An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools

Ryan S.J.d. Baker, Sujith M. Gowda
rsbaker@wpi.edu, sujithmg@wpi.edu
Department of Social Science and Policy Studies, Worcester Polytechnic Institute

Abstract. We study how student behaviors associated with disengagement differ between different school settings. Towards this, we investigate the variation in the frequency of off-task behavior, gaming the system, and carelessness in an urban school, a rural school, and a suburban school in the United States of America. This analysis is conducted by applying automated detectors of these behaviors to data from students using the same Cognitive Tutor educational software for high school Geometry, across an entire school year. We find that students in the urban school go off-task and are careless significantly more than students in the rural and suburban schools. Differences between schools in terms of gaming the system are less stable. These findings suggest that some of the differences in achievement by school type may stem from differences in engagement and problem behaviors.

1 Introduction

In recent years, intelligent tutoring systems have left the research laboratory, expanded beyond the research classroom, and have started to see large-scale use worldwide. To give two examples, Cognitive Tutors for high school Algebra and Geometry [cf. 14, 15] are now used by hundreds of thousands of students each year in the United States, and constraint-based tutors for SQL and database normalization are now used by tens of thousands of students each year worldwide [20]. Aplusix, another system that has seen particularly wide use, is now in use in at least 6 countries [cf. 21]. As the reach of intelligent tutors increases, they become available to an increasingly diverse population of students. Tutors are now used by students of very different economic and ethnical backgrounds, by students in suburbs, cities, and rural areas, by wealthy, middle-class, and poor students, by students from ethnic majority groups and students from minority groups [cf. 8, 14, 15, 23].

As the use of intelligent tutors spreads to a more diverse selection of populations, we gain the potential to use intelligent tutors as a research tool in yet another domain of educational research – comparisons of student learning and behavior between learners in schools with different demographic profiles. There is increasing evidence that students in different learning settings have radically different learning outcomes [cf. 11], but there is insufficient understanding about what factors mediate those learning outcomes. There are many differences between schools in different settings, including differences in teacher expertise [17], in the physical conditions of schools, and in students' backgrounds.

An additional possibility is that the differences in learning in schools may be influenced, at least in part, by disengagement. Given the known relationships between disengaged behaviors and learning (see [13] for an early review of this literature in traditional classrooms; see [2, 6, 10, 12, 22] for studies on this topic in classrooms using educational software), there is valid reason to think that disengaged behavior may mediate the differences in learning between different school settings. However, it has not yet been established whether there are significant differences in disengaged behavior between types of schools.

In this paper, we focus on this topic, studying how three types of student behavior associated with disengagement differ across different school settings. In specific, we investigate the variation in gaming the system, off-task behavior, and carelessness between urban, rural, and suburban classrooms in the United States of America, using "discovery with models" methods on data from students using the same Cognitive Tutor across an entire school year. Although there have been many studies on off-task behavior within rural, urban, and suburban settings ([13] offers an early review of this literature), we are not aware of any systematic comparisons of off-task behavior *between* types of schools. Similarly, we are not aware of any studies on gaming the system that span urban, rural, and suburban schools, and are not aware of systematic attempts to measure carelessness in any type of school prior to the model advanced in [3].

Studying these issues in the context of Cognitive Tutors several advantages. First, Cognitive Tutors have been deployed to a wide variety of settings, and learning gains have been demonstrated in urban, rural, and suburban schools [7, 8, 15, 23]. Second, Cognitive Tutors collect extensive logs of every student answer or help request within the software [16]. For Cognitive Tutors, sufficient log data to study these research questions is currently available in the Pittsburgh Science of Learning Center (PSLC) DataShop [16], in a format compatible with existing detectors of disengaged behaviors [2, 3, 4, 5].

In this paper, we utilize data from the PSLC DataShop in order to compare USA schools with 3 very different profiles (urban, suburban, and rural), according to 3 metrics (percent of time off-task, percent of time spent gaming the system, and average slip probability). We do so in the context of an entire year of use of the exact same Cognitive Tutor for Geometry.

2 Methods

Data from 3 schools was obtained through the PSLC DataShop [16] in order to compare students' behavior within urban, rural, and suburban settings, across an entire school year. While a sample of 3 schools is clearly not enough data to be able to draw conclusions about all schools from these categories, it enables us to statistically compare the behavioral profile of 3 schools that are representative of each of these categories. This study also provides a methodological template for future expanded investigations of these issues, which can be conducted as data becomes available from a broader range of schools.

The three schools studied are each public high schools in Southwestern Pennsylvania. As shown in Table 1, each school has a distinctive demographic profile; the urban school is overwhelmingly African-American, whereas the suburban and rural school are

Table 1. School Population Demographics

	Urban school	**Suburban school**	**Rural school**
% African-American	100%	<1%	2%
% White	0%	98%	97%
% Hispanic	0%	<1%	<1%
Free or Reduced-Price Lunch	99%	4%	Not Reported
% Proficient on state math exam	20%	77%	Not Reported
Median household income in community	$26,621	$60,307	$32,206
% Children under poverty line	30.8%	2.5%	18.4%

overwhelmingly White Non-Hispanic (the ethnic majority group in the region). There is evidence of considerable poverty in the urban school's population, with a low median household income for the United States, and a high percentage of children under the poverty line. Similarly, the rural school has a low median household income for the United States, and a substantial percentage of children under the poverty line. The suburb has considerably less poverty, with a median household income over double the urban school's median household income, and only 2.5% of children under the poverty line.

Data from each school was obtained through the PSLC DataShop [16]. In each case, high school students took their Geometry courses using the same Cognitive Tutor for Geometry [7, 14, 15], and all data was collected in the PSLC DataShop. Data was collected for the entire school year, from August 2005 to May 2006. 434 students in the rural school used the software, 88 students in the suburban school used the software, and 34 students in the urban school used the software. The three schools each assigned the software to students who were neither in special needs nor gifted classes – the difference in the number of students using the software is solely based on school size, and how many teachers chose to use the software (in particular, the rural school is a large regional school, as is increasingly common in the USA in rural areas, whereas the urban and suburban schools serve smaller populations).

In all schools, the software was used by groups of students in a computer laboratory, working individually at separate computers, at their own pace. Students in the rural school used the software an average of 9 hours, students in the suburban school used the software an average of 35 hours, and students in the urban school used the software an average of 51 hours. Hence it appears that the teachers in each school chose to have their students use the software in different amounts. This difference represents a selection bias in our data, but it is a difficult confound to resolve; for instance, restricting analysis to students who used the software above a time cutoff introduces a different selection bias. In particular, the difference in usage is a natural one, reflecting genuine implementation in each type of school.

To address this selection bias stemming from teacher choice, we analyze the data in two ways – using all data (the more ecologically valid choice), and using a time-slice consisting of the 3rd-8th hours (minutes 120-480) of each student's usage (this time-slice will not be as representative of the usage in each school, but avoids this confound). The 3rd-8th hours were selected, because the initial 2 hours likely represent interface learning, and therefore may not be representative of overall tutor use (and interface learning is likely to be dependent on prior experience with educational software, which is likely to be greater for wealthier students). In general, implementational differences between schools are likely to exist in year-long comparisons. In the long-term, this problem can probably be best addressed by conducting analyses of this nature across large numbers of schools, in order to average across implementational differences orthogonal to the type of school (though some implementational differences may be characteristic to certain types of schools – for instance, urban schools might use specific pieces of educational software more heavily due to having lower resources to provide a wide range of educational software in their classrooms).

Each action in each data set was labeled using detectors of gaming the system, off-task behavior, and carelessness. The gaming detector used was trained using data from students using a Cognitive Tutor for Algebra [5], using an age-similar population and an approach validated to generalize between students and between Cognitive Tutor lessons [4]. The off-task detector used was trained using data from students using a Cognitive Tutor for Middle School Mathematics. The off-task detector was validated to generalize to new students, and to function accurately in several Cognitive Tutor lessons [2]. Although the age range was moderately older in this study than in the original training data, off-task behavior is similar in nature within these populations – it involves ceasing to use the software for a significant period of time without seeking help (which can be detected in the log files by the behavior occurring before and after an idle pause). Carelessness was detected using the slip detector from [3], which was trained on data from Cognitive Tutor Geometry. This use of contextual slip is in line with theoretical work by Clements [9], who argues that making errors despite knowing the skills needed for successful performance should be considered evidence of carelessness. It is important, however, to note that contextual slip could potentially also be an indicator of shallow knowledge that does not apply to all items in the tutor, even if they are labeled as involving the same skill.

3 Results

In discussing results, we will first discuss our analyses conducted across the full year of tutor data, and then discuss the same analyses conducted across only a time-slice including the 3rd to 8th hours of tutor usage.

3.1 Analyses Across Full Data Set

Across the full data set, representing data collected during the entire school year, the pattern of off-task behavior was highly different between the three schools. Students in the suburban school were off-task an average of 15.4% of the time (past research in traditional classrooms has averaged 15-20% of time off-task [cf. 18, 19]). Students in the

Table 2. Average incidence of each indicator per school. Parentheses give standard deviation.

	Urban school	Suburban school	Rural school
% Off-Task	34.1% (18.0%)	15.4% (20.7%)	20.4% (13.3%)
% Gaming the System	7.4% (2.2%)	6.9% (3.1%)	6.6% (1.7%)
% Slip Probability	0.50 (0.07)	0.32 (0.11)	0.27 (0.13)

rural school were off-task an average of 20.4% of the time. Students in the urban school were off-task an average of 34.1% of the time. Hence, students at the urban school were off-task 67% more than students at the rural school (a 1.0 SD difference), and over double as much as students at the suburban school (a 0.9 SD difference). The overall difference in off-task behavior between schools was statistically significant between schools, F(2,553)= 18.80, p<0.01. The model predicting time off-task by school predicted 6.4% of the variance in time off-task. The pairwise differences between schools were all statistically significant, using Tukey's HSD to control for multiple comparisons.

The frequency of gaming the system had smaller differences between the three schools, although there were still significant differences. Students in the suburban school gamed 6.9% of the time, students in the rural school gamed 6.6% of the time, and students in the urban school gamed 7.4% of the time. In other words, students in the urban school gamed only 13% more than students in the rural school (a 0.47 SD difference), and 9% more than students in the suburban school (a 0.16 SD difference). The overall difference in gaming the system between schools was statistically significant, F(2,553)= 3.12, p=0.05. The model predicting time spent gaming the system by school predicted 1.1% of the variance in gaming, considerably less than is predicted by individual differences between students or by the differences between tutor lessons [1]. According to Tukey's HSD, the rural school had significantly less gaming than the urban school, but the other differences in gaming were not statistically significant.

The pattern of carelessness was highly different between the three schools. Students in the suburban school had a probability of 0.32 of slipping despite knowing a skill, students in the rural school had a probability of 0.27 of slipping despite knowing a skill, and students in the urban school had a probability of 0.50 of slipping despite knowing a skill. The overall difference in slipping between schools was statistically significant, F(2,553)= 54.50, p<0.001. The model predicting slip probability by school predicted 16.5% of the variance in slip probability. The pairwise differences between schools were all statistically significant, using Tukey's HSD to control for multiple comparisons.

3.2 Analyses Across Data From Hours 3-8

Within the restricted time-slice of data from hours 3-8, the differences in the frequency of off-task behavior were qualitatively similar to the analysis across the full data set, although the difference between the urban school and the other schools was smaller.

Table 3. Average incidence of each indicator per school. Parentheses give standard deviation.

	Urban school	**Suburban school**	**Rural school**
% Off-Task	25.7% (22.8%)	16.5% (27.5%)	21.0% (16.5%)
% Gaming the System	4.7% (1.9%)	5.9% (7.3%)	6.4% (2.2%)
% Slip Probability	0.53 (0.08)	0.44 (0.17)	0.33 (0.18)

Students in the suburban school were off-task an average of 16.5% of the time, very similar to the 15.4% reported across all data. Students in the rural school were off-task an average of 21.0% of the time, very similar to the 20.4% reported across all data. However, students in the urban school were off-task an average of 25.8% of the time, substantially lower than the 34.1% reported across all data, a statistically significant difference, t(33)=-2.55, p=0.02, for a two-tailed paired t-test. This result suggests that off-task behavior increased during the year in the urban school.

Nonetheless, even during this earlier time-slice, off-task behavior was higher in the urban school than the suburban school. The overall difference in off-task behavior between schools was statistically significant, F(2,484)= 3.01, p=0.05. The model predicting time off-task by school predicted 1.2% of the variance in time off-task. According to Tukey's HSD, the urban school had significantly more off-task behavior than the suburban school, but the rural school was not significantly different from either of the other two schools.

The pattern of gaming the system was highly different within the restricted time-slice of data from hours 3-8, as compared to the entire data set: Gaming the system was much rarer in the urban school. Students in the urban school gamed 4.7% of the time in the restricted time-slice, compared to 7.4% of the time in the full data set, a significant difference, t(33)=8.14, p<0.001, for a two-tailed paired t-test. Gaming was also less common in this time-slice in the other two schools, but to a much lower degree. Students in the suburban school gamed 5.9% of the time, compared to 6.9% of the time in the full data set, which was not quite statistically significant, t(71)=1.51, p=0.13, for a two-tailed paired t-test. Students in the rural school gamed 6.4% of the time, compared to 6.6% of the time in the full data set.

The overall difference in gaming the system between schools was statistically significant, F(2,484)= 4.09, p=0.02. The model predicting time spent gaming the system by school predicted 1.7% of the variance in gaming, considerably less than is predicted by individual differences between students or by the differences between tutor lessons [1]. According to Tukey's HSD, the rural school had significantly more gaming than the urban school – the exact opposite of the result across the entire data set – but the other differences in gaming were not statistically significant.

The pattern of carelessness retained the same ordering within the restricted time-slice of data from hours 3-8, and the entire data set, but the degree of carelessness was significantly higher for all three groups of students, t(71)=6.32, p<0.001 in the suburban school, t(374)=10.08, p<0.001 in the rural school, and t(33)=2.04, p=0.05 in the urban school. In all cases, a two-tailed paired t-test was used.

The overall difference in slipping between schools was statistically significant, F(2,478)=29.78, p<0.001. The model predicting slip probability by school predicted 11.1% of the variance in slip probability. The pairwise differences between schools were again all statistically significant, using Tukey's HSD to control for multiple comparisons.

4 Discussion and Conclusions

In this paper, we have analyzed the prevalence of three student behaviors associated with disengagement in urban, suburban, and rural classrooms in the USA: off-task behavior, gaming the system, and carelessness. These students used the exact same learning software for high school Geometry across the same school year. However, the students used the software for different amounts of time in each school, a common phenomena in real-world use of educational software, where usage decisions are made by teachers and school administrators, rather than researchers and curriculum developers. To address this difference, we compared between these schools in two fashions. First, we compared all data to get a fully ecologically valid comparison. Second, we compared within a time-slice consisting of each student's 3rd to the 8th hours of usage in each school, in order to control for confounds stemming from differences in implementation between schools.

The two versions of the analysis agreed that the urban school had more off-task behavior and carelessness than the suburban school and rural school. In terms of these behaviors, students in the rural and suburban schools were more similar to each other than either school was to the urban school. One interesting note is that carelessness dropped significantly more over the course of the school year in the suburban and rural schools than in the urban school, suggesting that some influence or factor caused the suburban and rural students to become more diligent during the school year, but that this influence or factor was significantly less relevant in the urban school.

As both the rural school and the urban school had significant poverty, it appears that some aspect of these schools other than simply socio-economic status explains the higher frequency of off-task behavior and carelessness in the urban school. There are several potential hypotheses what other aspects may explain these behavioral differences, including differences in teacher expertise (which is often lower in urban schools [17]), differences in schools' facilities, equipment (e.g. computers), and physical environment, and differences in students' cultural backgrounds. Determining whether one of these factors explains the differences in off-task behavior and carelessness will be an important topic for future research.

A contradictory finding between analyses was found for gaming the system. Across the whole data set and entire school year, gaming the system was most frequent in the urban school. However, within the 3rd to 8th hours of tutor usage, gaming the system was least frequent in the urban school. This suggests that students in the urban school gamed more, later in the year. This may just be an artifact of the lessons encountered, as tutor lesson predicts a substantial portion of the variance in gaming behavior in Cognitive Tutors [1]. However, it may also be that the novelty of Cognitive Tutors reduces gaming initially in American students. There is evidence for this possibility in the finding that gaming was lower in all schools during the 3rd-8th hours. The difference in gaming behavior between the early time-slice and the overall data set was more pronounced in the urban school, but this may be due to lower familiarity with educational technology in general, a finding worth investigating further.

In future years, we plan to replicate these analyses with a larger number of schools in each of these settings, using the research presented here as a methodological template for that later research. Automated machine-learned detectors provide an essential tool for analysis of this sort, in this author's opinion a better tool than existing alternatives. For example, it is not tractable to use observational, text replay annotation, or video methods at this sort of scale. [5] presents a use of text replay methods to analyze a single behavior among 58 students over an entire school year; though text replay methods are significantly faster than live observation or video coding methods, the coding needed for this analysis took over 200 hours. Utilizing text replays to annotate the 3 school sample used in this paper would have taken over 2000 hours, assuming a rate of observation equal to that in [5]. Video coding and field observation would have taken even longer.

That said, it is worth noting that automated detectors have important challenges not present when using human labels. It is important to validate the generalizability of detectors across students, schools, and learning materials, a task which has been only partially completed for the detectors used in this paper, and which has received insufficient attention in the literature in general. Construct validity is also a key issue in the use of machine-learned detectors, and is more a risk in detectors that are based on theoretically determined training labels (e.g. the model of carelessness), compared to detectors based on human judgments shown to have good inter-rater reliability (e.g. the detectors of off-task behavior and gaming the system). It is worth noting that automated detectors produced with a common alternative to machine learning, knowledge engineering, are likely to be prone to the same challenges to generalizability and construct validity as machine-learned detectors. Current practice with knowledge engineering often does not check detectors against human labels or across contexts, a potentially significant risk to using these models in discovery with models analyses.

As research applying detectors across contexts goes forward, it has significant potential to support progress in studying the impact of school context. By further study of which school contexts – and what attributes of those contexts – are associated with greater frequencies of disengaged behavior, we may be able to better understand the differences in learning between different learning settings. This may in turn support education researchers and practitioners in designing curricula, learning software, and interventions tailored to different schools – a potentially key step towards developing educational software that is equally effective for all students, whether they are in urban schools, rural schools, suburban schools, or elsewhere.

Acknowledgements

This research was supported by NSF grant REC-043779 to “IERI: Learning-Oriented Dialogs in Cognitive Tutors: Toward a Scalable Solution to Performance Orientation”, and by the Pittsburgh Science of Learning Center (National Science Foundation) via grant “Toward a Decade of PSLC Research”, award number SBE-0836012. We would also like to thank Adriana de Carvalho and Ma. Mercedes T. Rodrigo for valuable comments and suggestions, and the members of the Pittsburgh Science of Learning Center DataShop, Alida Skogsholm in particular, for their support and assistance.

References

[1] Baker, R.S.J.d.: Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model. *Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007*, 76--80.

[2] Baker, R.S.J.d. Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of ACM Computer-Human Interaction*, 2007, 1059--1068

[3] Baker, R.S.J.d., Corbett, A.T., Aleven, V. More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 2008, 406-415.

[4] Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R.. Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction, 18* (3), 2008, 287--314.

[5] Baker, R.S.J.d., de Carvalho, A. M. J. A.: Labeling Student Behavior Faster and More Precisely with Text Replays. *Proceedings of the 1st International Conference on Educational Data Mining*, 2008, 38-47.

[6] Beck, J. Engagement tracing: using response times to model student disengagement. *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005)*, 88-95.

[7] Butcher, K., & Aleven, V. Integrating visual and verbal knowledge during classroom learning with computer tutors. In D.S. McNamara & J.G. Trafton (Eds.), Proceedings of the 29th Annual Cognitive Science Society, 2007, 137-142.

[8] Carnegie Learning, Inc. *Results from The Colony, TX* (Cognitive Tutor Research Report TX-00-02), 2001. Pittsburgh, PA: Carnegie Learning, Inc.

[9] Clements, M.A. Careless Errors Made by Sixth-Grade Children on Written Mathematical Tasks. *Journal for Research in Mathematics Education, 13* (2), 1982, 136-144.

[10] Cocea, M., Hershkovitz, A., Baker, R.S.J.d.: The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 2009, 507--514

[11] Fan, X., Chen, M. J.: Academic achievement of rural school students: A multi-year comparison with their peers in suburban and urban schools. *Journal of Research in Rural Education*, 15, 1999, 31--46.

[12] Gobel, P.: Student off-task behavior and motivation in the CALL classroom. *International Journal of Pedagogies and Learning, 4* (4), 2008, 4-18.

[13] Karweit, N., Slavin, R.E.: Time-On-Task: Issues of Timing, Sampling, and Definition. *Journal of Experimental Psychology*, 74 (6), 1982, 844--851.

[14] Koedinger, K. R., Anderson, J.R., Hadley, W., Mark, M. Intelligent tutoring goes to school in the big city. *Proceedings of the 7th International Conference on Artificial Intelligence and Education*, 1995, 421--428.

[15] Koedinger, K. R., Corbett, A. T. Cognitive tutors: Technology bringing learning sciences to the classroom. R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences.* New York, NY: Cambridge University Press, 2006.

[16] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. A Data Repository for the EDM commuity: The PSLC DataShop. To appear in Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining*, in press. Boca Raton, FL: CRC Press.

[17] Lankford, H., S. Loeb, J. Wyckoff. Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis, 24* (1), 2002, 37—62.

[18] Lee, S.W., Kelly, K.E., Nyre, J.E. Preliminary Report on the Relation of Students' On-Task Behavior with Completion of School Work. *Psychological Reports*, *84*, 1999, 267--272.

[19] Lloyd, J.W., Loper, A.B.: Measurement and Evaluation of Task-Related Learning Behavior: Attention to Task and Metacognition. *School Psychology Review, 15* (3), 1986, 336—345.

[20] Mitrovic, A. An Intelligent SQL Tutor on the Web. *International Journal of Artificial Intelligence in Education, 13* (2), 2003, 173--197.

[21] Nicaud J.F., Bittar M., Chaachoua H., Inamdar P., Maffei L. Experiments With Aplusix In Four Countries. *International Journal for Technology in Mathematics Education, 13* (1), 2006.

[22] Rowe, J., McQuiggan, S., Robison, J., Lester, J.. Off-Task Behavior in Narrative-Centered Learning Environments. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Education*, 2009, 99-106.

[23] Sarkis, H. *Cognitive Tutor Algebra 1: Program Evaluation: Miami-Dade County Public Schools*, 2004. Lighthouse Point, FL: The Reliability Group.