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

*Authors contributed equally to this publication.

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636904

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

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.

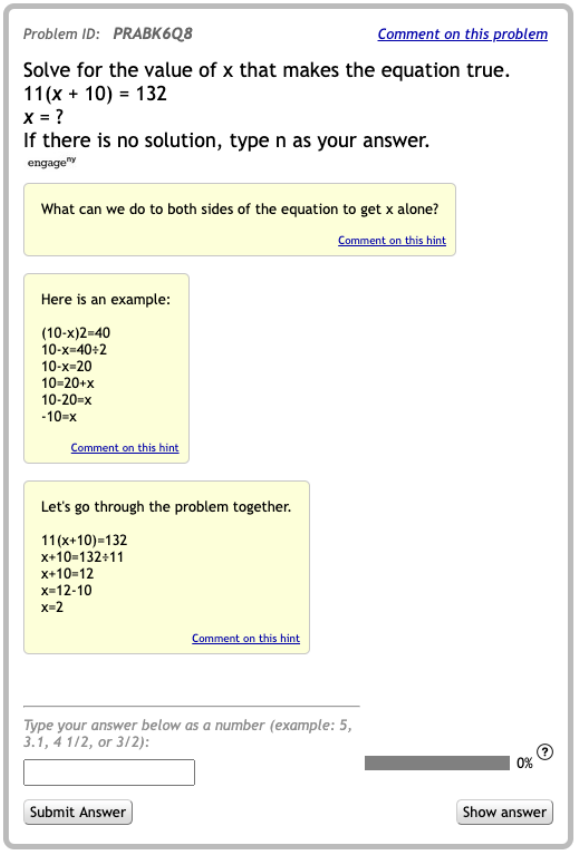

Figure 1: Immediate Feedback Condition Example

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.

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

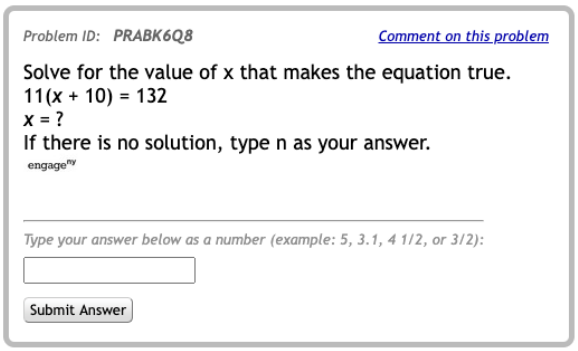

Figure 2: Delayed Feedback Condition Example

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.

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.

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

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

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

We impute α_i for C with random draws from a normal distribution with mean $\sum_k \gamma_k P_{ki} + \mu_{t[i]} + \mu_{s[i]}$, 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 ν_t and for school be ν_s.

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

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 (γ_{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 (γ_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 (γ_7 = -0.04, $P(< 0)$ = .92). Students with a higher negative reaction toward math (γ_8 = 0.19, $P(> 0)$ = .99) and with higher numeric confidence (γ_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 (β_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 (β_2 = -0.37, $P(< 0) > .99$).

The interaction between students' propensity to game the system and the Immediate Condition was likely negative (β_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