Item Response Theory-Based Gaming Detection

Yun Huang[1], Steven Dang[2], J. Elizabeth Richey[3], Michael Asher[4], Nikki G. Lobczowski[3], Danielle Chine[1], Elizabeth A. McLaughlin[1], Judith M. Harackiewicz[4], Vincent Aleven[1] and Kenneth Koedinger[1]

[1] Carnegie Mellon University
yunhuanghci@cmu.edu, {dchine, mimim}@andrew.cmu.edu, aleven@cs.cmu.edu, koedinger@cmu.edu

[2] Lexia Learning
steven.dang@lexialearning.com

[3] University of Pittsburgh
jelizabethrichey@gmail.com, ngl13@pitt.edu

[4] University of Wisconsin-Madison
{mwasher, jmharack}@wisc.edu

ABSTRACT

Gaming the system, a behavior in which learners exploit a system's properties to make progress while avoiding learning, has frequently been shown to be associated with lower learning. However, when we applied a previously validated gaming detector across conditions in experiments with an algebra tutor, the detected gaming was not associated with learning, challenging its construct validity. Our iterative exploratory data analysis suggested that some contextual factors that varied across and within conditions might contribute to this lack of association. We present a latent variable model, *item response theory-based gaming detection* (IRT-GD), that accounts for contextual factors and estimates latent gaming tendencies as the degree of deviation from normative behaviors across contexts. Item response theory models, widely used in knowledge assessment, account for item difficulty in estimating latent student abilities: students are estimated as having higher ability when they can get harder items correct than when they only get easier items correct. Similarly, IRT-GD accounts for contextual factors in estimating latent gaming tendencies: students are estimated as having a higher gaming tendency when they game in less commonly gamed contexts than when they only game in more commonly gamed contexts. IRT-GD outperformed the original detector on three datasets in terms of the association with learning. IRT-GD also more accurately revealed intervention effects on gaming and revealed a correlation between gaming and perceived competence in math. Our approach is not only useful for others wanting to apply a gaming assessment in their context but is also generally applicable in creating robust behavioral measures.

Keywords

Gaming the system, item response theory, behavior modeling

Y. Huang, S. Dang, J. E. Richey, M. Asher, N. G. Lobczowski, D. Chine, E. A. McLaughlin, J. M. Harackiewicz, V. Aleven, and K. Koedinger. Item response theory-based gaming detection. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 251–262, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853093

1. INTRODUCTION

Assessing students' engagement levels or motivation from their interaction behaviors in digital learning environments is a compelling challenge both practically and theoretically. Practically, valid behavioral assessment of student engagement can drive adaptations that adjust to students' needs, leading to greater learning and motivation; theoretically, valid behavioral assessment of student engagement can be used to better understand when and why interventions or system designs work for enhancing student learning or motivation. One frequently explored behavioral indicator of student engagement is "gaming the system", which 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" [6]. Many studies have demonstrated that gaming the system (abbreviated as *gaming* in this paper) is associated with poor learning outcomes in the short term or the long term [2, 5, 12, 30]. Prior research suggests that interventions directly targeting gaming can reduce gaming behaviors [4, 33] and improve learning [4], demonstrating the practical value of gaming detection. Recent work [28] has also shown that the positive effect of learning with an educational game was fully mediated by lower levels of gaming the system, showcasing the theoretical value of gaming detection for understanding how a specific intervention influences learning.

Past research has leveraged two classes of approaches to model gaming behaviors: knowledge engineering where experts develop rational rules that identify gaming behaviors [21, 25, 26] and machine learning where the model designer creates a set of features first and then a supervised learning algorithm is used to select features for predicting human coded gaming labels [6, 32]. Mainly the emphasis has been put on student features [21, 25, 26], such as how students utilize help (e.g., help abuse [1]) and make errors (e.g., systematic guessing [32]). Task or system features have been investigated to a limited extent although they have been found to be important contextual factors for gaming. For example, [8] found that system features explained more variance in gaming behaviors than student characteristics on a year-long log dataset with 22 different lessons of Cognitive Tutor Algebra. In particular, the results showed that gaming was more frequent in lessons that were abstract, ambiguous, and had unclear presentation of the content or task. Another study [22] also found that differences in gaming behaviors were more strongly associated with the learning environments than with student populations. Some machine-learned models have incorporated one or two task features in the initial set of features [6, 32]. For example, [32] included features related to question types (top-level or follow-through helping questions) and interfaces (multiple-choice or textbox) in the initial feature set, yet they did not mention whether these features remained in the final model. Meanwhile, in knowledge-engineered gaming detectors, task features typically are not (explicitly) considered [21, 25], i.e., rules to identify gaming behaviors are described in a task type-independent way. Such limited consideration of task features may limit the generalizability of gaming detectors to new contexts with task or system design substantially different from that of the original context. In addition, some research has identified cases where detected gaming behaviors were not harmful [6, 12] or were even good learning behaviors [31], further suggesting the necessity to look carefully into the contexts for identifying or interpreting gaming behaviors. However, such unconventional findings (unconventional in the sense that gaming has frequently been shown to be associated with lower learning) still have not received enough attention in the development of gaming detectors.

Obtaining a student-level gaming estimate has been valuable for studying the relation between student attributes and gaming or intervention effects on gaming. Prior work has predominantly used direct aggregation of detected (or observed) gaming by computing the proportion of gamed transactions or the average of predicted probabilities of gaming for each student [6, 22, 27]. However, [13] showed that the observation-level simple average failed to reveal correlations between motivation and gaming (except for one motivational measure), while a simple latent variable model that estimated a latent gaming tendency for each student controlling for the effect of curricular sections on detected gaming yielded strong associations between a range of motivational measures and gaming tendencies. Inspired by this prior work, we identified an overlooked connection between existing behavior modeling paradigms and knowledge modeling paradigms: latent variable models widely used for estimating student abilities or knowledge levels can also be used to obtain more valid student-level behavioral constructs, thanks to their capacity to account for both task and student features in a single framework. One widely used latent variable modeling paradigm for knowledge assessment in psychometrics is *item response theory* (IRT) [14]. IRT models the observed correctness on each item (e.g., problem steps) of each student as a function of item difficulty and student ability. Instead of using the proportion of correctly answered items as the measure of student ability, IRT accounts for item difficulty in estimating latent student abilities. In essence, students are estimated as having higher ability when they can get harder items correct than when they only get easier items correct. IRT models have been further extended to model dynamic student knowledge by considering the temporal aspect [17], and also by decomposing items into knowledge components (e.g., skills, concepts) shared across items [11].

With the increasing demand of learning engineering efforts towards building effective, engaging learning systems, the generalizability, interpretability, and development cost of gaming detectors are becoming increasingly important. Recent studies [23, 24] compared three previously separately validated gaming detectors across multiple systems: a knowledge-engineered model [25], a machine-learned model [7], and a hybrid model [24] that combines both knowledge engineering and machine learning. In particular, the knowledge-engineered model was developed by using cognitive task analysis to elicit knowledge about how experts code students as gaming or not in Cognitive Tutor Algebra [19]. It consists of 13 patterns of students' systematic guessing and help abuse behaviors. The comparisons [23, 24] focused on predictive performance of expert labels of gaming in held-out test sets in the original data and two new datasets collected from two other learning environments [3, 27]; the comparison also considered the interpretability of models. Results showed that the knowledge-engineered model achieved greater generalizability to new datasets and interpretability than the machine-learned model, and achieved comparable to slightly better generalizability and interpretability than the hybrid model. Although there was initial cost (higher than that of the machine-learned model) in developing the knowledge-engineered model, it could be directly used in new datasets without further cost (since actions that match any of the 13 patterns can be directly labeled as gaming). However, one may need to retrain the machine-learned or hybrid model (that needs a machine-learned model as input), given the much lower (and even unacceptable) predictive performance of the machine-learned model than the knowledge-engineered model on new datasets [23]. Thus, this knowledge-engineered gaming detector [25], which is referred to as KE-GD in this paper, appears to be the best choice (to build on) among the three detectors, considering generalizability, interpretability, and development cost in a new context altogether; it also represents a broad class of behavioral detectors that are built based on rational rules specified by experts. However, predictive performance of expert labels is only one aspect of construct validity; the establishment of construct validity of a gaming detector also requires examining the association between detected gaming and learning. Past studies [23, 24, 25] have not examined the association between detected gaming by KE-GD with learning, while other studies on other detectors have frequently shown that a higher detected gaming level is associated with lower learning [5, 15, 20, 21, 28].

In this work, we propose a latent variable model, *item response theory-based gaming detection* (IRT-GD), that estimates a latent gaming tendency for each student accounting for contextual factors (i.e., task and student features): students are estimated as having a higher gaming tendency when they game in less commonly (or frequently) gamed contexts than when they only game in more commonly (or frequently) gamed contexts. IRT-GD builds on a previously validated knowledge-engineered gaming detector (KE-GD) that focuses on students' action features and the predictiveness of human labels. We started with applying KE-GD on a dataset collected from experimentation with an algebra tutor, and examined the association between detected gaming and learning, an important aspect for the construct validity of gaming measures. Observing the lack of association with learning, we conducted an iterative exploratory data analysis, and found that this lack of association might result from some contextual factors not considered in KE-GD that varied across and within conditions. Without complex human feature engineering, we integrated contextual factors as predictors in a mixed effect model predicting whether a transaction was detected as gaming by KE-GD, and extracted the student random intercepts as the latent gaming tendencies. We compared KE-GD and IRT-GD by the association with learning in nine contexts, obtained from three datasets and three condition configurations per dataset. Finally, we demonstrated two applications of IRT-GD: to study whether there was a difference in the level of gaming between the two conditions from our experimentation with the tutor, and to explore the relation between gaming and motivation. The development and evaluation process of IRT-GD is explained as follows.

2. DEVELOPMENT OF IRT-GD

2.1 The Tutor

We used datasets collected from an algebra intelligent tutoring system for middle and high school students [18]. Students learn about writing algebraic expressions in story problems in various task (problem) formats: writing an expression in a textbox with dynamic scaffolding steps that appear if a student fails in the original question (*text* format); writing expressions in a table where the main question step and scaffolding steps are accessible at any time and are all required (*table* format); explaining a set of expressions extracted from a given equation by choosing the matching textual description from a dropdown menu for each expression (*menu* format); and given an equation, writing a set of expressions that match a given set of textual descriptions (*flipped-menu* format). These tasks also vary in the complexity of the expressions involved (e.g., one or two operators).

The algebra tutor was continuously redesigned and tested in three experiments with different student populations across three years. In each experiment (eight sessions over four weeks), we compared two versions of the tutor corresponding to two conditions differing in task design and sequencing. The control (CT) condition, corresponding to the original tutor, provided a *normal deliberate practice* schedule. Students received *full* tasks representing the full version of the problem requiring filling in all steps (including scaffolding steps) given a cover story. There were three consecutive units: the first unit contained all the table tasks, the second unit contained less complex menu and flipped-menu tasks, and the third unit contained more complex menu and flipped-menu tasks. Steps were labeled with coarser-grained knowledge components (KCs; skills). Students received individualized practice until reaching mastery of all KCs in a unit before moving on to the next unit. Across the three experiments, the design of the control condition remained the same. The experimental (EXP) condition, the *data-tuned adaptive* condition, corresponds to a redesigned tutor with redesign decisions drawn from data mining outcomes of student log data. It provided an *intense deliberate practice* schedule with task design and sequencing based on a refined, larger KC model revealing hidden difficulties (i.e., original KCs were split to differentiate easier and harder use cases). *Focused* tasks were introduced to reduce over-practicing easier KCs and target particularly difficult KCs. Examples of focused tasks include: text format tasks asking for the final expression without the mandatory intermediate steps required in table task; text format tasks that further remove the story and focus on learning algebraic grammar rules; and simpler menu and flipped-menu tasks with equations less complex than the original equations. There were three or more learning units where different task formats or task types (full or focused) were interleaved in each unit. Students received individualized practice until reaching mastery of all KCs in a unit before moving on to the next unit. Across the three experiments, the design of the experimental condition was continuously refined aiming at promoting greater learning. Our prior work has shown that the experimental condition led to better learning outcomes compared to the control condition [18]. Here, we are interested to see whether intense deliberate practice (i.e., the experimental condition) also led to higher behavioral engagement, particularly lower levels of gaming the system, and also whether gaming was linked with motivation. We started our investigation with the first dataset collected from the first experiment explained below.

2.2 A Previously Validated Knowledge-Engineered Gaming Detector Did Not Generalize

We chose a previously validated knowledge-engineered gaming detector, KE-GD, as the starting point for studying students' behavioral engagement when using the algebra tutor. KE-GD contains 13 interpretable patterns modeling systematic guessing and help abuse. For example, one pattern is "the student enters an incorrect answer, enters a similar and incorrect answer in the same part of the problem and then enters another similar answer in the same part of the problem". It is coded as "incorrect → [similar answer] [same context] & incorrect → [similar answer] & [same context] & attempt", consisting of constituents such as "[similar answer]" (judged by Levenshtein distance), and action types such as "attempt" (correct or incorrect) or "help". If a sequence of transactions (i.e., student-step interactions considering multiple attempts per step) matches any one of the 13 patterns, then all transactions involved are labeled as gaming. Details of the patterns and the validation of KE-GD could be found in [23, 25].

We used KE-GD to label transactions as gaming or not and then examined the construct validity of detected gaming in our dataset. We defined two metrics of construct validity in the current study, both of which evaluate the association between gaming and learning. The primary metric was the correlation between gaming levels and normalized learning gains over students. For each student, we computed a gaming level using the proportion of gamed transactions (referred to as *proportion of detected gaming* or *detected gaming (proportion)*) for KE-GD, or the estimated gaming tendency for IRT-GD (explained in Section 2.3.2); we computed the normalized learning gain using the widely adopted formula, *(posttest - pretest) / (1- pretest)*. We used Spearman correlation (*rho*) because it is less sensitive to outliers than Pearson correlation. As a supplementary metric, we conducted a regression analysis predicting posttest scores controlling for pretest scores and gaming levels over students and examined the coefficient of the variable of gaming levels. We considered negative correlations and coefficient values at a significance level of 0.10 as acceptable construct validity. Prior studies have used significance levels of 0.05 and 0.10 for correlation analyses involving behavior measures [5, 13, 31].

Two observations emerged. First, the detected gaming proportion 18% (last column in Table 1) was much higher than the previously reported proportions (3.5% in [13] and 6.8% in [25]) of the same detector in other math intelligent tutoring systems. Second, there was a lack of association between detected gaming and learning (correlation: *rho*=-.02, *p*=.86; regression coefficient: *b*=0.07, *p*=.69), challenging KE-GD's construct validity in our context.

Table 1. Statistics of the Fall 2019 dataset including the proportion of gamed transactions (considering all attempts of all steps) detected by KE-GD.

#stu	#transaction (tx)	#tx of 1st attempts of steps w/ KCs	Avg proportion of gamed tx over stu
129	98,176	32,419	.18 (*SD*=.08)

2.3 Identifying and Integrating Contextual Factors to Improve Construct Validity

Next, we conducted iterative exploratory data analysis on the first dataset to identify contextual factors that might explain the lack of association between detected gaming by KE-GD and learning, and integrated the contextual factors through latent variable modeling analogous to item response theory modeling, explained as follows.

2.3.1 Identifying the effect of task formats

One notable feature of our dataset compared to other datasets for developing gaming detectors is that it was collected from experimentation with two conditions with substantial differences in task design and sequencing. So, we first conducted a moderation analysis to test whether the condition moderated the relation between detected gaming and learning. We constructed a regression model predicting posttest scores for each student given the pretest scores, the condition indicator, detected gaming proportion and an interaction term between the condition and detected gaming proportion. The interaction was significant (*b*=-0.98, *p*=.007) and the control condition showed a relation opposite to theoretical prediction: higher proportion of detected gaming was associated with higher posttest scores (Figure 1).

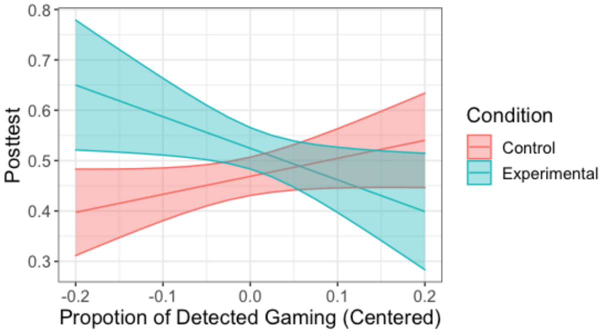

Figure 1. The interaction plot between the condition and detected gaming proportion of the regression model predicting posttest scores with pretest scores controlled for.

To understand this interaction, we started an exploratory data analysis on the overall dataset to examine when and how students gamed according to KE-GD. We used the unit of analysis normally used for modeling student learning, knowledge components (KCs), for better drawing insights into the relation between gaming and learning. We used the KC model previously validated for this dataset [18] based on model fitness. It includes 26 KCs shared by both conditions. We first examined whether students gamed much more on some KCs than on others, and if so whether there was a pattern in this variation. A pattern emerged (see Figure 2) showing that students gamed substantially more on KCs required in menu and flipped-menu formats than those required in table and text formats. Meanwhile, we knew that the control condition positioned menu and flipped-menu tasks in later units, whilst the experimental condition interleaved menu and flipped-menu tasks with text and table tasks in earlier units. Having in mind that higher detected gaming was associated with higher posttest scores in the control condition (Figure 1), we wondered whether this association was because students with higher abilities (who usually also have higher posttest scores) progressed faster to later units and thus accessed a higher proportion of menu and flipped-menu steps, which were highly-gamed contexts, than students with lower abilities. We approximated students' abilities by pretest scores and investigated this relation. Indeed, as shown in Figure 3, students with higher pretest scores in the control condition accessed a higher proportion of menu and flipped-menu steps than students with lower pretest scores (which was not the case for the experimental condition), and as a result, they might appear to game more than students with lower pretest score. Thus, the positive association between detected gaming and posttest scores in the control condition was spurious due to a confounder, the proportion of highly-gamed format steps a student accessed. The association between detected gaming revealed by KE-GD and posttest scores was biased. If we introduce task formats to account for (part of) the detected gaming, then this bias may be reduced.

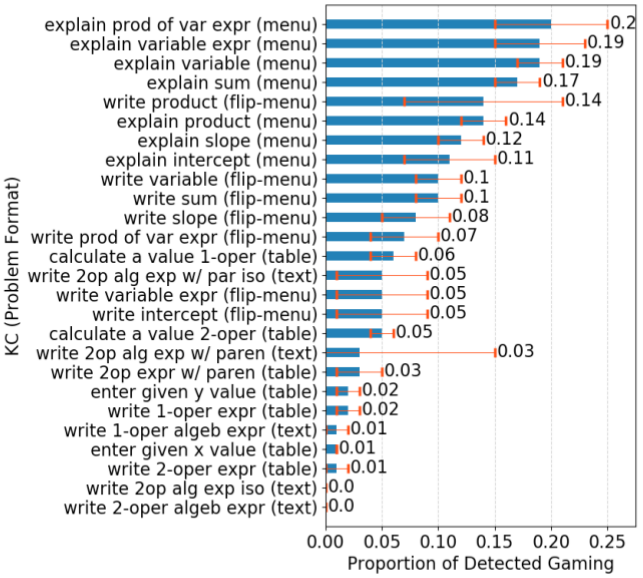

Figure 2. Detected gaming proportion by KCs averaged over students. 95% confidence intervals are plotted. (Only first attempts of steps with KCs are considered.)

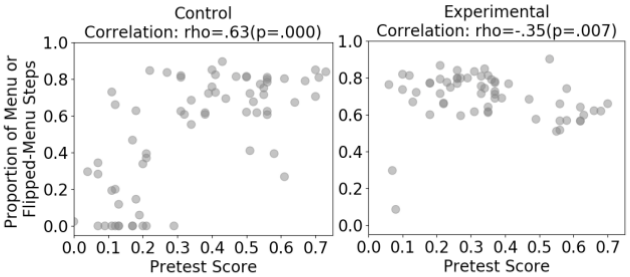

Figure 3. Correlations between pretest scores and proportion of highly-gamed formats (menu, flipped-menu) per condition.

2.3.2 The basic latent variable model accounting for task formats

Based on the first set of exploratory data analyses, we formulated a basic latent variable model, the simplest form of our proposed IRT-GD, that explains detected gaming by both task formats and students' latent gaming tendencies, analogous to explaining item performance by both item difficulties and students' latent abilities in Rasch model [14], the simplest form of item response theory (IRT) models. To illustrate our model, a student with a high proportion of detected gaming due to having a high proportion of menu steps will not be estimated as having a high gaming tendency if he or she does not game more than the average level of the student population on format steps. The model predicts the binary detected gaming label per transaction (i.e., an attempt on a student-step) *G* asserted by KE-GD, given the student identity and the current format (using a generalized linear mixed model):

$$\text{Detected gaming: } G \sim (1|\text{Student}) + \text{Format} \qquad (1)$$

$$\text{Gaming tendency: } \alpha = \exp(\theta) \qquad (2)$$

where the student identity is modeled as a random factor and the format of the current step is modeled as a fixed factor. Formula (1) is written using the syntax of R’s lme4 package for better replicability; a formal mathematical description is that the log odds of a transaction being labeled as gaming by KE-GD is a linear function of the student’s identity (of which the coefficient is the student’s random intercept θ) and the current format. In formula (2), a student’s gaming tendency α is obtained by exponentiating the student’s random intercept θ from formula (1), converting log odds scale to odds scale. This basic model improved over KE-GD in terms of the sign and strength of the association with learning (Table 2 row #1), but the statistical significance was insufficient, demanding further investigation.

2.3.3 *Identifying other contextual factors*

Based on prior literature, we hypothesized that students’ prior and dynamic knowledge levels (i.e., learning) might also account for detected gaming. The theoretical foundation can be found in several studies: [29] showed that avoiding help and failing repeatedly (which may be considered as systematic guessing, a form of gaming) is associated with better learning than seeking help on steps for which students have low prior knowledge; [31] suggested that the behavior of bypassing abstract hints in search of a concrete solution (traditionally considered as help abuse, a form of gaming) may be an engaged learning behavior where students use bottom-out hints as worked examples; [13] also suggested that detected gaming can be a desirable adaptive learning behavior when students encounter challenges far beyond their abilities. Students’ dynamic knowledge levels were often included as features in machine-learned gaming detectors [6, 32] but absent in KE-GD. Thus, we conducted further exploratory data analysis to examine the effect of prior knowledge (approximated by pretest scores) and dynamic knowledge (approximated by practice opportunities) on detected gaming. More specifically, since we had already identified task formats as an important contextual factor, we hypothesized that there might be interactions between prior knowledge and formats, as well as between practice opportunities and formats, on detected gaming.

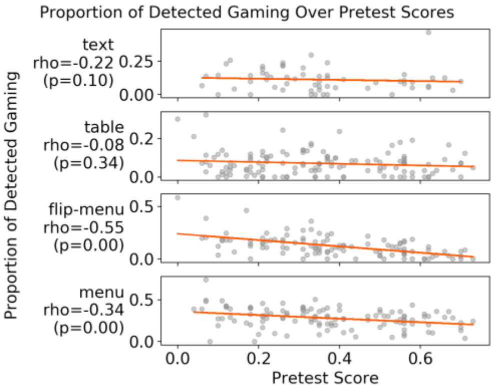

Figure 4. Correlations between pretest scores and detected gaming proportion per task format over students (considering all attempts of all steps).

Figure 4 shows that on flipped-menu and menu formats, students with lower pretest scores gamed much more than students with higher pretest scores, while this was not the case for other formats. Figure 5 shows that on table formats, students were more likely to game on earlier than later opportunities and reduced gaming quickly over opportunities. Discussion of these findings can be found later in Section 5. Based on this second set of exploratory data analysis, we integrated the discovered contextual factors into a latent variable model explained below.

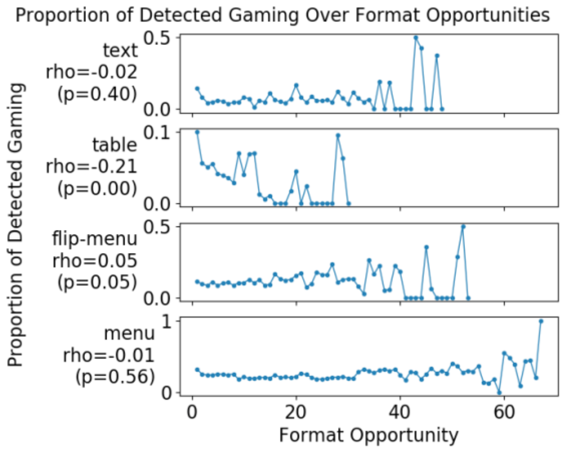

Figure 5. Correlations between practice opportunities and detected gaming per task format. Each point corresponds to the average proportion of detected gaming at an opportunity over students (considering all attempts of all steps). The blips at the end of the curves are due to small sample sizes.

2.3.4 *The full latent variable model accounting for critical contextual factors*

Based on the second set of exploratory data analyses, we identified two groups of contextual factors that might be important to explain general gaming behaviors: the first group captures the effect of pretest scores adjusted by formats; the second group captures the effect of learning adjusted by formats. We then estimated students’ latent gaming tendencies accounting for these contextual factors. The underlying rationale of our model can also be explained as follows. Since the proportion of detected gaming is usually a low proportion of a full dataset under a study (typically less than 7% in past studies and less than 20% in our datasets), it is sound to assume that a model describing well the detected gaming behaviors of a full dataset captures the *normative* behaviors of a population, and the deviation from the normative behaviors represents the intended gaming construct. Essentially, students are estimated as having a higher gaming tendency when they game in less frequently gamed contexts than when they only game in more frequently gamed contexts according to the general behaviors. This is analogous to IRT models where students are estimated as having higher ability when they can get harder items correct than when they only get easier items correct. The full formulation of our latent variable model IRT-GD is as follows (using a generalized linear mixed model):

$$\begin{aligned}\text{Detected gaming: } G \sim\ & (1|\text{Student}) + \text{Format}\\ & + \text{Pretest} + \text{Pretest:Format}\\ & + \text{Opportunity} + \text{Opportunity:Format} \qquad (3)\end{aligned}$$

$$\text{Gaming tendency: } \alpha = \exp(\theta) \qquad (4)$$

Formula (3) is written using the syntax of R’s lme4 package for better replicability; a formal mathematical description is that the log odds of a transaction being labeled as gaming by KE-GD is a linear function of the student’s identity (of which the coefficient is the student’s random intercept θ), the format of the current step, the student’s pretest score, the interaction between the pretest score and the format, the practice opportunity count of a format of the student (note that all steps of a task are considered as having the same opportunity count of the corresponding format), and the interaction between the opportunity count and the format. Except for the student identity modeled as a random factor, all other predictors are modeled as fixed factors. In formula (4), a student's gaming tendency α in odds scale is obtained by exponentiating the student's random intercept θ from formula (3).

Table 2. Associations between gaming tendencies estimated by different variants of IRT-GD and learning. Correlations with normalized learning gains and coefficients of gaming tendency variables in regression predicting posttest scores are reported (p<.10: *boldfaced and italicized*; p<.05: boldfaced).

ID	Level	Predictors	Cor with NLG	Post~ Pre+G

1	Format	(1|Stu)+F	rho=-.07
			p=.44
				b=-.02
				p=.31

2	Format	(1|Stu)+F+Pre	rho=-.14
			p=.11
				b=-0.03
				p=.16

3	Format	(1|Stu)+F+Pre+Pre:F[1]	***rho=-.16***
			p=.07
				b=-0.03
				p=.14

4	Format	(1|Stu)+F+Pre+Pre:F+Opp	***rho=-.16***
			p=.07
				b=-0.03
				p=.14

5	Format	(1|Stu)+F+Pre+Pre:F+Opp+F:Opp (The final chosen model)	**rho=-.18** **p=.04**	***b=-0.04*** ***p=.09***
6	Format	(1|Stu)+F+Pre+Pre:F+Opp+F:Opp (1st attempts of steps w/ KCs)	**rho=-.26** **p=.00**	**b=-0.08** **p=.01**
7	KC	(1|Stu)+K+Pre+Pre:K+Opp+K:Opp[2] (1st attempts of steps w/ KCs)	**rho=-.25** **p=.00**	**b=-0.07** **p=.01**

Table 2 shows the construct validity metrics of full models (row #5-#7) as well as reduced models (row #1-#4) of IRT-GD. All the seven variants reached higher validity than KE-GD in terms of having stronger associations with learning, and the three full models reached desirable statistical significance (row #5-#7). The five predictors increasingly strengthened the association (except when adding the single opportunity term in row #4 before adding the interaction term) and were necessary for reaching acceptable validity in this dataset. In formulating the full models, we explored two other configurations: one that used KCs as the unit (row #7) and fit the model using first attempts of steps with KC labels (without modifying the detected gaming labels associated with these transactions); another that used the same data subset as the KC-level model to fit the model but maintaining the unit of format. We found that a format-level modeling worked as well as the KC-level modeling, when using the same subset (row #6 vs. #7). We also found that using the subset with only first attempts of steps labeled with KCs could improve validity compared to using all attempts of all steps (row #6 vs. #5) in this dataset. However, using all attempts of all steps does not require additional KC labels, so we chose to fit IRT-GD with all attempts of all steps for potentially greater generalizability. The final chosen model for the rest of the paper was the one in row #5 in Table 2. We further examined the fitted parameters of the chosen model (Table 3) and found that they had high consistency with the patterns observed in our exploratory data analyses (note that some differences may be due to the differences in statistical methods and data processing used in the two kinds of analyses). We thus concluded the formulation of IRT-GD for valid gaming detection in our tutor.

Table 3. Parameters of the chosen full model of IRT-GD (row #5 in Table 2). Categorical variables were dummy coded and continuous variables were standardized for reducing multicollinearity. The coefficients are in log odds scale.

Modeling purpose	Regression term	Coefficient
Effect of format	Intercept (Text)	β=-2.09, p<.001 ***
	Table	β=-1.50, p<.001 ***
	FlipMenu	β=0.09, p=.03 *
	Menu	β=1.12, p<.001 ***
Effect of prior knowledge adjusted by formats	Pretest (Pretest:Text)	β=-0.09, p=.13
	Pretest:Table	β=0.04, p=.37
	Pretest:FlipMenu	β=-0.19, p<.001 ***
	Pretest:Menu	β=-0.10, p=.01 *
Effect of learning adjusted by formats	Opp (Opp:Text)	β=0.01, p=.71
	Opp:Table	β=-1.13, p<.001 ***
	Opp:FlipMenu	β=-0.00, p=.99
	Opp:Menu	β=-0.01, p=.84

3. GENERALIZABILITY OF IRT-GD

In the previous section, we conducted exploratory data analysis and validity evaluation on the same dataset and on a single dataset, which might risk overfitting to the dataset. In this section, we tested the generalizability of IRT-GD to two new datasets. We looked into conditions separately and together for all three datasets, resulting in nine contexts across different populations and designs of the system. The two new datasets were collected in 2020 Spring (20S) and 2021 Fall (21F) from the second and third experiments with the tutor with some design changes derived from data mining in the experimental (EXP) condition: new units were introduced for providing focused practice on prerequisite KCs; a lower proportion of menu and flipped-menu tasks was positioned in earlier units compared to the first dataset; a new task format was introduced in the 21F dataset involving interactions with animations. The four task formats identified in the first dataset were still present in the two new datasets. On the other hand, the control condition remained the same.

Table 4 shows statistics of all datasets including detected gaming by KE-GD. Again, the detected gaming proportions were high (16%) in the new datasets. When applying both KE-GD and IRT-GD to the nine contexts (see Table 5), IRT-GD consistently outperformed KE-GD in reaching higher associations with learning in all nine contexts, except one (21F dataset the EXP condition) where the correlation of IRT-GD was slightly weaker but of the same level of significance as KE-GD. In particular, when examining both conditions together and the EXP condition, IRT-GD reached high construct validity (i.e., *rho*<0 and *p*<.05) in all six contexts, while KE-GD only reached construct validity in half of the contexts. When examining the control condition, IRT-GD also improved on KE-GD by reversing positive correlations to the theoretically consistent negative correlations for all datasets and reached acceptable significance on the 20S dataset, although the correlations did not reach acceptable significance in other datasets. We conducted further investigation next.

[1] In R's lme4 package, a colon : is used to denote an interaction term.

[2] We treated the KC variable as a random factor. R formula:
G~(1|Stu)+(1+Pre+Opp|K)+Pre+Opp.

Table 4. Statistics of datasets including detected gaming by KE-GD (CT/EXP: control/experimental condition).

Data	#stu			#transactions	Avg proportion of gamed tx over stu
	All	CT	EXP		
19F	129	69	60	98,176	.18 (*SD*=.08)
20S	222	106	116	109,193	.16 (*SD*=.11)
21F	99	46	53	59,703	.16 (*SD*=.11)

Table 5. Associations between gaming from KE-GD or IRT-GD with learning across nine contexts. The Gaming variables in regression models predicting posttest scores for KE-GD and IRT-GD are of different scales. (p<.10: *boldfaced and italicized*; p<.05: boldfaced; NLG: normalized learning gain; CT: control condition; EXP: experimental condition.)

Da-ta	Detect-or	All		CT		EXP	
		Cor w/ NLG	Post~ Pre+G	Cor w/ NLG	Post~ Pre+G	Cor w/ NLG	Post~ Pre+G
19 F	KE-GD	rho=-.02 p=.86	b=0.07 p=.69	rho=.14, p=.25	b=0.34 p=.11	**rho=-.29 p=.02**	**b=-0.71 p=.02**
	IRT-GD	**rho=-.18 p=.04**	***b=-0.04 p=.09***	rho=-.02. p=.86	b=-0.01 p=.62	**rho=-.41 p=.00**	**b=-0.10 p=.02**
20 S	KE-GD	rho=-.04 p=.55	b=0.01 p=.92	rho=.16, p=.10	b=0.25 p=.10	rho=-.13 p=.17	b=-0.32 p=.12
	IRT-GD	**rho=-.20 p=.00**	**b=-0.04 p=.00**	***rho=-.19 p=.05***	**b=-0.05 p=.03**	**rho=-.21 p=.02**	***b=-0.03 p=.06***
21 F	KE-GD	**rho=-.29 p=.00**	**b=-0.45 p=.00**	rho=.00 p=.98	b=-0.04 p=.85	**rho=-.52 p=.00**	**b=-0.83 p=.00**
	IRT-GD	**rho=-.36 p=.00**	**b=-0.05 p=.00**	rho=-.20 p=.18	b=-.02 p=.19	**rho=-.48 p=.00**	**b=-0.09 p=.00**

3.1 Identifying deeper task format effects for refining the input detector

To understand and address the lack of validity of IRT-GD in the control condition in two datasets (Table 5), we conducted further investigation on the 19F dataset where IRT-GD showed the weakest association with learning. We wondered whether the bottleneck lay in the input detector KE-GD. If the gaming labels (i.e., values of the dependent variable for fitting IRT-GD) were too noisy, it would be hard to get accurate tendency estimates by any means. If we decompose a gaming label, it is the union of 13 gaming labels corresponding to 13 gaming patterns defined in KE-GD. Could some of the patterns under some formats be better considered as not gaming in our control condition context? In other words, we hypothesized that there might be deeper task format effects in students' interaction patterns. We conducted a third set of exploratory data analysis where we examined the associations between detected gaming proportions of each of the 13 patterns from KE-GD with learning under each format. We used a *local* normalized learning gain computed using tasks related to a specific format rather than all tasks in the pretest and posttest. The results in Table 6 suggest that on different formats, the same gaming pattern could be helpful or harmful for learning, supporting our hypothesized deeper format effect. To account for this contextual factor, we updated the detected gaming labels from KE-GD in the control condition by using the union of only the patterns that were negatively associated with learning (regardless of statistical significance) for each format while maintaining the labels of the experimental condition. This was a change in the dependent variable rather than the predictors in IRT-GD. We used the updated dataset to fit new IRT-GD variants, referred to as IRT-GD-PR, and estimated gaming tendencies for the control condition and the overall dataset. Table 7 shows that IRT-GD-PR achieved acceptable validity for the control condition and also boosted the validity for the overall dataset compared to IRT-GD and KE-GD. We leave for future work to further improve and test this local refinement method.

Table 6. Correlations between local normalized learning gains and proportion of each gaming pattern detected by KE-GD in the control condition in the 19F dataset. Pattern #8 was omitted due to its absence. (+: rho>0, -: rho<0, ·: p<.10)

Format	1	2	3	4	5	6	7	9	10	11	12	13
Avg prop	.01	.08	.01	.00	.13	.00	.01	.01	.01	.00	.01	.02
Table	-	+	-	-	+	-·	-	-	-	+	-	+
Menu	+	-	+·	-	-	-	+	-	-	-	-·	-
Flip-M	+	-	+	na	-	-	+	+	-	+	-	+

Table 7. Associations between gaming (from KE-GD, IRT-GD, or IRT-GD-PR) with learning. Rho and p values are reported for correlation with normalized learning gains; coefficients and p values of Gaming variables are reported for regression.

Detector	All		CT (control condition)	
	Cor w/ NLG	Post~Pre+G	Cor w/ NLG	Post~Pre+G
KE-GD	-.02(.86)	0.07(.69)	.14(.25)	0.34(.11)
IRT-GD	**-.18(.04)**	***-0.04(.09)***	-.02(.86)	-0.01(.62)
IRT-GD-PR	**-.27(.00)**	**-0.07(.01)**	***-.23(.06)***	**-0.06(.04)**

4. APPLICATIONS OF IRT-GD

In this section, we demonstrated two applications of IRT-GD. We used the estimated gaming tendencies from IRT-GD to study whether there was a difference in the level of gaming between the two conditions from our experimentation with the tutor, and to explore the relation between gaming and motivation.

4.1 Intervention effects on gaming

Our prior work [18] has shown that the data-tuned adaptive condition (that provided intense deliberate practice) led to greater learning outcomes compared to the control condition (that provided normal deliberate practice) in the first experiment (19F dataset); we are interested to see whether the intervention also led to higher behavioral engagement, particularly lower levels of gaming the system. We conducted a regression analysis predicting levels of gaming over students given the condition indicator on the three datasets. The two detectors had contradicting results on the 19F and 20S datasets. On the 19F dataset, KE-GD showed that the intervention led to significantly higher levels of gaming while IRT-GD showed that there was no statistical difference (Table 8 the 2nd column). The suggested intervention effect of increased gaming levels by KE-GD contradicted the previously validated intervention effect of improved learning, since higher levels of gaming are usually associated with lower learning. Thus, IRT-GD more accurately revealed the intervention effect on this dataset. We hypothesized that this could be due to KE-GD not being able to account for the task format effect. We computed the proportion of highly-gamed formats over transactions and the normalized learning gain per student per condition. We found that the EXP condition had a higher average proportion of highly-gamed formats (Table 9 the 2nd column), consistent with our hypothesis. On the

20S dataset, KE-GD showed that the intervention led to significantly lower levels of gaming while IRT-GD showed that there was no statistical difference (Table 8 the 3rd column). However, both conditions have similar normalized learning gains, and the control condition had a much higher average proportion of highly-gamed formats (Table 9 the 20S columns). This again suggests that KE-GD provided biased gaming assessment by using direct proportion of gaming without accounting for formats. This set of analyses shows that IRT-GD more accurately revealed intervention effects on gaming than KE-GD in our experiments.

Table 8. Intervention effects on gaming examined by regression predicting gaming proportions or tendencies given the condition variable (Control: 0, Experimental: 1). Coefficients of the condition variable are reported.

Detector	19F	20S	21F
KE-GD	**b=0.02, p=.03**	**b=-0.09, p<.001**	b=0.03, p=.15
IRT-GD	b=-0.05, p=.68	b=0.03, p=.79	b=0.16, p=.48

Table 9. The proportion of highly-gamed formats (PHGF) in transactions and normalized learning gain per condition. Mean and *SD* are reported. Higher values are in boldface.

Cond	19F		20S		21F	
	PHGF	NLG	PHGF	NLG	PHGF	NLG
CT	.38(.22)	.16(.29)	**.31(.22)**	.12(.34)	.16(.17)	.14(.20)
EXP	**.44(.09)**	**.24(.28)**	.06(.13)	**.14(.39)**	**.40(.16)**	**.15(.25)**

4.2 Motivation and gaming

The investigation of the relation between motivation and gaming contributes to understanding why students game and developing behavioral measures of motivation. Prior work [9] indicated that students' attitudes and interest towards the domain was related to detected (observed) gaming frequency. More recent work [13] applying a simple latent variable model identified strong associations between several motivational measures and estimated gaming tendencies. Our investigation of the relation between motivation and gaming adds to the limited empirical evidence in this space. On our datasets, motivational surveys with four scales (Table 10) were collected at the first and the last sessions of each month-long experiment. Each question used a 7-point Likert rating; responses for each scale were averaged to present students' motivation along the scale. Table 11 shows correlations between motivational measures from surveys and estimated gaming tendencies over students. Among the four scales, only perceived competence in math (PC) showed consistent significant correlations with gaming and only in the experimental condition across three datasets; the sign of the correlations was negative as theoretically predicted. The correlations between PC and gaming did not appear to be due to students' abilities approximated by pretest scores, because we did not find correlations between pretest scores and gaming tendencies. To understand why PC was only associated with gaming in the experimental condition that provided intense deliberate practice but not in the control condition that provided normal deliberate practice, we compared objective difficulties measured by the proportion correct of first attempts and subjective difficulties measured by the difference between the final and the initial values of PC between the conditions (Table 12). We found that the experimental condition had lower objective difficulties but higher subjective difficulties. We discussed the results in the next section.

Table 10. Motivational survey inventory.

Scale	Question
Perceived competence in math (PC)	How good at math are you?
	Compared to most of your other school subjects, how good are you at math?
Math utility value (UV)	How important is it to you to learn math?
	How important do you think math will be to you in the future?
Interest in math (IM)	How interesting is math to you?
Interest in tutor (IT)	How excited are you to do math on a computer?

Table 11. Correlations between motivational measures from surveys and estimated gaming tendencies. Correlations with pretest scores were added for contrast.

Scale	Cond	19F	20S	21F
PC	CT	-.00(.97)	-.10(.31)	-.03(.84)
	EXP	**-.26(.046)**	**-.18(.05)**	**-.32(.02)**
	All	-.11(.20)	**-.15(.03)**	-.12(.25)
UV	CT	-.11(.38)	.00(.98)	**-.31(.04)**
	EXP	.09(.50)	-.03(.78)	-.17(.22)
	All	.01(.94)	-.02(.78)	**-.21(.04)**
IM	CT	.04(.73)	-.03(.75)	-.05(.73)
	EXP	-.07(.60)	-.13(.18)	-.11(.41)
	All	.02(.87)	-.08(.21)	-.04(.66)
IT	CT	.11(.37)	-.03(.75)	-.01(.97)
	EXP	.06(.64)	-.06(.55)	-.01(.93)
	All	.12(.19)	-.04(.54)	-.01(.93)
Pretest	CT	-.04(.74)	-.01(.91)	-.06(.68)
	EXP	-.01(.92)	-.03(.74)	-.06(.66)
	All	-.01(.90)	-.04(.61)	-.07(.52)

Table 12. Objective difficulties measured by the proportion correct of first attempts (prop cor) and subjective difficulties measured by the difference of PC between the final value and the initial value (*Δ*PC) per condition. Mean and *SD* are reported. Higher values are in boldface.

Cond	19F		20S		21F	
	prop cor	*Δ*PC	prop cor	*Δ*PC	prop cor	*Δ*PC
CT	.60(.17)	**.02(.86)**	.60(.14)	**-.03(.83)**	.61(.16)	**.02(1.11)**
EXP	**.62(.10)**	-.22(.91)	**.69(.11)**	-.09(.86)	**.70(.10)**	-.22(.98)

5. DISCUSSION AND CONCLUSION

In this paper, we demonstrate a latent variable model for more valid and robust gaming assessment, item response theory-based gaming detection (IRT-GD), that estimates latent student gaming tendencies accounting for contextual factors. We started with applying a previously validated knowledge-engineered gaming detector (KE-GD) to a dataset collected from an algebra tutor with varying task design and sequencing across conditions. However, the detected gaming level by KE-GD was not associated with learning, challenging its construct validity in our context. We conducted exploratory data analyses and identified contextual factors that could capture the normative interaction behaviors of the population that might explain this lack of association. We then built an IRT-GD model that explains detected gaming from KE-GD by both contextual factors and students' intrinsic gaming tendencies; it estimates a student-level latent gaming tendency as the degree of deviation from normative behaviors of a population across contexts. We tested the generalizability of IRT-GD and found that it outperformed KE-GD on three datasets across different contexts in construct validity measured by associations with learning. Our approach is not only useful for others wanting to apply a gaming assessment in their context, but is also generally applicable in creating more robust behavioral measures.

There are two notable features of our approach that may be particularly relevant for anyone building or using behavioral detectors. One is that our modeling approach adapts an existing behavioral detector to new contexts without complex feature engineering, which may be attractive for the learning engineering community to maximally build on past methods and adapt them to new contexts. For example, the learning effect on detected gaming is incorporated through practice opportunity counts without an additional process to estimate dynamic knowledge as in [6, 32]. Another feature is that our modeling and evaluation approaches do not require extra human labeling and focus on the association between the behavior measure and learning. Many past works constructed and validated detectors solely by predictions of human labels; although human labels have undeniable merits, they may contain bias. For example, in the development of KE-GD [25], experts examined each clip, which consists of five consecutive actions, from a set of clips randomly selected from log data and decided whether the clip would be coded as gaming or not. A clip was shown in a textual format giving *individual-level* information about the actions *within* the clip (e.g., each action's time, the problem context, the input entered, the relevant skill, whether the input was right, wrong, a help request or a "bug"), and experts made judgements about gaming without *population-level* information (e.g., the median time of the step of the population), or information *outside* the clip from previous or future clips. This may increase the speed and ease of labeling, yet it may risk introducing bias. For example, if the student did not deviate much from the general behavior of the population or if the student could get a similar step correct in a future clip on their first attempt, then it may be better to label this clip as not gaming. Thus, behavioral detectors validated solely by predictions of human labels looking at isolated clips may not always reliably capture unproductive or harmful behaviors for learning. Our approach reduces bias and enhances the support for learning when applying a behavioral detector by considering contextual factors that were not considered in the original human labeling process, but are important for identifying behaviors harmful for learning. Although further examination of generality, stability, and reliability (as elaborated later) of IRT-GD may be needed to strengthen the validity claim of our approach, we think current evidence suffices to suggest that IRT-GD and our latent variable modeling approach can *enhance* (rather than replace) existing behavior measures for more valid and most robust behavioral assessment. One may consider using IRT-GD and our latent variable modeling approach when an existing behavior measure lacks validity in a specific context.

We identified strong contextual factors, i.e., the task format and its interaction with students' practice opportunities, aligning with previous research. The menu format led to the highest detected gaming, which coheres with prior work hinting at the high propensity of the multiple-choice format (which also involves selecting an option given a set of options) for triggering detected gaming. One explanation may be that the cognitive cost [16] in making attempts in menus is low since it does not require typing, and different descriptions may only have subtle differences, so students might have developed a trial-and-error strategy with genuine engagement. This explanation could also be applied to the second highest gamed format, flipped-menu, where students might enter several expressions extracted from a given equation corresponding to a description rather than writing expressions from scratch as in other formats. Cognitive cost of a format is implicitly considered in IRT-GD and may be worth more attention for others developing gaming detectors. Meanwhile, students more likely decreased detected gaming as their prior knowledge levels increased on menu and flipped-menu formats compared to other formats, suggesting that certain game-like learning strategy (e.g., a trial-and-error strategy) may only be likely when the cognitive cost is low and students are of low prior knowledge. Moreover, we found that students decreased detected gaming faster on the table format over successive practice opportunities than on other formats, suggesting the reasons for students to game on this format might be different from the reasons they gamed on menus or flipped-menus. Examining the interface, one explanation may be that there are no clear instructions on how to fill in the various cells of the table, e.g., under the column labeled as "Show your work", it is not clear whether a student could enter *15+10* (graded as wrong) instead of *3*5+10* (the correct answer). This coheres with prior work suggesting that students gamed more when the presentation is unclear [8], and that students may game as a way to obtain worked examples [31]. Further analysis on student answers may support our hypothesized explanations. A final remark regarding task formats is that in the tutor we studied, the interpretation of task formats requires caution since a task format is not only coupled with a specific interface design (as the name *format* suggests), but also a specific scaffolding design (e.g., fixed or dynamic scaffolding) as well as specific KCs. A future direction is to study them separately through experimentation.

In a context where IRT-GD did not reach statistically significant associations with learning, we conducted local refinement of the input detector, KE-GD, by considering deeper format effects, i.e., the interaction between formats and specific interaction patterns. Our refinement led to acceptable validity and further confirms the importance of task features and demonstrates the flexibility of our latent variable approach. A next step is to test whether this local refinement approach is robust in other contexts. In some contexts, KE-GD already reached acceptable validity, although IRT-GD further improved on it by reaching stronger associations with learning. A next step is to apply IRT-GD to other learning environments and to study more automatic ways to identify contextual factors important in a specific context.

One aspect that needs further examination is whether and how well a fitted IRT-GD model extrapolates to unseen students or formats. This aspect is especially relevant to online intervention where the tutor has to react to gaming as designed for new students or formats. In Section 4, we conducted one kind of generalizability checking where we used the same independent variables in the IRT-GD model for the first dataset to construct IRT-GD models for new datasets fitted to the complete set of the new datasets. The new format (*animation*) was handled through adding a new dummy coded variable. We showed that the structure (i.e., predictors) of IRT-GD generalizes to new students and new versions of the system. This checking is most relevant if one uses IRT-GD to conduct offline student-level analysis as was done in Section 4. However, we have not examined how well a fitted IRT-GD model extrapolates to unseen students or formats, i.e., predicts detected gaming or estimates gaming tendencies for unseen students or formats, which is important when using IRT-GD for online intervention. In theory a fitted IRT-GD model can extrapolate to an unseen or newly seen student: we first plug in the values of the fixed factors which equates to using the population mean to obtain a prediction; after observing at least one data point of the new student, we can (repeatedly) reestimate the parameters with the accumulated data with a random intercept added for the new student. Meanwhile, the extrapolation to unseen formats is also feasible: we can first treat a new format as a seen, similar format, and after observing at least one data point of the new format, we can (repeatedly) reestimate the parameters with a parameter fitted for the new format[3]. A promising modification of IRT-GD that enables greater generalizability is to replace the dummy coded format variables with a variable that describes key properties of formats, e.g., *whether a response set is given or can be easily inferred*. As for the question of *how well*, we plan to test the "online" predictiveness of IRT-GD used in the aforementioned ways for extrapolation as a next step.

A related examination that could further support the validity of IRT-GD is to examine the stability and reliability of estimated gaming tendencies. To examine stability, we may check whether gaming tendencies estimated from the first half of students' temporally ordered interactions correlate with those estimated from the second half of the student interactions, i.e., whether students who tend to game more earlier also tend to game more later. To examine reliability, we may check whether gaming tendencies estimated from interactions of a set of formats correlate with those estimated from interactions of other formats, i.e., whether students who tend to game more in some formats also tend to game more in other formats. The higher the validity and reliability, the higher the truthfulness of the underlying assumption of IRT-GD about a latent, stable gaming tendency construct and the soundness of the identified contextual factors.

Although we have focused on a student-level gaming estimate by IRT-GD, it can also give a transaction-level gaming estimate for online intervention. For example, we can first fit an IRT-GD model to past data using the full formulation. Then, we can apply the fitted model without using the random student intercepts to predict whether an average student may game (as defined by KE-GD) on a step according to current contextual factors. Then we compare this population-level prediction (considering an interval of uncertainty) to the gaming label by KE-GD to identify cases where gaming is not acceptable, i.e., deviating too much from the norm, and finally activate the pre-designed intervention.

One seemingly conflicting result with prior studies is that we did not find an association between gaming tendencies with pretest scores (Table 11), where prior studies have shown that lower prior knowledge levels were associated with higher gaming frequencies [5, 20]. This is because IRT-GD already includes pretest scores and relevant interactions as predictors for detected gaming. IRT-GD is intentionally designed to extract latent gaming tendencies that are not (primarily) triggered by prior knowledge, but by other factors such as students' motivation or metacognitive skills. This may lead to tutor design that focuses on promoting students' motivation or metacognition. However, our latent variable modeling approach is flexible in that one could consider dropping the pretest scores related predictors if they are interested in gaming tendencies triggered by prior knowledge.

One finding seemingly less consistent with prior work and harder to interpret is the link between motivation and gaming. We found a negative correlation between perceived competence in math and gaming in the experimental condition (i.e., the intense deliberate practice condition), consistent with the reported negative correlation between self-efficacy in math and gaming in [13], but we did not find any correlations between other motivational measures and gaming, such as students' interest towards the domain and gaming reported in [9, 13], or any correlations in the control condition (i.e., the normal deliberate practice condition)[4]. Rather than prematurely attributing the general lack of correlation between motivation and gaming to the lack of validity of estimated gaming tendencies, we hypothesize several reasons. There may be interactions between different student attributes (measured or unmeasured in the current study) or between student attributes and system attributes not considered in a simple zero-order correlation we did here. Additionally, the motivational survey was deployed at the first session but was used to correlate with month-long accumulated behaviors. After all, there is still limited empirical evidence of the relation between motivation and gaming, so further investigation is needed. To explain why there was a negative correlation between perceived competence in math and gaming in the intense deliberate practice condition but not in the normal deliberate practice condition, we conducted a preliminary exploration and found that the objective difficulty (measured by the proportion correct of first attempts) of the intense deliberate practice condition was lower than the normal deliberate practice condition but the subjective difficulty (measured by perceived competence in math) of it was higher. One hypothesis is that the patterns of successes or failures may matter more than the proportion of success for students' perceived competence. The intense deliberate practice driven by a more fine-grained and larger KC model may have more constantly pushed students to work on their weak spots in new tasks (i.e., put them on the edge of competence), challenging their perceived competence. It may be worth considering letting students to occasionally work on already mastered skills to boost their perceived competence, or preparing students better for desirable difficulties or failures. Combining this finding with the finding that intense deliberate practice alone did not reduce gaming tendencies, one promising direction is to introduce motivational interventions or designs that could maintain or promote perceived competence or self-efficacy in the task domain under intense deliberate practice, to reach a potential multiplier effect of both cognitive and motivational interventions.

6. ACKNOWLEDGMENTS

This work was supported by Bill and Melinda Gates Foundation Prime Award #OPP1196889. We also thank all reviewers for their valuable feedback which helped us improve our manuscript.

7. REFERENCES

[1] Aleven, V., McLaren, B., Roll, I. and Koedinger, K. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), pp.101-128.

[2] Almeda, M.V. and Baker, R.S. 2020. Predicting student participation in STEM careers: The role of affect and

[3] Treating a categorical variable with few levels as a random factor may lead to imprecise estimates [10]. Thus, we do not consider this as a next step when the number of formats is small (e.g., <10).

[4] The positive (rather than the expected negative) correlation rho values in some cells under the 19F column in Table 11 are considered as a result of statistical noise, since none of these values are even marginally significant.

engagement during middle school. *Journal of Educational Data Mining, 12*(2), 33-47.

[3] Baker, R.S., Corbett, A.T., and Koedinger, K.R. 2004. Learning to distinguish between representations of data: A cognitive tutor that uses contrasting cases. In *Proceedings of the International Conference of the Learning Sciences* (pp. 58-65).

[4] Baker, R.S., 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. June. Adapting to when students game an intelligent tutoring system. In *International conference on intelligent tutoring systems* (pp. 392-401). Springer, Berlin, Heidelberg.

[5] 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 SIGCHI conference on Human factors in computing systems* (pp. 383-390).

[6] Baker, R.S., Corbett, A.T., Roll, I., and Koedinger, K.R. 2008. Developing a generalizable detector of when students game the system. In *User Modeling and User- Adapted Interaction, 18*(3), 287-314.

[7] Baker, R. S. and de Carvalho, A. M. J. A. 2008. Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 38-47).

[8] Baker, R. S., de Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T., and Koedinger, K. R. 2009. Educational software features that encourage and discourage "gaming the system". In *Proceedings of the 14th international conference on artificial intelligence in education* (pp. 475-482).

[9] 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.

[10] Bolker, Ben. 2022. GLMM FAQ. https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#should-i-treat-factor-xxx-as-fixed-or-random

[11] Cen, H. 2009. *Generalized Learning Factors Analysis: Improving Cognitive Models with Machine Learning.* Doctoral Thesis. Carnegie Mellon University.

[12] Cocea, M., Hershkovitz, A., and Baker, R. S. 2009. The impact of off-task and gaming behaviors on learning: immediate or aggregate?. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A. Graesser (Eds.), *Artificial Intelligence in Education* (pp. 507-514). IOS Press.

[13] Dang, S. and Koedinger, K. 2019. Exploring the Link Between Motivations and Gaming. *Proceedings of The 12th International Conference on Educational Data Mining*, pp. 276 - 281.

[14] De Boeck, P. and Wilson, M. 2004. *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.

[15] Fancsali, S. 2014. Causal discovery with models: behavior, affect, and learning in cognitive tutor algebra. In *Educational Data Mining 2014*.

[16] Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., and Welsh, M. E. 2015. Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology, 41*, 232-244.

[17] González-Brenes, J., Huang, Y., and Brusilovsky, P. 2014. General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In *The 7th International Conference on Educational Data Mining* (pp. 84-91).

[18] Huang, Y., Lobczowski, N.G., Richey, J.E., McLaughlin, E.A., Asher, M.W., Harackiewicz, J.M., Aleven, V., and Koedinger, K.R. 2021. A general multi-method approach to data-driven redesign of tutoring systems. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 161-172).

[19] Koedinger, K.R. and Corbett, A. 2006. Cognitive tutors: Technology bringing learning sciences to the classroom. In *The Cambridge handbook of the learning sciences*, 61-77.

[20] Mogessie, M., Elizabeth Richey, J., McLaren, B.M., Andres-Bray, J.M.L., and Baker, R.S. 2020. Confrustion and gaming while learning with erroneous examples in a decimals game. In *International Conference on Artificial Intelligence in Education* (pp. 208-213). Springer, Cham.

[21] Muldner, K., Burleson, W., Van de Sande, B., and VanLehn, K. 2011. An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts. *User Modeling and User-Adapted Interaction, 21*(1), 99-135.

[22] Paquette, L. and Baker, R. S. 2017. Variations of gaming behaviors across populations of students and across learning environments. *International Conference on Artificial Intelligence in Education* (pp. 274-286). Springer, Cham.

[23] 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), 585-597.

[24] Paquette, L., Baker, R.S., de Carvalho, A.M.J.A., and Ocumpaugh, J. 2015. Cross-system transfer of machine learned and knowledge engineered models of gaming the system. In *Proceedings of the 23rd Conference on User Modeling, Adaptation and Personalization* (pp. 183-194).

[25] Paquette, L., de Carvalho, A. M., and Baker, R. S. 2014. Towards Understanding Expert Coding of Student Disengagement in Online Learning. In *Proceedings of the 36th Annual Cognitive Science Conference*, 1126- 1131.

[26] Pardos, Z.A., Baker, R.S., 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.

[27] Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K.R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., and Livak, T. 2005. The ASSISTments project: Blending assessment and assisting. In *Proceedings of the 12th annual conference on artificial intelligence in education* (pp. 555-562).

[28] 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 *International Conference on*

Artificial Intelligence in Education (pp. 342-355). Springer, Cham.

[29] Roll, I., Baker, R. S. D., Aleven, V., and Koedinger, K. R. 2014. On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, *23*(4), 537-560.

[30] San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., and Heffernan, N.T. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Proceedings of the 6th International Conference on Educational Data Mining,* 177-184.

[31] Shih, B., Koedinger, K. R., and Scheines, R. 2008. A Response Time Model For Bottom-Out Hints as Worked Examples. *Educational Data Mining 2008,* 117-126.

[32] Walonoski, J. A. and Heffernan, N. T. 2006a. Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. *Proc of ITS 2006*, 382-391.

[33] Walonoski, J.A. and Heffernan, N.T. 2006b. Prevention of off-task gaming behavior in intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 722-724), Jhongli, Taiwan. Berlin:Springer-Verlag.