condition, these effects likely have different antecedents due to the differences in the programs.

FH2T was most effective for students with a high propensity to persist, and students with a low propensity to persist would likely have benefited from the Active Control condition in which they did not have any opportunities to reattempt problems. Notably, students with average reattempt behavior and average pretest covariates likely experience only marginal effects of the program – less than a tenth of a standard deviation on the post-test. These findings suggest that the program's efficacy may be contingent upon students engaging in productive persistence by reattempting problems after sub-optimal performance.

The larger moderation effect of reattempt propensity for FH2T compared with the other platforms may be attributable to the prominence of the “retry" button on the reward screen when informing students of sub-optimal performance. Notably, DragonBox's reattempt button did not have the same prominence. It's possible that students in DragonBox would have reattempted more problems and experienced larger effects from the program had the option to reattempt problems been more conspicuous.

Overall, DragonBox had a more substantial effect on students with average reattempt behavior and pretest covariates than the other platforms. DragonBox also showed smaller effect heterogeneity compared with FH2T. Only the students with the lowest propensities to persist likely experienced negative effects from this program. The lower heterogeneity may be due to the differences in gamification features between DragonBox and FH2T. This inference is consistent with previous research showing that gamification is often insufficient to explain behavioral and learning differences between programs (Sailer & Homner, 2020; Sailer et al., 2017). Students with a lower propensity to persist in FH2T may have been more motivated to replay problems in DragonBox, due to narrative features and the puzzle-like presentation of algebraic problems. Furthermore, DragonBox's use of non-mathematical symbols to scaffold students' learning of algebraic concepts may have helped students with lower propensities to persist still benefit.

Limitations

Although we provide evidence of effect heterogeneity associated with students' persistence tendencies, the analysis has some limitations. There remains uncertainty regarding whether some of the main and interaction effects in the model significantly differ from zero. Overall, for the estimates we considered likely to be different from zero, there was at most a 10% posterior probability that the true parameter had the opposite sign compared to the estimate presented in this paper. Thus, it is important that future research substantiate these findings to increase the certainty of their veracity.

Furthermore, there are some limitations in the interpretation of the FLPS results. First, although we are making inferences about how the reattempt behavior may influence the efficacy of a program, the moderation model does not address this directly. Rather, FLPS indicates whether an effect varies based on an estimated propensity of the behavior. From the study design, we infer that some conditions enabled students to engage in persistent behaviors more than others, but this difference is not measured directly. However, it is possible that our measure of students' propensity to reattempt problems actually assesses another latent construct by which the effect sizes also vary. Future studies could measure these differences in reattempt behavior in similar programs and medication analyses to evaluate how the behavior change may cause outcome differences.

Finally, it is important to acknowledge the constraints inherent to FLPS. Although this method has been employed in both empirical and simulation studies, which support the method's robustness (Sales & Pane, 2021; Sales et al., 2022), the estimated impacts using FLPS heavily depend on the inherent quality of the model itself. It remains unclear how inaccuracies in estimating the model could potentially lead to biased outcomes. Future research should focus on gaining a more accurate understanding of how to assess these models to confirm that they furnish impartial appraisals of intervention effects.

Conclusion

In conclusion, the findings of this study underscore the critical interplay between students' persistence tendencies, program institutional designs, and efficacy. For students to merely possess a high propensity to persist is not sufficient; the design of the CBLP must facilitate opportunities for students to engage in productive behaviors such as reattempting problems. Therefore, educators and platform designers should incorporate features that actively encourage students to persist through challenges, providing multiple attempts and diverse pathways to problem-solving.

Furthermore, the study highlights the pressing need for a nuanced approach to designing educational technologies. The efficacy of CBLPs is not uniform; rather, it is moderated by how these platforms cater to persistent behaviors. For instance, FH2T's prominent reattempt feature showed a significant advantage for students with high persistence, whereas DragonBox's narrative-driven, puzzle-like approach benefited a broader range of students, including those with lower persistence tendencies. These insights challenge us to design educational technologies that account for students' diverse behavioral tendencies, engaging us in the process of creating more effective learning platforms.

More generally, our study shows the importance of understanding how efficacy may vary by a population's behavioral tendencies. Analysis of behavior-moderated effects like this one holds promise for allowing research beyond assessing whether some populations benefit more than others by pointing toward which aspects of the programs may drive students toward positive learning behaviors. Conversely, analysis like these can allow us to understand the potential barriers to these behaviors that must be overcome to create programs that impact students equitably.

References

Abdelrahman, G., Wang, Q., & Nunes, B. P. (2022). Knowledge tracing: A survey. (arXiv:2201.06953). http://arxiv.org/abs/2201.06953

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The future of embodied design for mathematics teaching and learning. *Frontiers in Education*, *5*. https://www.frontiersin.org/articles/10.3389/feduc.2020.00147

Adjei, S. A., Baker, R. S., & Bahel, V. (2021). Seven-year longitudinal implications of wheel spinning and productive persistence. *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I*, 16–28.

Arezooji, D. M. (2020). *A markov chain monte-carlo approach to dose-response optimization using probabilistic programming (rstan)*.

Bloom, B. S. (1968). Learning for mastery. instruction and curriculum. regional education laboratory for the carolinas and virginia, topical papers and reprints, number 1 [issue: 2 container-title: Evaluation Comment volume: 1 ERIC Number: ED053419]. *Evaluation Comment*, *1*(2). https://eric.ed.gov/?id=ED053419

Boedeker, P., & Henson, R. K. (2020). Evaluation of heterogeneity and heterogeneity interval estimators in random-effects meta-analysis of the standardized mean difference in education and psychology. *Psychological Methods*, *25*(3), 346.

Botelho, A. F., Van Inwegen, E. G., Varatharaj, A., & Heffernan, N. T. (2019). Refusing to try: Characterizing early stopout on student assignments. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, 391–400. https://doi.org/10.1145/3303772.3303806

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(88), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bye, J., Lee, J., Chan, J., Closser, A., Shaw, S., & Ottmar, E. (2022). Perceiving precedence: Order of operations errors are predicted by perception of equivalent expressions. *Poster presented at the annual meeting of the American Educational Research Association (AERA)*.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cayton-Hodges, G. A., Feng, G., & Pan, X. (2015). Tablet-based math assessment: What can we learn from math apps? *Journal of Educational Technology Society*, *18*(2), 3–20.

Cheung, A. C. K., & Slavin, R. E. (2011). The effectiveness of education technology for enhancing reading achievement: A meta-analysis. *Center for Research and Reform in Education*. https://eric.ed.gov/?id=ed527572

Chiu, L.-H., & Henry, L. L. (1990). Development and validation of the mathematics anxiety scale for children. *Measurement and evaluation in counseling and development*, *23*(3), 121–127.

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *User modeling 2001* (pp. 137–147). Springer. https://doi.org/10.1007/3-540-44566-8_14

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/BF01099821

Decker-Woodrow, L. E., Mason, C. A., Lee, J.-E., Chan, J. Y.-C., Sales, A., Liu, A., & Tu, S. (2023a). The impacts of three educational technologies on algebraic understanding in the context of covid-19. *AERA Open*, *9*, 23328584231165919. https://doi.org/10.1177/23328584231165919

Decker-Woodrow, L. E., Mason, C. A., Lee, J.-E., Chan, J. Y.-C., Sales, A., Liu, A., & Tu, S. (2023b). The impacts of three educational technologies on algebraic understanding in the context of covid-19. *AERA open*, *9*, 23328584231165919.

Denden, M., Tlili, A., Chen, N.-S., Abed, M., Jemni, M., & Essalmi, F. (2022). The role of learners' characteristics in educational gamification systems: A systematic meta-review of the literature. *Interactive Learning Environments*, 1–23.

Dieter, K. C., Studwell, J., & Vanacore, K. (2020). Differential responses to personalized learning recommendations revealed by event-related analysis. *International Conference on Educational Data Mining (EDM)*, *13*. https://eric.ed.gov/?id=ED607826

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Fadda, D., Pellegrini, M., Vivanet, G., & Zandonella Callegher, C. (2022). Effects of digital games on student motivation in mathematics: A meta-analysis in k-12. *Journal of Computer Assisted Learning*, *38*(1), 304–325. https://doi.org/10.1111/jcal.12618

Feng, M., Huang, C., & Collins, K. (2023). *Technology-based support shows promising long-term impact on math learning: Initial results from a randomized controlled trial in middle schools*. https://www.wested.org/wp-content/uploads/2023/07/ASSISTments-Long-Term-Effects-_07-11-23_FINAL-ADA.pdf

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29.

Gallop. (2019). Education technology use in schools. https://www.newschools.org/wp-content/uploads/2020/03/Gallup-Ed-Tech-Use-in-Schools-2.pdf

Gaston, J., & Cooper, S. (2017). To three or not to three: Improving human computation game onboarding with a three-star system. *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*, 5034–5039.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, *2*(2), 265–284. https://doi.org/10.1111/j.1756-8765.2009.01055.x

Goldstone, R. L., Marghetis, T., Weitnauer, E., Ottmar, E. R., & Landy, D. (2017). Adapting perception, action, and technology for mathematical reasoning. *Current Directions in Psychological Science*, *26*(5), 434–441. https://doi.org/10.1177/0963721417704888

Gray, L., & Lewis, L. (2021). Use of educational technology for instruction in public schools: 2019–20. https://nces.ed.gov/pubs2021/2021017Summary.pdf

Gurung, A., Vanacore, K., Mcreynolds, A. A., Ostrow, K. S., Worden, E., Sales, A. C., & Heffernan, N. T. (2024). Multiple choice vs. fill-in problems: The trade-off between scalability and learning. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 507–517. https://doi.org/10.1145/3636555.3636908

Heffernan, N. T., & Heffernan, C. L. (2014). The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Huang, Y., Dang, S., Elizabeth Richey, J., Chhabra, P., Thomas, D. R., Asher, M. W., Lobczowski, N. G., McLaughlin, E. A., Harackiewicz, J. M., Aleven, V., & Koedinger, K. R. (2023). Using latent variable models to make gaming-the-system detection robust to context variations. *User Modeling and User-Adapted Interaction*, *33*(5), 1211–1257. https://doi.org/10.1007/s11257-023-09362-1

Hyde, D. C., Liu, A., Sella, F., Prado, J., & Vanacore, K. (2022). Developmental perspectives: Digital interventions and mathematics learning in typical and atypical populations. *Proceedings of 2022 International Mind, Brain and Educations Society Conference*, 19.

Jagušt, T., Botički, I., & So, H.-J. (2018). Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Computers & education*, *125*, 444–457.

Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, *103*(481), 101–111.

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining*, *10*(11), 36–71. https://doi.org/10.5281/zenodo.3344810

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education [arXiv:2007.05443 [cs]]. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education*. Taylor Francis. http://arxiv.org/abs/2007.05443

Kluge, A., & Dolonen, J. (2015). Using mobile games in the classroom: The good and the bad of a new math language. In *Mobile learning and mathematics* (pp. 106–121). Routledge.

Kozierkiewicz-Hetmańska, A., & Nguyen, N. T. (2010). A computer adaptive testing method for intelligent tutoring systems. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 281–289.

Kuchipudi, N. (2024). *Various design projects in the assistments foundation* [Doctoral dissertation, Worcester Polytechnic Institute].

Kulik, C.-L., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*(2), 265–299.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming*, *45*(6), 752–768.

Lavoué, E., Ju, Q., Hallifax, S., & Serna, A. (2021). Analyzing the relationships between learners' motivation and observable engaged behaviors in a gamified learning environment. *International Journal of Human-Computer Studies*, *154*, 102670.

Lee, J.-E., Chan, J. Y.-C., Botelho, A., & Ottmar, E. (2022). Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational technology research and development*, *70*(5), 1575–1599.

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*(5), 635–652. https://doi.org/10.1080/03054985.2012.734800

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, *44*(7), 1585–1595. https://doi.org/10.1016/j.paid.2008.01.014

Litman, J. A., & Jimerson, T. L. (2004). The measurement of curiosity as a feeling of deprivation. *Journal of Personality Assessment*, *82*(2), 147–157. https://doi.org/10.1207/s15327752jpa8202_3

Liu, A., Vanacore, K., & Ottmar, E. (2022). How reward-and error-based feedback systems create micro-failures to support learning strategies. *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 1633-1636*.

Liu, Z., Cody, C., & Barnes, T. (2017). The antecedents of and associations with elective replay in an educational game: Is replay worth it? *Proceedings of the 10th International Conference on Educational Data Mining*.

Long, Y., & Aleven, V. (2017). Educational game and intelligent tutoring system: A classroom study and comparative design analysis. *ACM Transactions on Computer-Human Interaction*, *24*(3), 1–27. https://doi.org/10.1145/3057889

Malkiewich, L. J., Lee, A., Slater, S., Xing, C., & Chase, C. C. (2016). No lives left: How common game features could undermine persistence, challenge-seeking and learning to program. https://repository.isls.org//handle/1/115

McClelland, D. C. (1961). *Achieving society* (Vol. 92051). Simon; Schuster.

Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L. H., Anderman, E., & Roeser, R. (2013, April). Patterns of adaptive learning scales [Institution: American Psychological Association DOI: 10.1037/t19870-000]. https://doi.org/10.1037/t19870-000

Newton, K. J., Star, J. R., & Lynch, K. (2010). Understanding the development of flexibility in struggling algebra students. *Mathematical Thinking and Learning*, *12*(4), 282–305. https://doi.org/10.1080/10986065.2010.482150

of Education, G. D. (2020).

O'Rourke, E., Peach, E., Dweck, C. S., & Popovic, Z. (2016). Brain points: A deeper look at a growth mindset incentive structure for an educational game. *Proceedings of the third (2016) acm conference on learning@ scale*, 41–50.

Ottmar, E., Lee, J.-E., Vanacore, K., Pradhan, S., Decker-Woodrow, L., & Mason, C. A. (2023). Data from the efficacy study of from here to there! a dynamic technology for improving algebraic understanding. *Journal of Open Psychology Data*, *11*(1), 5. https://doi.org/10.5334/jopd.87

Owen, V. E., Roy, M.-H., Thai, K. P., Burnett, V., Jacobs, D., & Baker, R. S. (2019). Detecting wheel-spinning and productive persistence in educational games. *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019).*

Page, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*, *5*(3), 215–244.

Page, L. C., Feller, A., Grindal, T., Miratrix, L., & Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, *36*(4), 514–531. https://doi.org/10.1177/1098214015594419

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127–144. https://doi.org/10.3102/0162373713507480

Pane, J. F., McCaffrey, D. F., Slaughter, M. E., Steele, J. L., & Ikemoto, G. S. (2010). An experiment to evaluate the efficacy of cognitive tutor geometry. *Journal of Research on Educational Effectiveness*, *3*(3), 254–281. https://doi.org/10.1080/19345741003681189

Pane, J. F., Seaman, D., & Doss, C. J. (2023, September). *Students using lexia® core5® reading show greater reading gains than matched comparison students*. https://www.rand.org/pubs/research_reports/RRA2859-1.html

Park, S. (2023). Discovering unproductive learning patterns of wheel-spinning students in intelligent tutors using cluster analysis. *TechTrends*, *67*(3), 489–497. https://doi.org/10.1007/s11528-023-00847-9

Pradhan, S., Gurung, A., & Ottmar, E. (2024). Gamification and deadending: Unpacking performance impacts in algebraic learning. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 899–906. https://doi.org/10.1145/3636555.3636929

Prihar, E., Sales, A., & Heffernan, N. (2023). A bandit you can trust. *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 106–115.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Richey, J. E., Zhang, J., Das, R., Andres-Bray, J. M., Scruggs, R., Mogessie, M., Baker, R. S., & McLaren, B. M. (2021). Gaming and confrustion explain learning advantages for a math digital learning game. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *In artificial intelligence in education: 22nd international conference* (pp. 342–355). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_28

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, *2*(4), 2332858416673968. https://doi.org/10.1177/2332858416673968

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in human behavior*, *69*, 371–380.

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Sales, A. C., & Pane, J. F. (2019). The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics*, *13*(1), 420–443. https://doi.org/10.1214/18-AOAS1196

Sales, A. C., & Pane, J. F. (2021). Student log-data from a randomized evaluation of educational technology: A causal case study. *Journal of Research on Educational Effectiveness*, *14*(1), 241–269.

Sales, A. C., Vanacore, K., & Ottmar, E. R. (2022). Geepers: Principal stratification using principal scores and stacked estimating equations. *arXiv preprint arXiv:2212.10406*.

Schudde, L. (2018). Heterogeneous effects in education: The promise and challenge of incorporating intersectionality into quantitative methodological approaches. *Review of Research in Education*, *42*(1), 72–92.

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., Ventura, M., & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers Education*, *86*, 224–235. https://doi.org/10.1016/j.compedu.2015.08.001

Siew, N. M., Geofrey, J., & Lee, B. N. (2016). Students' algebraic thinking and attitudes towards algebra: The effects of game-based learning using dragonbox 12 + app. *The Research Journal of Mathematics and Technology*, *5*(1). https://doi.org/2163-0380

Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, *40*, 41–54.

Star, J. R., & Seifert, C. (2006). The development of flexibility in equation solving. *Contemporary Educational Psychology*, *31*(3), 280–300. https://doi.org/10.1016/j.cedpsych.2005.08.001

Stekhoven, D. J., & Buehlmann, P. (2012). Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Torres, R., Toups, Z. O., Wiburg, K., Chamberlin, B., Gomez, C., & Ozer, M. A. (2016). Initial design implications for early algebra games. *Proceedings of the 2016 Annual Symposium* *on Computer-Human Interaction in Play Companion Extended Abstracts*, 325–333. https://doi.org/10.1145/2968120.2987748

Vanacore, K., Gurung, A., Sales, A., & Heffernan, N. (2024). Effect of gamification on gamers: Evaluating interventions for students who game the system: Evaluating interventions for students who gaming the system. *Journal of Educational Data Mining*, *16*(11), 112–140. https://doi.org/10.5281/zenodo.11549799

Vanacore, K., Gurung, A., Sales, A., & Heffernan, N. T. (2024). The effect of assistance on gamers: Assessing the impact of on-demand hints feedback availability on learning for students who game the system. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 462–472. https://doi.org/10.1145/3636555.3636904

Vanacore, K., Lee, J.-E., Egorova, A., & Ottmar, E. (2023). Beyond performance analytics: Using learning analytics to understand learning processes that lead to improved learning outcomes. In *Perspectives on learning analytics for maximizing student outcomes* (pp. 168–187). IGI Global.

Vanacore, K., Ottmar, E., Liu, A., & Sales, A. (2024). Remote monitoring of implementation fidelity using log-file data from multiple online learning platforms. *Journal of Research on Technology in Education*, 1–21. https://doi.org/10.1080/15391523.2024.2303025

Vanacore, K., Sales, A., Liu, A., & Ottmar, E. (2023). Heterogeneous effects of game-based failure on student persistence in an online algebra game. https://sree.confex.com/sree/2023/meetingapp.cgi/Paper/4653

Vanacore, K., Sales, A. C., Hansen, B., Liu, A., & Ottmar, E. (2024). Effect of game-based failure on productive persistence: An application of regression discontinuity design for evaluating the impact of program features on learning-related behaviors. (4789291). https://doi.org/10.2139/ssrn.4789291

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved r hat for assessing convergence of mcmc (with discussion). *Bayesian analysis*, *16*(2), 667–718.

Ventura, M., Shute, V., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers & Education*, *60*(1), 52–58.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

WWC. (2020). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized bayesian knowledge tracing models. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 171–180). Springer. https://doi.org/10.1007/978-3-642-39112-5_18

Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, *30*, 100326.

Appendix A

Prior to running the full FLPS model, we validated our measurement model using a ten-fold cross-validation method. Figure 7 presents the sampling and modeling method for this cross-validation process. We specifically sought to understand how well our model performed under two conditions: (1) when the random intercept for the students is estimated empirically and (2) when the random intercepts for the students are imputed with random draws from the posterior distribution. Sampling for the cross-validation procedure required randomly dividing students from the FH2T condition into groups and partitioning the student's associated problem-level data. This process created samples of problem-level data from distinct student populations. In k-fold cross-validation, one sample is held out from the training sample each time the model is estimated, and predictions are made on the held-out sample. In this case, we added an additional step by holding out a random sampling 10% of the problems from students assigned to the training data set. Therefore, for each of the ten models, we have two hold-out samples: (1) a sample from the same students as the training data set but with problem data that is not included in model estimation (i.e. Test Data 1) and (2) a sample of students who were not included in the model estimation (Test Data 2). Predictions on these two samples are used to assess model performance under the two conditions.

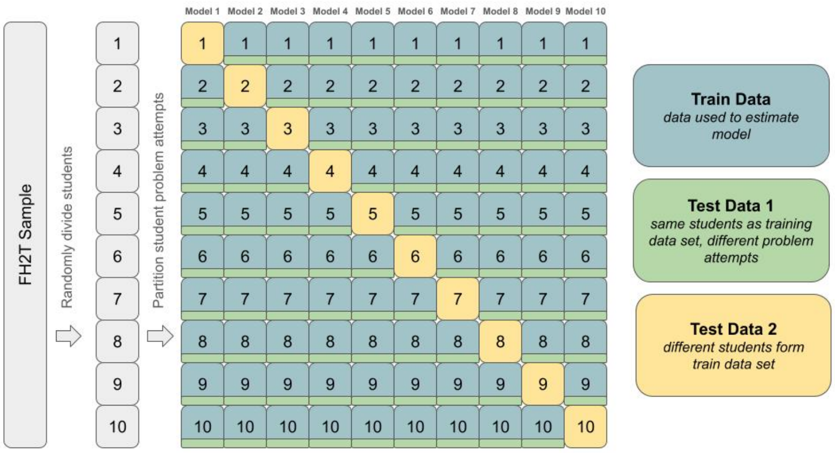

FH2T Sample
Randomly divide students
1
2
3
4
5
6
7
8
9
10
Partition student problem attempts
Model 1
Model 2
Model 3
Model 4
Model 5
Model 6
Model 7
Model 8
Model 9
Model 10
Train Data
data used to estimate model
Test Data 1
same students as training data set, different problem attempts
Test Data 2
different students form train data set

Figure 7

The results yielded acceptable model performance. Table 8 provides prediction statistics for each of the samples. The AUC on the training data is 0.91. The AUC on the test data in which the student random effect was not imputed was 0.88 (Test Data 1). The AUC on the test data in which the student random effect was imputed was 0.68 (Test Data 2).

Table 8

Caption

Sample	AUC
Training Data	0.91
Test Data 1	0.88
Test Data 2	0.68

Appendix B

As a robustness check for the FLPS model, we created fake condition samples by randomly sampling the FH2T sample and rerunning the model. In this simulation, the main and interaction effect parameters from the outcomes models should be effectively zero. We ran this procedure of sampling and re-estimating the model participants five times. Figure 8 presents the key parameter estimates and their credible intervals. All credible intervals for main and interaction effects included zero, suggesting that FLPS is robust to false positives.

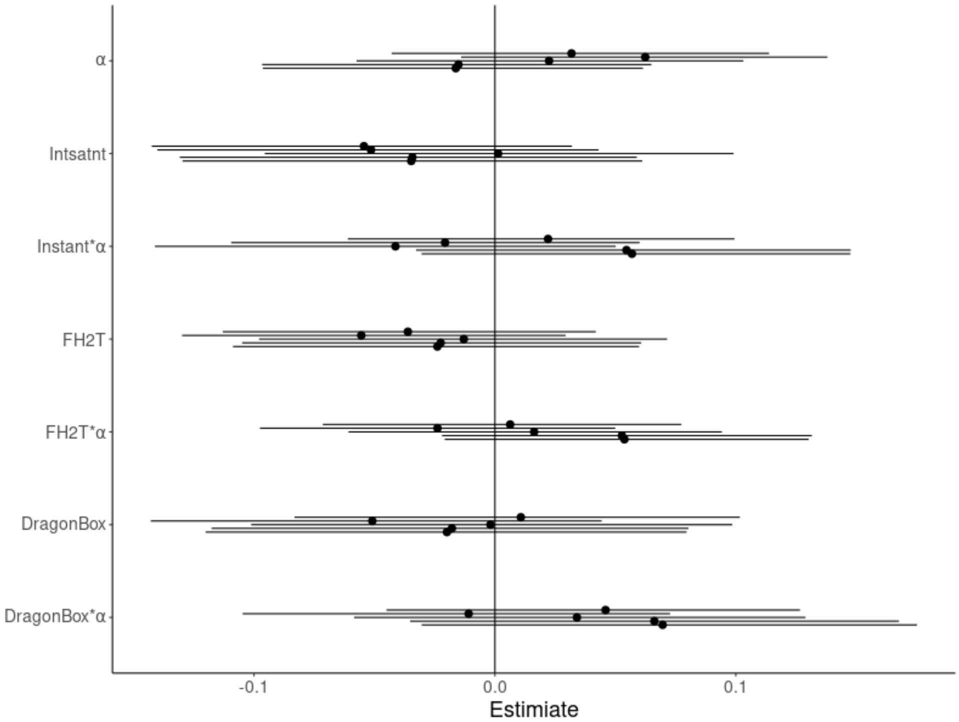


Figure 8

Parameter estimates generated from five FLPS models run with fake condition samples. All credible intervals for main and interaction effects included zero.

Chapter 5

Conclusion

Developing effective, equitably educational programs requires that they benefit students regardless of their identities, abilities, or behavioral tendencies. Often, equity research in education focuses on ensuring that programs are beneficial to students of all identities. Similarly, education researchers have put substantial emphasis on evaluating whether programs close the gap between lower- and higher-performing students. However, less work has been dedicated to understanding how students' engagement may influence the impact of a program. Throughout this dissertation, I provide evidence of effect heterogeneity across multiple CBLPs based on students' propensities towards specific engagement patterns. These findings point towards the need to leverage how students use programs when evaluating their effectiveness across the entire student population.

I introduce this idea in the book chapter *Beyond Performance Analytics: Using Learning Analytics to Understand Learning Processes that Lead to Improved Learning Outcome*, which explores how Learning Analytics (LA) can use rich data sources produced by CALPs to study and enhance students' learning processes, including their problem-solving strategies and behavioral engagement (Section 1.3). The chapter emphasizes that to maximize student outcomes, we must consider how students engage with those programs beyond their accuracy. I, along with my coauthors, point to the need for more work connecting the behaviors that students display in CALPs and the impact of those programs. This dissertation provides examples of such work.

First, I consider students who have disengaged while working in CALPs by studying the variance of impacts of CALPs based on students' tendencies towards gaming the system behaviors. I find that delayed feedback and hints can mitigate some of the negative gaming effects of students' gaming tendencies, but the method is insufficient to help gamers to perform at the level of their peers (Section 2.1). I also find that students' gaming tendencies interact inconstantly with different versions of gamification (Section 2.2). Specific designs of gamification within programs seem to influence student behavior and learning, as some gamified environments prove more effective for 'gamers' than others. The paper highlights the need for targeted interventions to support students who game the system, emphasizing that some features of gamification help 'gamers' whereas others are associated with adverse outcomes.

Next, I present research highlighting the interplay between students' persistence tendencies, program design, and program efficacy. I find inducing game-based failure through low-stakes, yet direct performance-based feedback increases the likelihood that students will engage in productive persistence by reattempting problems after the suboptimal performance (Chapter 3). However, this effect was not statistically significant for the most complex problems or for the lowest-performing students, indicating that the context and abilities of students determine how they respond to elements of gamification. In the final analysis, I find that the efficacy of three CBLPs is likely moderated by students' productive persistence tendencies. Notably, merely having a high propensity to persist is insufficient to improve students ouctomes; the design of CBLPs must provide opportunities for productive engagement, such as multiple problem reattempts. This finding of behavioral-moderated effects points to the need to create educational technologies that account for diverse behavioral tendencies.

Collectively, this work demonstrates the importance of understanding how student engagement profiles interact with digital learning environments to impact learning. The work highlights that evaluations of educational programs should consider the variance in impacts based on how different students interact with program features. The insights gained from behavior-moderated efficacy research can inform the development of targeted instructional strategies and equitable learning opportunities that cater to diverse student needs. In order for CBLPs to have consistent effectiveness across a population CBLPs their features need to account for variability in students' behavioral tendencies. Ultimately, this work points to the need to create environments that either adapt to students' engagement styles or are universally accessible and effective regardless of students' identities, abilities, and behavioral tendencies.

Unlike many heterogeneity analyses, focusing on behavior-moderated effects increases our understanding of the connection between how students use programs and the programs' impacts. These analyses indicate who benefits from different learning environments while also providing insights into how program features and elements of instructional design may help students improve their learning-related behaviors and outcomes. Taken as a whole, the analyses portray the complexities of engagement and the nuances of how to assist students with different behavioral tendencies to productively engage in learning behaviors through CBLPs.

5.1 Engagement Dichotomy

At the beginning of this dissertation, I posited that gaming the system and productive persistence were at opposite ends of the engagement spectrum: gaming the system is characterized by specifically trying not to learn, whereas productive persistence is a drive to learn despite difficulties. If true, students' propensities for these theoretically antithetical behaviors should be negatively associated. Figure 5.1 showed the association between students' propensities to game the system in the Immediate Condition and to reattempt problems after suboptimal performance in FH2T. As expected, there is a moderate negative correlation between these propensities ($r = 0.42$, $p < 0.001$), indicating that if a student is high in one propensity, they are likely low on the other. The heat map (right plot) shows the distribution of students in different quartiles for each engagement propensity. Notable, the densest regions of the plot are in the upper left and lower right corners. These are the areas that indicate if students are in the top quartile (Q4) on one engagement propensity and are most likely to be in the lowest quartile on another (Q1). Alternatively, very few students are in the lower left and upper right regions: Q1 on both or Q4 on both. This

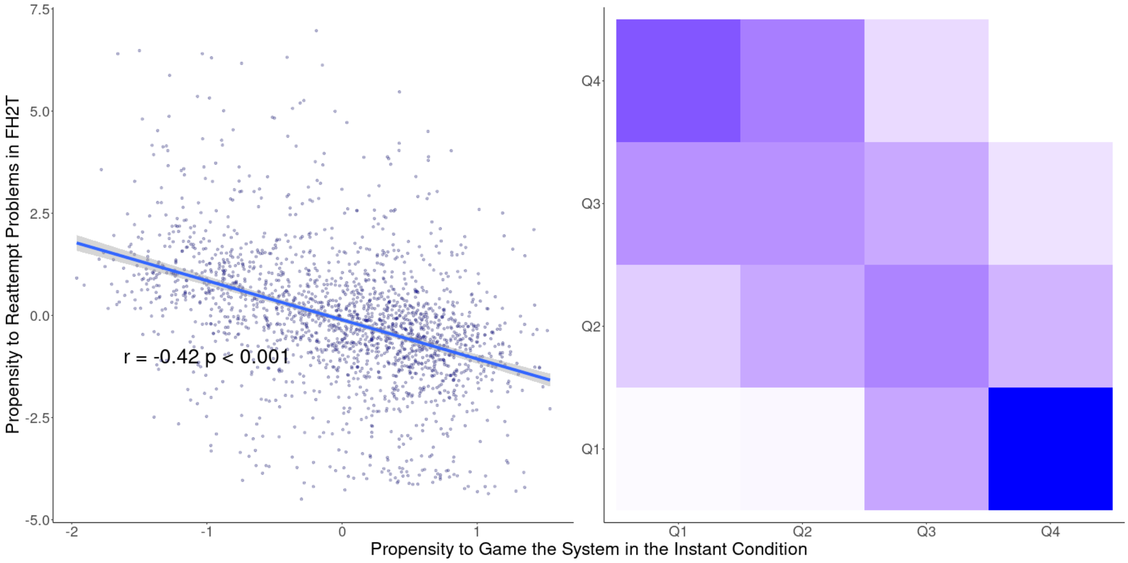


Figure 5.1: Association between students' propensities to game the system in the Immediate Condition and to reattempt problems in FH2T. The left graph plots the two propensities for each student and their correlation. The right plot is a heat map of the alignment between the two measures by quartile.

relationship supports the hypothesis that these behaviors represent different ends of the engagement spectrum.

Although the relationship between these behaviors is strong, they are not perfectly predictive of one another. Thus, some students display complex propensities for these engagement behaviors. For example, a student may tend to game the system on some problems in the Immediate Condition but also would be likely to reattempt problems in FH2T (e.g., those in Q4 on both measures). This is consistent with research suggesting that engagement behaviors are not all or nothing but rather highly context-dependent, with students' tendencies varying across platforms [2].

The FLPS measurement models used to estimate students' propensities provide some indication of student characteristics that predict different engagement tendencies. Gaming the system is more associated with general mathematical abilities, whereas productive persistence is associated with skills that are specific to the mathematical task (e.g., algebraic flexibility and perceptual sensitivity to equivalence in algebraic expressions). This suggests that students may game the system because they lack general math skills to complete the problems. On the other hand, in order for students to engage in productive presence, they may need highly specific skills for the task at hand. Notably, numerical inconfidence was positively associated with both behaviors, providing some insight into the overlap between the two behavioral profiles. Perhaps gamified conditions encourage students who lack confidence in their math skills to persist by helping them imagine new ways to engage with math. In contrast, the traditional CBLP may drive these students to become discouraged and disengage. These nuanced findings highlight that more work is necessary to understand the complex cognitive and affective antecedent or engagement.

Engagement may be viewed as a product of students' own volition. Under this view, education may be characterized by the old axiom: ”You can lead a horse to water, but you can't make it drink." Similarly, education researchers have debated whether engagement is a transitory state or fixed trait [4, 41]. The research presented in this dissertation indicates that, although engagement propensities can be estimated for a specific context, changing the context affects whether students with different engagement propensities learn. For example, a student who is likely to game the system when they do problems with on-demand hints and automatic feedback is likely to learn more when these features are removed. Thus, while latent traits of engagement exist, the manifestations of these traits are highly context-dependent.

5.2 Behavior-Moderated Effects

As stated above, analyses of effect heterogeneity are necessary to ensure that all students experience the positive effects of educational programs. We must specifically ensure that programs do not disproportionately affect the most advantaged students over the least advantaged [22]. A common finding is that typical effect heterogeneity analyses may reveal that disadvantaged groups of students do not benefit as much as advantaged groups [34, 45]. Although these analyses are important indications of problems, they often fail to address why some students benefit and others do not. Analyses of behavior-moderated effects can point toward which behaviors may be driving these disparities and what programmatic aspects may help mitigate them. Furthermore, estimating behavior-moderated effects may reveal unexpected heterogeneity patterns that provide insights into the nuanced connection between student behaviors and the benefits of educational programs.

The analyses of behavior-moderated effects in this dissertation provide insights into why some students benefit, and others do not and what types of environments might be necessary to help students engage in productive learning behaviors. Overall, these findings also suggest that the impact of more innovative programs may depend on specific engagement tendencies generally held by high-performing students. FH2T involves an innovative method of teaching problems by allowing them to move components of the problems around following the order of operations and algebraic rules. Although FH2T has a positive average treatment effect on students' allergic knowledge, this effect is smaller for students with lower pretest knowledge [13] and negative for students who tend to game the system (Section 2.2) or do not engage in productive persistence (Chapter 4). This suggests that FH2T teaching methods are gap-increasing innovation [22], as disadvantaged students – lower performing and less engaged – either benefit less or are adversely impacted by the program, whereas advantaged students – higher performing and more engaged – benefit more from the program.

Although simply interacting pretest knowledge with the FH2T condition, as done by [13], is sufficient to identify FH2T as a gap-increasing innovation, it does not indicate why and leaves the designers to speculate on how to mitigate this problem. Alternatively, behavior-moderated effects can point toward potential causes and solutions. For example, this research suggests that the effects of FH2T may be contingent upon whether a student is likely to engage in productive performance by reattempting problems after the suboptimal performance. Thus, it is likely that increasing students' propensity to engage in productive persistence may increase student learning. Yet, I found that although game-based failure increases the probability that students would engage in productive persistence, the effect was not significant for the lowest-performing students (Chapter 3). Thus, the students who needed the most encouragement to persist were least responsive to a game-based failure intervention. Thus, other methods of encouraging low-performing students to attempt problems after suboptimal performance could be tested to see if they increase students' persistence and improve the impact of the program.

These findings indicate that gamification is not a panacea for disengagement, even though certain gamified features may help encourage positive behaviors and possibly migrate negative behaviors. The patterns of heterogeneity between the two gamified conditions (FH2T and DragonBox) suggest the specifics of gamification matter more than whether a program is gamified at all. FH2T had estimated negative effects on students with high gaming tendencies, whereas DragonBox positively affected students regardless of the propensity to game the system. On the other hand, the moderating effect of the propensity to engage in productive persistence was likely greater for FH2T than for Dragonbox. Put differently, DragonBox had an estimated positive impact on a wider array of students with differing engagement profiles, whereas FH2T only had an estimated positive impact on students who tended to be highly engaged (e.g., low likelihood of gaming the system and high likelihood of persistence). This pattern exists despite the fact that the two programs share a key innovation: dynamic manipulation of equations. Dragonbox differentiated itself by including native goals and puzzle-based math primers. These features may be critical in helping students engage in learning activities.

A benefit of the analysis of behavior-moderated effects is that they point toward which features may positively or negatively impact students' learning depending on their behavioral engagement tendencies. Table 5.1 provides some key features to the programs and Table 5.2 indicates whether each program uses those features. By examining the features of the CBLPs along with our understanding of which behavioral tendencies are associated with the program's effectiveness, we can hypothesize whether features may be the "active ingredients" in producing differing results for students. For example, on-demand hints and immediate feedback likely drive the negative effects on students who are likely to game the system. The differences between the gamified conditions can also provide some indications of which features may help specific subsets of students learn. Notably, DragonBox uses pedagogical techniques similar to those used in FH2T, but students tend to benefit from this program regardless of their ability or engagement tendencies. There are two possible explanations for the trend. First, DragonBox may implement the innovation differently, making it more accessible and engaging to all students. Second, DragonBox's other features – including a native goal (freeing the dragon) and teaching students algebra through non-mathematical symbols before introducing more traditional equations (puzzle-based math primers) – may alleviate the gap-increasing nature inherent to the underlying pedagogical technique.

This work also provides a better understanding of the relationship between assistance features in non-gamified CALPs and students' engagement tendencies. Both studies of engagement behavior moderated effects suggest that students with low engagement tendencies (i.e., those who are less likely to persist and more likely to game) may benefit from restricted access to on-demand hints and immediate feedback. This suggests that these students are not using these features in ways that maximize their learning potential. Thus, to increase the impact of CALPs on disengaged students, instructional designs may ned to include features that directly address how these students interact with assistance.

Table 5.1: Gamified and traditional CBLP feature definitions

Features	Definition	Feature of Gamification*
Hints	Subtle suggestions to guide problem-solving process	No
Embedded feedback	Feedback that occurs during the problem-solving process	No
Performance-based feedback	Feedback on performance (e.g., correct, incorrect, 90%)	No
Performance-based rewards	Rewards given based on performance (e.g., stars)	Yes
Reattempt problem	Ability to try a problem again	No
Embedded gestures	Ability to dynamically manipulate problem elements	Yes
Narrative goals	Goals unrelated to the underlying concepts, but achieved through performance (e.g., freeing the dragon)	Yes
Diverse representations of concepts	Underlying concepts are presented in many ways (e.g., teaching algerbia concepts through puzzels)	No

* Features more typically associated with games than traditional learning activities

Table 5.2: CBLP Features

	Platforms			
Features	FH2T	DargonBox	Immediate Condition	Delayed Condition
Hints	Yes	Yes	Yes	No
Embedded feedback	Yes	Yes	No	No
Performance-based feedback	Yes	Yes	Yes	No
Performance-based rewards	Yes	Yes	No	No
Reattempt problem	Yes	Yes	Yes	No
Embedded gestures	Yes	Yes	No	No
Narrative goals	No	Yes	No	No
Diverse representations of concepts	No	Yes	No	No

5.3 Future Directions

Overall, these examples of research on the connection between instructional design, engagement, and learning impacts provide indications of how programs can be improved to be highly impactful for all students. As explained above, this work points toward potential solutions for pernicious effect heterogeneity (e.g., gap-increasing effect variance). However, future work must test these solutions to ensure that they improve effect equity within innovative CBLPs.

Finding, deploying, and testing ways to increase the persistence behaviors of all students, specifically those who are lower performing, will likely increase the impact of FH2T for students who generally struggle with persistence behaviors. Similarly, restricting access to hints and feedback for students who tend to game the system may help these students engage in more productive learning behaviors. However, features like these need to be tested to ensure that they have the desired effects.

Finally, behavior-moderated effects analyses also have the potential for more efficient use of A/B testing as they point towards specific features that may produce effect differentials based on students' behaviors. For example, A/B testing adaptive features in traditional CBLPs that restrict access to assistance for 'gamers' should likely be tested. Also, testing whether requiring students to try multiple paths to solutions in general or only on problems in which students perform sub-optimally would help refine our understanding of the learning mechanisms behind reattempt features. This understanding could help the field develop better features that leverage the benefits of students generating multiple solutions. Similarly, evaluating whether the diverse representations of concepts used in DragonBox might alleviate some of the heterogeneity in FH2T. Testing the addition of puzzle-based math primers in FH2T could help establish whether this is the curtail difference between the programs that makes DargonBox more accessible to a wider array of students. In general, behavior-moderated effects produce hypotheses about the mechanisms of learning in CBLPs that may be distilled into testable features for future research.

Bibliography

[1] Amrein, A. L., and Berliner, D. C. The effects of high-stakes testing on student motivation and learning. *Educational Leadership 60*, 5 (2003), 32–38. ERIC Number: EJ660880.

[2] Baker, R., and Carvalho, A. Labeling student behavior faster and more precisely with text replays.

[3] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., and Koedinger, K. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research 19*, 2 (2008), 185–224.

[4] Baker, R. S. Is gaming the system state-or-trait? educational data mining through the multi-contextual application of a validated behavioral model. In *Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007* (2007), vol. 2007, User Modeling Inc Boston, MA, pp. 76–80.

[5] Baker, R. S. J. d., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G. W., Ocumpaugh, J., and Rossi, L. *Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra.* June 2012. ERIC Number: ED537205.

[6] Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., and Salas, E. Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation & Gaming 43*, 6 (2012), 729–760.

[7] Bickman, L., and Reich, S. M. Randomized controlled trials. *What counts as credible evidence in applied research and evaluation practice 51* (2008).

[8] Boedeker, P., and Henson, R. K. Evaluation of heterogeneity and heterogeneity interval estimators in random-effects meta-analysis of the standardized mean difference in education and psychology. *Psychological Methods 25*, 3 (2020), 346.

[9] Botelho, A. F., Van Inwegen, E. G., Varatharaj, A., and Heffernan, N. T. Refusing to try: Characterizing early stopout on student assignments. In *PervasiveHealth: Pervasive Computing Technologies for Healthcare* (Mar. 2019), ICST, p. 391–400.

[10] Bryan, C. J., Tipton, E., and Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour 5*, 88 (Aug. 2021), 980–989.

[11] Butler, A. C., and Woodward, N. R. *Toward consilience in the use of task-level feedback to promote learning*, vol. 69. Academic Press, Jan 2018, p. 1–38.

[12] Chan, J. Y.-C., Ottmar, E. R., and Lee, J.-E. Slow down to speed up: Longer pause time before solving problems relates to higher strategy efficiency. *Learning and Individual Differences 93* (Jan. 2022), 102109.

[13] Decker-Woodrow, L. E., Mason, C. A., Lee, J.-E., Chan, J. Y.-C., Sales, A., Liu, A., and Tu, S. The impacts of three educational technologies on algebraic understanding in the context of covid-19. *AERA Open 9* (Jan. 2023), 23328584231165919.

[14] Deterding, S., Khaled, R., Nacke, L., and Dixon, D. Gamification: Toward a definition. In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems* (Jan. 2011), ACM, p. 12–15.

[15] Durlak, J. A., and DuPre, E. P. Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology 41*, 3–4 (June 2008), 327–350.

[16] Dynarski, M. Using research to improve education under the every student succeeds act. *Evidence Speaks Reports 1*, 8 (2015), 1–6.

[17] Emerson, A., Min, W., Azevedo, R., and Lester, J. Early prediction of student knowledge in game-based learning with distributed representations of assessment questions. *British Journal of Educational Technology n/a*, n/a.

[18] Fadda, D., Pellegrini, M., Vivanet, G., and Zandonella Callegher, C. Effects of digital games on student motivation in mathematics: A meta-analysis in k-12. *Journal of Computer Assisted Learning 38*, 1 (2022), 304–325.

[19] Hale, S., Dunn, L., Filby, N., Rice, J., and Van Houten, L. Evidence-based improvement: A guide for states to strengthen their frameworks and supports aligned to the evidence requirements of essa. *WestEd* (2017).

[20] Juul, G. J. *Fear of Failing? The Many Meanings of Difficulty in Video.* Routledge, 2008.

[21] Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., Heffernan, N., et al. Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining 10*, 1 (2018), 36–71.

[22] Kizilcec, R. F., and Lee, H. *Algorithmic fairness in education.* Taylor Francis, 2022. arXiv:2007.05443 [cs].

[23] Koedinger, K. R., Pavlik, P., McLaren, B. M., and Aleven, V. Is it better to give than to receive? the assistance dilemma as a fundamental unsolved problem in the cognitive science of learning and instruction. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (Austin, TX, 2008).

[24] Kratochwill, T., Hitchcock, J., Horner, R., Levin, J., Odom, S., Rindskopf, D., and Shadish, W. What works clearinghouse. *Retrieved Sep 25* (2010), 2011.

[25] Krumm, A. E., Beattie, R., Takahashi, S., D'Angelo, C., Feng, M., and Cheng, B. Practical measurement and productive persistence: Strategies for using digital learning system data to drive improvement. *Journal of Learning Analytics 3*, 22 (Sept. 2016), 116–138.

[26] Landers, R. N. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation Gaming 45*, 6 (Dec. 2014), 752–768.

[27] Lee, J.-E., Chan, J. Y.-C., Botelho, A., and Ottmar, E. Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational technology research and development 70*, 5 (Oct. 2022), 1575–1599.

[28] Lendrum, A., and Humphrey, N. The importance of studying the implementation of interventions in school settings. *Oxford Review of Education 38*, 5 (Oct. 2012), 635–652.

[29] Lu, X., Sales, A., and Heffernan, N. T. Immediate versus delayed feedback on learning: Do people's instincts really conflict with reality? *Journal of Higher Education Theory and Practice 21*, 1616 (Dec. 2021).

[30] Lu, X., Wang, W., Motz, B. A., Ye, W., and Heffernan, N. T. Immediate text-based feedback timing on foreign language online assignments: How immediate should immediate feedback be? *Computers and Education Open 5* (Dec. 2023), 100148.

[31] Malone, T. W. Toward a theory of intrinsically motivating instruction. *Cognitive science 5*, 4 (1981), 333–369.

[32] McClelland, D. C. *Achieving society*, vol. 92051. Simon and Schuster, 1961.

[33] Ottmar, E., Lee, J.-E., Vanacore, K., Pradhan, S., Decker-Woodrow, L., and Mason, C. A. Data from the efficacy study of from here to there! a dynamic technology for improving algebraic understanding. *Journal of Open Psychology Data 11*, 1 (Apr. 2023), 5.

[34] Pane, J. F., Seaman, D., and Doss, C. J. *Students Using Lexia® Core5® Reading Show Greater Reading Gains Than Matched Comparison Students.* Sept. 2023.

[35] Razzaq, L., Heffernan, N. T., and Lindeman, R. W. What level of tutor interaction is best?

[36] Richey, J. E., Zhang, J., Das, R., Andres-Bray, J. M., Scruggs, R., Mogessie, M., Baker, R. S., and McLaren, B. M. Gaming and confrustion explain learning advantages for a math digital learning game. In *In Artificial Intelligence in Education: 22nd International Conference* (Utrecht, The Netherlands, 2021), I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, Eds., Lecture Notes in Computer Science, Springer International Publishing, p. 342–355.

[37] Rollings, A., and Morris, D. *Game architecture and design.* Coriolis, 200.

[38] Rubin, D. B. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association 103*, 484 (2008), 1350–1353.

[39] Schoenfeld, A. H. What doesn't work: The challenge and failure of the what works clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher 35*, 2 (2006), 13–21.

[40] Schudde, L. Heterogeneous effects in education: The promise and challenge of incorporating intersectionality into quantitative methodological approaches. *Review of Research in Education 42*, 1 (2018), 72–92.

[41] Shernoff, D. J., and Shernoff, D. J. Engagement as an individual trait and its relationship to achievement. *Optimal learning environments to promote student engagement* (2013), 97–126.

[42] Shute, V. J. Focus on formative feedback. *Review of Educational Research 78*, 1 (Mar 2008), 153–189.

[43] Slavin, R. E. Evidence-based reform in education. *Journal of Education for Students Placed at Risk (JESPAR) 22*, 3 (2017), 178–184.

[44] Soderstrom, N. C., and Bjork, R. A. Learning versus performance: An integrative review. *Perspectives on Psychological Science 10*, 2 (Mar. 2015), 176–199.

[45] Walberg, H. J., and Tsai, S.-L. Matthew effects in education. *American Educational Research Journal 20*, 3 (Jan. 1983), 359–373.