Variations of Gaming Behaviors Across Populations of Students and Across Learning Environments

Luc Paquette[1], Ryan S. Baker[2]

[1]University of Illinois at Urbana-Champaign, Champaign, IL
[2]University of Pennsylvania, Philadelphia, PA
lpaq@illinois.edu, rybanker@upenn.edu

Abstract. Although gaming the system, a behavior in which students attempt to solve problems by exploiting help functionalities of digital learning environments, has been studied across multiple learning environments, little research has been done to study how (and whether) gaming manifests differently across populations of students and learning environments. In this paper, we study the differences in usage of 13 different patterns of actions associated with gaming the system by comparing their distribution across different populations of students using Cognitive Tutor Algebra and across students using one of three learning environments: Cognitive Tutor Algebra, Cognitive Tutor Middle School and ASSISTments. Results suggest that differences in gaming behavior are more strongly associated to the learning environments than to student populations and reveal different trends in how students use fast actions, similar answers and help request in different systems.

Keywords: Gaming the system, intelligent tutoring system, student populations

1 Introduction

Studies of students who “game the system”, a disengaged behavior in which students "attempt 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" [1], have shown its relationship with poorer learning outcomes [2, 3, 4, 5], increased boredom [6] and lower long-term levels of academic attainment [7]. Research on gaming the system has applied machine learning [1, 8, 9] and knowledge engineering [9, 10, 11, 12, 13, 14] approaches to build models able to detect gaming from data sets collected from a specific population of students in a specific learning environment. Although data collected across multiple learning environments provides us with information about how often students typically game different environments, little work has focused on explicitly comparing gaming across environments and populations of students. In one exception, Baker and Gowda [15] compared the incidence of disengaged behaviors across populations of students using Cognitive Tutor Algebra, finding that the incidence of gaming was different across populations. However, their work did not investigate whether the nature of the gaming behaviors differed across populations.

adfa, p. 1, 2011.

Recent work in the creation of models of student affect (which are often similar to models of gaming the system) has found that they may not always transfer between different populations of students [16]. By contrast, recent work has suggested that models of gaming the system can, in some cases, function reliably in new learning environments, albeit with some degradation in performance [17]. To better understand this degradation and how we may be able to develop more universal models of this important behavior, we study whether the specific ways that gaming the system manifests varies between environments and populations of students.

In this paper, we study differences in the patterns of gaming behaviors demonstrated by students from three different populations of students using the Cognitive Tutor Algebra [18] environments and within three different learning environments: Cognitive Tutor Algebra [18], Cognitive Tutor Middle School [18] (an earlier version of Cognitive Tutor Bridge to Algebra), and ASSISTments [19]. To do so, we compared the relative frequencies of 13 patterns of actions associated with gaming that were previously identified in a cognitive model of gaming behaviors [20]. Results from our study showed stronger differences in gaming patterns across environments than across student populations, suggesting that the ways that gaming manifests behaviorally are more strongly associated with the environment than with specific student populations.

2 Method

2.1 Model

In previous work, we studied how experts identify whether students are gaming the system by analyzing text replays, textual representations of the students' actions in the learning environments [20]. To do so, we conducted a cognitive task analysis [21, 22] of how an expert identifies behaviors as gaming or not. This was achieved using a combination of active participation [22, 23] (in which the person performing the cognitive task analysis actively participated in the coding of text replays), think aloud observations [24] and interviews to explicate the coding process.

Using this technique, 13 patterns of actions were identified that each captured part of the gaming behavior in Cognitive Tutor Algebra. Those 13 patterns were used to build a cognitive model able to detect gaming from sequences of student actions in the learning environment. In this model, a sequence of actions, called a clip, was classified as gaming the system when the actions it contained matched at least one of the 13 patterns. Table 1 describes the 13 patterns that were identified as gaming (symbolic representations of the patterns are presented in [20]). The most frequent elements of those patterns include: quickly entering answers without thinking, re-entering the same answer in different parts of the problem, entering sequences of similar answers (defined as two consecutive answers with a Levenshtein distance [25] of 1 or 2), quickly asking for help without thinking about what to do next, and moving on to a new part of the problem before correctly solving the part that was previously attempted.

This cognitive model of gaming the system achieved a performance of 0.330 [20] for the Kappa [26] metric, a metric which indicates how much better the model is than chance at identifying gaming behaviors, when applied to an held-out test set. This performance was considerably higher than the previous best model of gaming for Cognitive Tutor Algebra, which obtained a Kappa of 0.24 [27] when cross-validated at the student level, but was less effective than a more recent hybrid model created by improving on the cognitive model using machine learning techniques [28]. This hybrid model combined traditional classification algorithms with the automatic generation and selection of patterns of actions that mirror the structure of those identified during the cognitive task analysis. This model achieved a Kappa of 0.457 under student-level cross-validation [28]. In the current paper, we use the cognitive model for two main reasons. First, the pattern structure of the cognitive model makes it easier to interpret than the hybrid model, an important consideration when comparing gaming behaviors across populations and learning environments. Second, a previous study [17] showed that the performance of the cognitive model was more robust than the hybrid model when applying it to data sets from two new learning environments: the scatter plot lesson in Cognitive Tutor Middle School and ASSISTments.

Table 1. List and descriptions of the 13 patterns contained in our model of gaming the system.

#	*Pattern*
1	The student enters an incorrect answer and then quickly re-enters the same incorrect answer in a different part of the problem.
2	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.
3	The student enters an incorrect answer, followed by a similar and incorrect answer and finally re-enters the second answer in a different part of the problem.
4	The student quickly enters an incorrect answer, followed by quickly entering a second incorrect answer and then, once again, quickly entering a different answer.
5	The student enters an incorrect answer, followed by a similar incorrect answer and then quickly enters a different answer.
6	The student asks for help and quickly looks for the answer in the hints provided by the learning environment, enters an incorrect answer and then enters a similar incorrect answer.
7	The student enters an incorrect answer, followed by the same incorrect answer in a different part of the problem, followed by the student attempting to answer, or requesting help for, a different part of the problem.
8	The student enters a known error (recognized by the system as a “bug”), then re-enters the same answer in a different part of the problem, gets the right answer, and then enters a new bug for a different part of the problem.
9	The student enters an incorrect answer, enters a similar incorrect answer and then moves on to a different part of the problem and enters another incorrect answer.
10	The student enters an incorrect answer, moves on to a different part of the problem, once again enters an incorrect answer and then enters a similar incorrect answer.
11	The student enters an incorrect answer, followed by a similar incorrect answer, doesn’t take the time to think about the error before asking for help and finally enters a similar incorrect answer.
12	The student asks for help, followed by a sequence of 3 incorrect answers with at least 2 of which are similar to each other.
13	The student enters a sequence of 3 incorrect answers, at least 2 of which are similar to each other, and then quickly asks for help without thinking about the errors.

2.2 Data

To study differences in the usage of the 13 patterns of gaming the system, we applied the cognitive model from [20] to 6 different data sets. The first three data sets (obtained from the Pittsburgh Science of Learning Center DataShop [29]) were all collected from students using the Cognitive Tutor Algebra digital learning environment [18]. Each data set was collected from one school and represents a different population of students: rural, suburban and urban students. All three schools were located in the same geographical region of the Northeastern United States with the same nearest urban center. Table 2 presents the number of different classes and teachers who used the system in each data set, a summary of the demographic information for the school and the distance between the school and the nearest urban center.

The second group of three data sets was collected from different learning environments: Cognitive Tutor Algebra, Cognitive Tutor Middle School and ASSISTments. The Cognitive Tutor Algebra data set was created by combining the rural, suburban and urban data sets. The Cognitive Tutor Middle School data set was collected from 2 suburban school districts in the Northeastern United States, in the same region as the Cognitive Tutor Algebra data set. The ASSISTments data set was collected from three school districts in a different part of the Northeastern United States. One of the ASSISTments schools was urban and two were suburban. It is important to note that, although the focus of the analysis for those data sets is the learning environment, the population of students will also vary due to the different geographical regions in which those systems are used and the different age groups targeted by each system.

Table 2. Descriptive statistics for each school that used the Cognitive Tutor Algebra system

	rural	*suburban*	*urban*
# classes	24	4	11
# teachers	7	3	6
White students	97.51%	97.61%	2.69%
Black students	1.79%	0.46%	96.58%
Hispanic students	0.39%	0.28%	0.24%
Asian/Pacific Islander students	0.08%	1.57%	0.49%
American Indian/Alaska Native students	0.23%	0.09%	0%
Reading proficient	44.00%	90.80%	31.20%
Math proficient	25.00%	84.00%	21.60%
Economically disadvantaged	26.72%	4.20%	99.30%
Distance from urban center	32.6 mi.	8.6 mi.	0.8 mi.

Each of the data sets were separated into clips on which our model of gaming the system was applied. The result is a list of clips for each student as well as an indicator of whether a clip contained each of the 13 patterns of gaming the system.

For the Cognitive Tutor Algebra and Cognitive Tutor Middle School data sets, clips were created from sequences of at least 5 actions with a minimum duration of 20 seconds. In cases where the 5 selected actions had a total duration of less than 20 seconds, additional sets of 5 actions were added until the duration of the clip was greater than 20 seconds. This is consistent with how clips were created in the data set that was used to develop the model of gaming the system.

Due to differences in how ASSISTments presents problems, clips in ASSISTments were defined using a different structure. Whereas the Cognitive Tutor platform requires students to solve multiple steps before completing a problem, ASSISTments's problems can be solved in one step when the first attempt is correct. Problems are scaffolded through additional sub-questions when the student answers incorrectly. As such, we defined a clip in ASSISTments as starting from the first action on an original unscaffolded problem to the last attempt before the next original problem. For this reason, clips in ASSISTments can be shorter or longer than clips in the Cognitive Tutors.

Table 3 presents the number of actions, clips and students for each data set as well as the average number of clips per student and average number of actions per clip. For each student, in each data set, we computed the *relative percentage* of time each pattern was observed for this student, shown in Table 3. This measure informs us as to which gaming pattern was most common in each data set.

Table 3. Descriptive statistics for our six data sets.

	Algebra - rural	*Algebra - suburban*	*Algebra - urban*	*Algebra - all*	*Middle School*	*ASSIST-ments*
# actions	474,150	385,628	1,048,294	1,908,072	865,439	681,105
# clips	80,337	61,510	178,662	320,509	126,434	240,450
# students	352	59	165	576	233	1,367
Avg. clips per student	228.23	1042.54	1082.80	556.44	542.64	175.90
Avg. actions per clip	5.90	6.27	5.87	5.95	6.84	2.83

Table 4. Mean (and *SD*) for the relative percentage of time each pattern was observed.

	Algebra - rural	*Algebra - suburban*	*Algebra - urban*	*Algebra - all*	*Middle School*	*ASSIST-ments*
Pattern 1	4.94% (5.87%)	6.48% (3.46%)	4.08% (3.59%)	4.85% (5.13%)	5.33% (4.27%)	4.64% (11.73%)
Pattern 2	25.10% (18.12%)	20.97% (6.94%)	30.98% (21.19%)	26.38% (18.53%)	23.85% (9.90%)	36.17% (18.55%)
Pattern 3	2.17% (3.80%)	2.08% (1.27%)	1.99% (2.10%)	2.11% (3.19%)	2.59% (1.87%)	0.01% (0.17%)
Pattern 4	7.59% (8.48%)	9.74% (4.53%)	5.79% (4.89%)	7.29% (7.34%)	12.98% (6.00%)	5.12% (6.50%)

Pattern 5	18.03% (11.46%)	20.00% (4.85%)	15.38% (10.75%)	17.47% (10.83%)	25.57% (6.47%)	29.38% (15.98%)
Pattern 6	2.73% (3.04%)	2.90% (1.97%)	2.88% (2.47%)	2.79% (2.78%)	1.18% (1.33%)	1.77% (6.51%)
Pattern 7	4.42% (5.66%)	5.22% (2.58%)	5.68% (11.21%)	4.87% (7.52%)	5.97% (6.12%)	0.29% (2.57%)
Pattern 8	1.44% (3.37%)	1.64% (1.47%)	2.57% (9.15%)	1.79% (5.61%)	0.00% (0.00%)	0.00% (0.00%)
Pattern 9	4.57% (5.66%)	4.68% (2.83%)	4.22% (3.41%)	4.48% (4.86%)	8.51% (3.96%)	0.14% (1.28%)
Pattern 10	4.64% (4.48%)	4.83% (2.68%)	4.52% (3.51%)	4.46% (4.06%)	7.90% (3.94%)	5.89% (8.35%)
Pattern 11	5.09% (4.95%)	4.74% (2.49%)	4.52% (3.24%)	4.89% (4.31%)	1.22% (1.88%)	3.30% (7.18%)
Pattern 12	7.06% (7.03%)	5.76% (2.12%)	6.47% (5.50%)	6.75% (6.27%)	3.10% (2.47%)	7.29% (10.40%)
Pattern 13	12.21% (10.27%)	10.96% (3.35%)	10.90% (9.45%)	11.70% (9.55%)	1.82% (2.04%)	5.99% (7.97%)

3 Results

Statistical analyses were conducted to compare the distributions of relative percentages of gaming patterns across populations and learning environments. This allowed us to investigate which gaming patterns were most common in each data set. Due to the non-normal distributions and non-homogeneous variance of our variables, we used Kruskall-Wallis tests, the non-parametric equivalent of ANOVA, to identify main effects across population and environments and Mann-Whitney tests, the non-parametric equivalent of the t-test, to compare pairs of data sets when a main effects were found.

Two different sets of analyses were conducted: 1) relative percentages of patterns across populations in Cognitive Tutor Algebra (Table 5); and 2) relative percentages of patterns across learning environments (Table 6). For each set of analyses, Kruskall-Wallis tests were conducted to identify statistically significant differences across three data sets, followed by Mann-Whitney tests to identify significant differences between pairs of data sets and to compute effect sizes (reported as the rank-biserial correlation *r*). Due to the use of many statistical significance tests, the false discovery rate for each set of analyses was controlled using the Benjamini and Hochberg procedure [30].

Table 5. Diffences in relative percentages of gaming patterns across populations of students. Dashes signify that pairwise comparisons were not conducted when no significant main effect was found. Significant results, after controlling the false discovery rate, are indicated using bold fonts.

	All data	***Rural vs. suburban***		***Rural vs. urban***		***Suburban vs. urban***	
	P	*p*	*r*	*p*	*r*	*p*	*r*
Pattern 1	**< 0.001**	**< 0.001**	**0.202**	0.954	0.003	**< 0.001**	**0.314**
Pattern 2	**< 0.001**	0.455	0.038	**< 0.001**	**0.182**	**< 0.001**	**0.254**
Pattern 3	**0.003**	**0.02**	**0.154**	**0.027**	**0.099**	0.199	0.087
Pattern 4	**< 0.001**	**0.001**	**0.167**	0.074	0.080	**< 0.001**	**0.355**
Pattern 5	**< 0.001**	**0.011**	**0.129**	**< 0.001**	**0.159**	**< 0.001**	**0.355**

Pattern 6	0.156	--	--	--	--	--	--
Pattern 7	**< 0.001**	**< 0.001**	**0.187**	**0.006**	**0.123**	0.085	0.116
Pattern 8	**< 0.001**	**< 0.001**	**0.255**	**< 0.001**	**0.224**	0.206	0.085
Pattern 9	0.087	--	--	--	--	--	--
Pattern 10	0.559	--	--	--	--	--	--
Pattern 11	0.945	--	--	--	--	--	--
Pattern 12	0.564	--	--	--	--	--	--
Pattern 13	0.085	--	--	--	--	--	--

Table 6. Diffences in relative percentages of gaming patterns across learning environments. Significant results, after controlling the false discovery rate are indicated using bold fonts.

	All data	***Algebra vs. Middle School***		***Algebra vs. ASSISTments***		***Middle School vs. ASSISTments***	
	P	*p*	*r*	*p*	*r*	*p*	*r*
Pattern 1	**< 0.001**	**0.010**	**0.092**	**< 0.001**	**0.304**	**< 0.001**	**0.356**
Pattern 2	**< 0.001**	0.494	0.024	**< 0.001**	**0.345**	**< 0.001**	**0.313**
Pattern 3	**< 0.001**	**< 0.001**	**0.209**	**< 0.001**	**0.714**	**< 0.001**	**0.916**
Pattern 4	**< 0.001**	**< 0.001**	**0.408**	**< 0.001**	**0.187**	**< 0.001**	**0.410**
Pattern 5	**< 0.001**	**< 0.001**	**0.500**	**< 0.001**	**0.467**	**< 0.001**	**0.134**
Pattern 6	**< 0.001**	**< 0.001**	**0.239**	**< 0.001**	**0.379**	**< 0.001**	**0.258**
Pattern 7	**< 0.001**	**< 0.001**	**0.165**	**< 0.001**	**0.746**	**< 0.001**	**0.876**
Pattern 8	**< 0.001**	**< 0.001**	**0.473**	**< 0.001**	**0.652**	1.000	0.000
Pattern 9	**< 0.001**	**< 0.001**	**0.461**	**< 0.001**	**0.763**	**< 0.001**	**0.906**
Pattern 10	**< 0.001**	**< 0.001**	**0.383**	0.173	0.032	**< 0.001**	**0.217**
Pattern 11	**< 0.001**	**< 0.001**	**0.426**	**< 0.001**	**0.302**	0.266	0.028
Pattern 12	**< 0.001**	**< 0.001**	**0.331**	**0.002**	**0.073**	**0.008**	**0.067**
Pattern 13	**< 0.001**	**< 0.001**	**0.603**	**< 0.001**	**0.366**	**< 0.001**	**0.101**

4 Discussion

4.1 Differences across populations of students

Table 7 summarizes our comparison of the distributions of relative percentages of each pattern of gaming across 3 populations of students using Cognitive Tutor Algebra. Significant main effects were found in 7 out of 13 gaming patterns with less than half of the pairwise comparison showing significant differences (counting pair-wise tests not run due to a non-significant main effect as themselves non-significant). The suburban population of students seemed to differ most in gaming behavior with the highest number of significant pairwise differences found when comparing rural and suburban students and the largest effect size being found when comparing suburban and urban students. Effect sizes were relatively low when comparing rural and suburban students (average r = 0.182) and rural and urban students (average r = 0.157). The comparison of suburban and urban students resulted in higher size (average *r* = 320 among significant results), but only 4 patterns showed significant differences.

Further inspection of the patterns for which significant differences were found provided us with information about the nature of gaming behaviors across population. The strongest effects were found in relationship to students quickly entering answers. Suburban students more often engaged in gaming that involved quick answers. This was observed for all three patterns that included quick answers (patterns #1, #4 and #5). Urban students also used more quick answers than rural students, but this difference was only significant for pattern #5. Significant differences for pattern #2 revealed that urban students used more long sequences of similar answers when gaming. Finally, significant differences for patterns #7 and #8 suggest that rural students engage less often in gaming that involve reusing the same answer in different parts of the problem.

Table 7. Number of significantly different distributions of relative percentages of gaming pattern across 3 populations of students and average effect size *r* significant.

	Relative percentages	
	Significant differences	Average *r*
Main effect	7 out of 13	--
Rural vs. Suburban	6 out of 13	0.182
Rural vs. Urban	5 out of 13	0.157
Suburban vs. Urban	4 out of 13	0.320

4.2 Differences across learning environments

Table 8 summarizes our comparison of the distributions of the relative percentages of each pattern of gaming across Cognitive Tutor Algebra, Cognitive Tutor Middle School and ASSISTments. Significant main effects were found in all patterns with a large majority of patterns showing significant pairwise comparison. Average effect sizes were also stronger than for populations of students. Overall, Gaming behaviors differed more between ASSISTments and both Cognitive Tutor environments than between Cognitive Tutor Algebra and Cognitive Tutor Middle School.

Significant differences in relative percentages suggested that quick answers were more common in Cognitive Tutor Middle School then in Cognitive Tutor Algebra. Both Cognitive Tutors had more gaming involving quick answers than ASSISTments. This was supported by significant differences in the relative percentages for patterns #1 and #4. However, pattern #5, which also involves quick answers, was more common in ASSISTments. Long sequences of similar answers (pattern #2) were more common in ASSISTments than in other environments and were more common in Cognitive Tutor Algebra than in Cognitive Tutor Middle School. Gaming Patterns which included students reusing the same answer in a different part of the problem tended to occur infrequently in ASSISTments. This is true for 3 out of 4 patterns containing such behaviors (patterns #3, #7 and #8) which had average relative percentages lower than 0.3% (Table 4); only pattern #1 had a higher average relative percentage (4.64%). This is probably because Cognitive Tutors tend to show several problem steps on the screen at once, whereas students complete one problem step at a time in ASSISTments, enabling students to quickly try the same answer in multiple places. Even pattern #1 was seen significantly more frequently in Cognitive Tutors than ASSISTments. Finally, gaming behavior containing help requests were most frequent in Cognitive Tutor Algebra, followed by ASSISTments. Cognitive Tutor Middle School had the least help-related gaming. This was true for 3 out of 4 patterns containing help requests (patterns #6, #11 and #13). However, pattern #11 did not show a statistically significant difference between Cognitive Tutor Algebra and ASSISTments. Pattern #12 was the only help related pattern for which ASSISTments had the highest relative percentage. This may reflect a lower overall rate of gaming for ASSISTments than the Cognitive Tutors, or it may suggest that students using ASSISTments use additional methods for gaming that were not uncovered in the qualitative research (conducted using Cognitive Tutor data) that led to this model.

Table 8. Number of significantly different distributions of relative percentages of gaming patterns across learning envrionments and average effect size *r* across significant differences.

	Relative percentages	
	Significant differences	Average *r*
Main effect	13 out of 13	--
CT Algebra vs. CT Middle School	12 out of 13	0.358
CT Algebra vs. ASSISTments	12 out of 13	0.442
CT Middle School vs. ASSISTments	11 out of 13	0.414

5 Conclusion

Results from our study showed significant differences in the nature of gaming behavior demonstrated by different populations of students and in different environments. However, differences between populations tend to be less frequent and weaker than those between environments. We also observed that, when comparing behaviors across learning environments, the degree of similarity between the environments seems to play a factor in the strength of the differences. Indeed, observed differences between the fairly similar Cognitive Tutor Algebra and Cognitive Tutor Middle School were weaker than the differences between either of those environments and ASSISTments.

The presented study allowed us to identify that student behaviors, more specifically gaming the system behaviors, can vary across different data sets. This information is important to consider when building models of student behaviors, whether for gaming the system or for other constructs. Although it is not surprising that students have different behaviors in different learning environments, the possibility of different behaviors across different schools is not as readily apparent. As such, it is important to keep in mind that models created using only a data set limited only to specific schools and/or regions might have biases based on the population of students it includes.

Those biases do not necessarily imply that models of student behaviors are unusable across different data sets. For example, a previous study [17] showed how gaming models were able to transfer, with some limitation, to new data sets. However, those biases could be a factor in the decreased performance we observed when applying our models to some of the new data sets. Similarly, they might explain why the model created solely based on expert knowledge was more stable across data sets than the hybrid model, created using a combination of machine learning and expert knowledge, despite being less accurate on the training data. The improved accuracy the hybrid models achieved within the systems they were created for may be due to them fitting to details of how gaming specifically occurs in that system, rather than capturing more general aspects of gaming behavior. Being aware of such biases in our models will offer us insight that will be useful as we attempt to improve the generalizability of our models of student behaviors, whether developed using knowledge engineering or machine learning approaches.

Our study presents a first step toward understanding how different factors can bias gaming behaviors in intelligent tutoring systems. We showed differences in the frequencies of gaming behavior across systems and populations. However, we did not investigate in detail which specific factors are most strongly associated to those differences (though there is good reason to believe that some of the difference between systems is related to different aspects of their design, specifically whether multiple problem steps are visible at the same time). One interesting direction that this suggests is that it may be possible to better tailor gaming detection for a new system by taking a model such as this one, and determining which behaviors are less feasible within the new system. These behaviors can then be omitted and the predictive strength of the more feasible behaviors can be increased within the model, under the assumption that students who want to game will find an alternate strategy for doing so [cf. 31].

When studying differences across student populations, our first step focused on comparing rural, suburban and urban schools. This was done for multiple varied reasons. First, a study in the field of affect detection [16] provided us with evidence that differences can be observed, even at such a coarse-grained level. Second, the demographic information presented in Table 2 indicates that there are variations in student populations across the different schools we selected. Finally, historically, fine-grained demographic information has usually not been collected in intelligent tutoring system studies. This lack of historical data significantly increases the difficulty of conducting large scale studies across populations of student. In this context, studying school level differences provided a good starting point for our study. Future study will need to ensure that detailed demographic information is collected. The absence of this information may have led to the result we observed, where although the differences between populations were statistically significant, the correlations were stronger between systems than between populations. It is possible that looking at more fine-grained characteristics of the student populations, such as gender, ethnicity or socio economic status, would provide a better fit for the distribution of gaming behaviors across students. Similarly, when studying differences across systems, future study will need to identify defining characteristics of the different systems and study how they are associated with the observed differences in behavior and will need to pay attention to how demographic differences in the populations of students using those learning environments impacts their behaviors. By doing so, we hope that we identify ways to improve the generalization of models of student behaviors, such as gaming the system, across learning systems and student population to speed their adoption by the broader community of learning engineers.

6 Acknowledgement

We would like to thank PSLC DataShop, Carnegie Learning and the ASSISTments team for providing us with access to student data. We would also like to thank support from NSF #DRL-1535340 and NSF #DRL- 1252297.

References

1. Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R.: Developing a Generalizable Detector of When Students Games the System. *User Modeling and User Adapted Interaction*, 18, 287-314, 2008.
2. Beck, J., Rodrigo, M.M.T.: Understanding Wheel Spinning in the Context of Affective Factors. *Proc. of the 12th Int'l Conference on Intelligent Tutoring Systems*, 162-167, 2014.
3. Cocea, M., Hershkovitz, A., Baker, R.S.J.d.: The Impact of Off-Task and Gaming Behaviors on Learning: Immediate or Aggregate? *Proc. of the 14th Int'l Conference on Artificial Intelligence in Education*, 507-514, 2009.
4. Fancsali, S.E.: Data-Driven Causal Modeling of "Gaming the System" and Off-Task Behavior in Cognitive Tutor Algebra. *NIPS Workshop on Data Driven Education*.
5. Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M.: 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, 2014.
6. Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: Better to be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States During Interactions with Three Different Computer-Based Learning Environments. *Int'l Journal of Human-Computer Studies*, 68, 223-241, 2010.
7. San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T.: Predicting College Enrolment from Student Interaction with an Intelligent Tutoring System in Middle School. *Proc. of the 6th Int'l Conference on Educational Data Mining*, 177-184, 2013.
8. Baker, R.S.J.d., Mitrovic, A., Mathews, M.: Detecting Gaming the System in Constraint-Based Tutors. *Proc. of the 18th Conference on User Modeling, Adaptation and Personalization*, 267-278, 2010.
9. Walonoski, J.A., Heffernan, N.T.: Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. *Proc. of the 8th Int'l Conference on Intelligent Tutoring Systems*, 382-391, 2006.
10. Beal, C.R., Qu, L., Lee, H.: Classifying Learner Engagement Through Integration of Multiple Data Sources. *Proc. of the National Conference on Artificial Intelligence*, 151-156, 2006.
11. Johns, J., Woolf, B.: A Dynamic Mixture Model to Detect Student Motivation and Proficiency. *Proc. of the National Conference on Artificial Intelligence*, 163-168, 2006.
12. Muldner, K., Burleson, W., Van de Sande, B., VanLehn, K.: An Analysis of Students' Gaming Behaviors in an Intelligent Tutoring System: Predictors and Impact. *User Modeling and User Adapted Interaction*, 21, 99-135, 2011.
13. Aleven, V., McLaren, B.M., Roll, I., Koedinger, K.R.: Towards Meta-Cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor. *Int'l Journal of Artificial Intelligence in Education*, 16, 101-130, 2006.

14. Gong, Y., Beck, J., Heffernan, N.T., Forbes-Summer, E.: The Fine-Grained Impact of Gaming (?) on Learning. *Proc. of the 10th Int'l Conference on Intelligent Tutoring Systems*, 194-203, 2010.
15. Baker, R.S.J.d., Gowda, S.M.: An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools. *Proc. of the 3rd Int'l Conference on Educational Data Mining*, 11-20, 2010.
16. Ocumpaugh, J., Baker, R.S., Gowda, S., Heffernan, N., Heffernan, C.: Population Validity for Educational Data Mining Models: A Case Study in Affect Detection. *British Journal of Educational Technology*, 45(3), 487-501.
17. Paquette, L., Baker, R. S., de Carvalho, A.M.J.A., Ocumpaugh, J.: Cross-System Transfer of Machine Learned and Knowledge Engineered Models of Gaming the System. *Proc. Of the 23rd Conference on User Modeling, Adaptation and Personalization*, 183-194, 2015.
18. Koedinger, K.R, Corbett, A.T.: Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R.K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences*, 61-77, 2006.
19. Razzaq, L., et al.: The Assistment Project: Blending Assessment and Assisting. *Proc. of the 12 Annual Conference on Artificial Intelligence in Education*, 555-562, 2005.
20. Paquette, L., de Carvalho, A.M.J.A, Baker, R. S.: Towards Understanding Export Coding of Student Disengagement in Online Learning. *Proc. of the 36th Annual Cognitive Science Conference*, 1126-1131, 2014.
21. Clark, R.E., Feldon, D., van Merriënboer, J., Yates, K., Early, S.: Cognitive Task Analysis. In J.M. Spector, M.D., Merrill, J.J.G. van Merriënboer, & M.P. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology (3rd ed.)*, 575-593, 2008.
22. Cooke, N.J.: Varieties of Knowledge Elicitation Techniques. *Int'l Journal of Human-Computer Studies*, 41, 801-849, 1994.
23. Meyer, M.A.: How to Apply the Anthropological Technique of Participant Observation to Knowledge Acquisition for Expert Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 22, 983-991, 1992.
24. Van Someren, M.W., Barnard, Y.F., Sandberg, J.A.C.: *The Think Aloud Method: A Practical Guide to Modeling Cognitive Processes*, 1994.
25. Levenshtein, A.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10 (8), 707-710, 1966.
26. Cohen, J.: A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20 (1), 37-46, 1960.
27. Baker, R.S.J.d., de Carvalho, A.M.J.A.: Labeling Student Behavior Faster and More Precisely with Text Replays. *Proc. of the 1st Int'l Conference on Educational Data Mining 2008*, 38-47, 2008.
28. Paquette, L., de Carvalho, A.M.J.A, Baker, R.S., Ocumpaugh, J.: Reengineering the Feature Distillation Process: A Case Study in the Detection of Gaming the System. *Proc. of the 7th Int'l Conference on Educational Data Mining*, 284-287, 2014.
29. Koedinger, K. R., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: *A Data Repository for the Community: The PLSC DataShop*. CRC Press, Boca Raton, 2010.
30. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society B., 57, 289-300, 2003.
31. Murray, C., VanLehn, K. Effects of Dissuading Unnecessary Help Requests While Providing Proactive Help. *Proceedings of the International Conference on Artificial Intelligence and Education*, 887-889, 2005.