Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Implementation Factors and Outcomes for Intelligent Tutoring Systems: A Case Study of Time and Efficiency with Cognitive Tutor Algebra

Stephen E. Fancsali[a], Steven Ritter[a], Michael Yudelson[b], Michael Sandbothe[a], Susan R. Berman[a]

[a] Carnegie Learning, Inc.
437 Grant Street, Suite 1906
Pittsburgh, PA 15219 USA
{sfancsali, sritter, msandbothe, sberman}@carnegielearning.com

[b] Carnegie Mellon University
Human-Computer Interaction Institute
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
myudelson@gmail.com

Abstract

While many expect that the use of advanced learning technologies like intelligent tutoring systems (ITSs) will substitute for human teaching and thus reduce the influence of teachers on student outcomes, studies consistently show that outcomes vary substantially across teachers and schools (Pane et al. 2010; Pane et al. 2014; Ritter et al. 2007a; Koedinger et al. 1997; Koedinger and Sueker 2014). Despite these findings, there have been few efforts (e.g., Schofield 1995) to understand the mechanisms by which teacher practices influence student learning on such systems. We present analyses of Carnegie Learning's Cognitive Tutor ITS data from a large school district in the southeastern United States, which present a variety of usage and implementation profiles that illuminate disparities in deployments in practical, day-to-day educational settings. We focus on differential effectiveness of teachers' implementations and how implementations may drive learner efficiency in ITS usage, affecting long term learning outcomes. These results are consistent with previous studies of predictors and causes of learning outcomes for students using Cognitive Tutor. We provide recommendations for practitioners seeking to deploy intelligent learning technologies in real world settings.

Introduction

Why are teachers important? Intelligent learning technologies like intelligent tutoring systems (ITSs) offer the promise of solving Bloom's two sigma problem (Bloom 1984) by providing instruction as effective as one-on-one tutoring. However, real-world ITSs like Carnegie Learning's Cognitive Tutor (CT) (Ritter et al. 2007) are not intended to replace qualified, engaged instructors or teachers in a classroom. Indeed, Bloom's task in his famous two sigma paper was to find methods of group instruction to achieve substantial improvement in learning outcomes, of which ITSs like CT are but one.

Nevertheless, data-driven research on ITSs and other learning technologies tends to focus on student- or learner-level usage patterns and factors as predictors and possible causes of learning outcomes. While quality of implementation is noted as an important factor that drives outcomes (Pane et al. 2010; Pane et al. 2014; Ritter et al. 2007a; Koedinger et al. 1997; Koedinger and Sueker 2014), there are few studies, with notable exceptions, that delve into specifics about how, precisely, teacher and administrative behaviors affect outcomes of students using learning technologies like ITSs. One notable, early exception considered how teacher and student behavior changed after the introduction of an ITS for geometry proofs (Schofield 1995). This study noted how teachers' roles became more of those of collaborators and helpers to students working through geometry content; further, more individual attention could be provided to students who were struggling most because of automated tutoring affordances of the ITS. Other work explores implementation fidelity and teacher "buy in" of learning technologies within the context of a random field trial of the ASSISTments system as a form of homework support in mathematics (Feng et al. 2014). Another recent study considers aspects of instructor implementation fidelity with the Reasoning Mind blended learning system for elementary and middle school mathematics, taking a data-driven approach to detecting, from log data, whether teachers are providing proactive remediation (Miller et al. 2015).

A recent study of CT also takes a data driven approach to explore elements of implementation fidelity by considering the extent to which teachers implement mastery learning within the ITS. Some teachers manually move students within the system so that they "stay with the class," rather than allowing students to move at a pace determined by mastery (Ritter et al. 2016). Students in classes with teachers who allowed mastery learning to run its course tended to have less variability in error rates, reflecting the fact that they were progressing at an appropriate pace with problems geared to their current level of knowledge as they moved through the CT curricula. In contrast, students who were moved forward in violation of mastery learning made progressively more errors, reflecting increased gaps in their prerequisite knowledge. The willingness or ability to implement pedagogical practices intended within the system, like mastery learning, is thus one way that teachers influence student outcomes when using an ITS. However, implementations differ from class-to-class and school-to-school for a bevy of reasons beyond whether teachers implement mastery learning. In this paper, we present a case study of another important way in which teachers matter, particularly teachers' differential engagement with students interacting with ITSs in their classrooms and computer labs. This engagement affects student time and efficiency using learning technologies, and, consequently, student learning outcomes.

Although many reasonably consider increased "time-on -task" as connected to improved outcomes, and there are many ways to conceptualize and measure time-on-task (see Kovanović et al. 2015 for a thorough review of the literature), researchers examining *school time* provide a nuanced (if perhaps imperfect) view. They define three categories of time in learning settings (Aronson, Zimmerman, and Carlos 1998): *instructional time*, *engaged time,* and *academic learning time*. School policy fundamentally determines instructional time, the amount of time scheduled for particular classes (e.g., the length of class periods in computer labs in which students can use an ITS like CT). Teachers also have influence on instructional time, since they may decide, for example, whether classes will make use of computer lab time scheduled for their math classes.

Research shows that providing students more instructional time may improve outcomes. Double periods of algebra instruction (i.e., doubling instructional time in algebra) were found to provide substantial short-term and long-term improvements for under-performing students in an analysis of data from Chicago Public Schools (Cortes, Goodman, and Nomi 2013). Other studies find short-term benefits of increased instructional time in math but diminishing long-term effects, while noting costs to such increases in math instructional time, namely that increases to instructional time in math "crowds out" important time for other subjects (Taylor 2014).

Teacher practices and, to a lesser extent, school policy determine engaged time, which refers to instructional time that is devoted to learning, as opposed to time spent on non-instructional tasks like student discipline issues and taking attendance, among others. The amount of class time lost to non-instructional tasks may be surprisingly high, with one estimate (Karweit 1985) that only 38% of time in school actually being engaged time.

The portion of engaged time in which students are doing work that is appropriate to their academic goals and knowledge level is called academic learning time. We posit that a fundamental role for teachers is to ensure that engaged time is mostly academic learning time as much as is possible, especially in situations in which necessary instructional resources like computer labs are not always available, for example. Of course, the goal is not to equate engaged time as academic learning time so much so that other important social, non-cognitive, and meta-cognitive learning factors are not nurtured in engaged time (e.g., allowing for students to reasonably and occasionally collaborate or help one another as they interact with the ITS in a computer lab). One way to think about the goal of adaptive learning systems is that they try to decrease the difference between student engaged time and academic learning time by presenting material that is appropriate to students' knowledge levels. More generally, teaching practices are fundamental determinants of the amount of engaged time that is academic learning time, and the proportion of engaged time that is academic learning time is one measure of classroom learning efficiency.

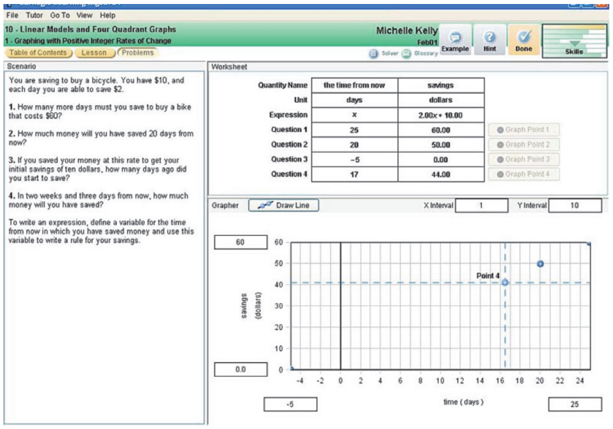
Since school policy is likely to largely determine instructional time and our present focus is the role of teachers, we shift attention to engaged time and academic learning time, upon which teachers have a substantial influence. Recent research demonstrates that student efficiency (Ritter et al. 2013) and content mastery (Sales and Pane 2015) in the classroom environment (i.e., a combination of sufficient academic learning time and of classroom learning efficiency) is one of the best predictors of longer-term learning outcomes.

In the present study, we explore data from use of CT at a large school district in the southeastern United States. We find substantial variation in both academic learning time and learning efficiency between schools, and between classes within a school. We explore associations between student time using the CT, efficiency of CT use, and learning outcomes on a state-level standardized test for algebra. Before providing these analyses, we introduce the CT for Algebra as well as prior work exploring measures of student efficiency and progress and their correlations with standardized test learning outcomes.

Cognitive Tutor Algebra & Curriculum

Carnegie Learning's Algebra curriculum provides a blended approach to learning, combining the adaptive CT ITS with consumable texts that support student-centered instruction in the classroom. Carnegie Learning recommends a 60%-40% split between time in classrooms with teachers employing progressive instructional techniques and time using CT. Since most CT use is in the classroom or a computer lab, the teacher is typically present; students use CT for most instruction but call the teacher over when they have questions that the software cannot answer. Recommended usage of CT amounts to about an hour and a half of time per school week (i.e., ideally, approximately fifty hours of CT interaction in an academic year).

CT focuses on mathematics problem solving. Students learn at their own pace in CT, receiving scaffolding and instruction customized to their own solution strategies. Using the probabilistic framework of Bayesian Knowledge Tracing (BKT) (Corbett and Anderson 1995), CT continually assesses student knowledge of fine-grained skills or knowledge components that are a part of CT's underlying cognitive model. Using BKT's assessment of student skill mastery, CT provides each student with activities that emphasize the skills that he or she needs to learn. Topical sections, each associated with sets of skills, are completed when students demonstrate mastery of relevant skills by solving problems without requiring hints or committing errors. CT breaks problems down into steps to which fine-grained skills are mapped while collecting correspondingly fine-grained usage data about how students are interacting with the ITS (Figure 1).



You are saving to buy a bicycle. You have $10, and each day you are able to save $2.

1. How many more days must you save to buy a bike that costs $60?

2. How much money will you have saved 20 days from now?

3. If you saved your money at this rate to get your initial savings of ten dollars, how many days ago did you start to save?

4. In two weeks and three days from now, how much money will you have saved?

To write an expression, define a variable for the time from now in which you have saved money and use this variable to write a rule for your savings.

Quantity Name	the time from now	savings
Unit	days	dollars
Expression	x	2.00x + 10.00
Question 1	25	60.00
Question 2	20	50.00
Question 3	-5	0.00
Question 4	17	44.00

Figure 1. Screenshot of mathematics problem solving in Cognitive Tutor Algebra. The student is presented a multi-step word problem and a table into which values are entered as responses to questions as well as a graphing tool to plot points and graph a line.

Prior Work: Time & Efficiency in CT

Prior studies of CT Algebra data have focused on facets of student time interacting with CT as well as efficiency with which students manage to make progress through CT content. These studies show that "clock" time using CT (roughly engaged time) is only weakly associated with improved outcomes as indicated by various standardized test scores (Ritter et al. 2013; Joshi et al. 2014; Sales and Pane 2015). These findings make intuitive sense; simply logging on to the software is not going to help students learn.

Instead, student mastery of topics presented in the ITS and the efficiency with which such mastery is achieved are more strongly associated with success. In analysis of experimental data from a large-scale effectiveness trial (Pane et al. 2014) of the CT, models that classified students according to time using CT did not predict as substantial gains on long-term learning outcomes relative to the control group as did models based on completion of content (Sales and Pane 2015). This suggests students who were placed in environments that fostered efficient progress and completion of CT curricula benefited most from the CT. Specifically, the study found that students who completed at least 27 sections of CT content experienced statistically significant (and substantively significant) improvement in learning outcomes compared to equivalent students in a matched control group. This is perhaps surprising because the completion of 27 sections of CT content can usually occur within roughly 13 hours of CT usage, a stark contrast to Carnegie Learning's recommended 50 hours of use over an academic year. Consistent with, and in addition to, these results from experimental data, observational data analyses suggest that the number of sections mastered per hour is also a strong predictor of standardized test scores (Ritter et al. 2013; Joshi et al. 2014). These results allude to a simple fact: students need to engage with the mathematics in order to achieve better outcomes. That is, engaged time, and with the help of the CT and engaging teaching practices, academic learning time, are necessary to improve learning outcomes.

The present study provides further evidence for and a real world illustration of the fact that student time using ITSs like CT, especially instructional time and engaged time, is insufficient for substantially improved learning; teacher practices must encourage increases in academic learning time and its efficient use by students for learning.

Data

We consider CT Algebra usage data from a large school district in the southeastern United States, which also provided data on whether individual students passed or failed a state-level Algebra exam at the end of their Algebra I course. Our original analysis of the dataset was prompted by a consulting request from the district about its implementation of the CT and how it might be improved. The dataset includes 2,025 students across 18 high schools and 3 middle schools. District level policy dictated that the high school students that failed their mathematics year-end standardized test in the previous academic year used CT in their high school Algebra course. Students in the three middle schools in the dataset are somewhat more likely to be higher performing, as they are in Algebra courses in middle school.

Analysis & Results

Figure 2 shows usage of CT (measured in number of hours per student) by school, along with the percentage of these students passing the end-of-year Algebra exam at that school. None of the schools show usage levels consistent with our recommended 50 hours per year, although four schools do show reasonable usage of more than 35 hours per year. The largest cluster of schools shows usage of less than 5 hours per student for the year. Prior research (Ritter et al. 2013) has shown fewer than five hours exposure to CT in a school year is insufficient to show any association or correlation with test scores.

Considering all schools in Figure 2, there is no evident relationship between pass rates and time. Such a relationship might obtain if we removed schools with very low usage and two outliers, but this is clearly a weak basis for claiming a relationship.

Figure 3 shows the relationship between number of sections (roughly math topics) mastered per hour and passing rates on the end-of-year exam, by school. The use of mastery per hour, as has been noted, is an efficiency measure: how much work do students accomplish when they are in class? Figure 3 demonstrates a stronger relationship between this efficiency measure and student scores. Figure 3 also explains Outlier 1 and Outlier 2 in Figure 2. Although students in these two schools have spent a relatively large amount of time logged on to the software (cf. Figure 2), they are at the bottom of the scale with respect to how much mathematics they master in that time.

As this observational study arises out of a real world, consulting use case, when school district administrators were presented data about these two outlier schools, they immediately provided a plausible explanation; these two districts had long-term substitute teachers in place for mathematics for much of the school year. These teachers were likely not invested in high quality implementation, lacked professional development from Carnegie Learning staff, and likely did not actively engage students using the CT. It is likely that such teachers believe that the ITS is responsible for teaching students, with little intervention from the teacher. This illustrates the vital role of teachers in the implementation of ITSs like the CT.

In a well-implemented computer lab, most of the students should be able to work on mathematics within the CT and make progress on their own, sometimes relying on hints within the software. This allows the teacher to talk one-on-one with students who have deep misunderstandings or other questions that are not well addressed within the software. Teachers also play a strong motivational role in the computer lab, setting expectations for the amount of work that students are expected to complete and how students may constructively interact, collaborate, and assist one another. Teachers also ensure that students remain on-task, and they keep students focused on improving their mathematics skills while avoiding detrimental behavior like so-called “gaming the system” whereby students attempt to make progress without deep understanding of mathematics content (Baker et al. 2004; Fancsali 2014).

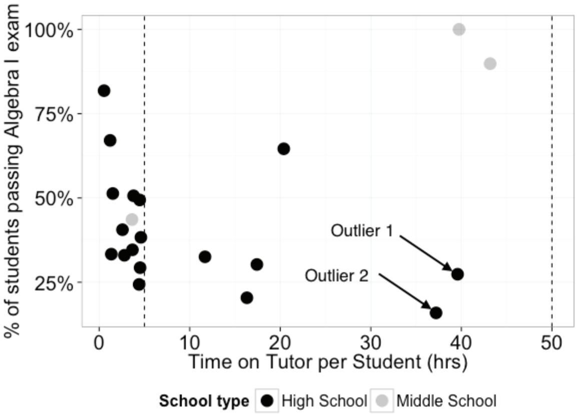


Figure 2. Pass rates on end-of-year Algebra exam, by school and Cognitive Tutor usage in the schools. The vertical bar at 5 hours represents minimal usage, below which we do not expect a meaningful relationship between usage and outcomes. The vertical bar at 50 hours represents our recommended yearly usage.

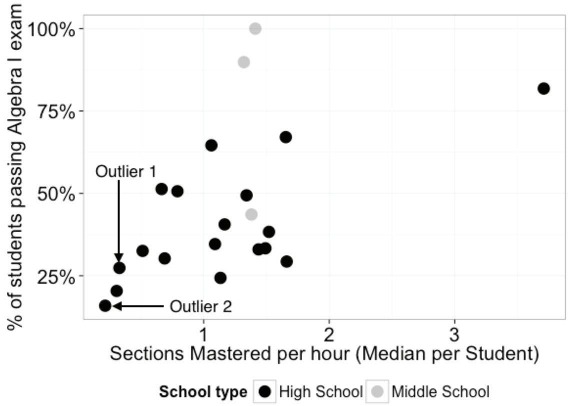


Figure 3. Pass rates on end-of-year Algebra exams, by school and sections mastered per hour. Two high schools with high rates of usage show very low rates of efficiency in section mastery, indicating that the time they are spending with the software is not being well used.

Figure 4 demonstrates a readily apparent relationship between time, sections mastered, and end-of-year Algebra exam outcomes at the individual student level. This graph shows that students who use CT for a sufficient amount of time (over five hours) and who use that time well (falling above the diagonal representing average sections mastered per hour) are highly likely to succeed on the year-end exam.

As noted, a recent study (Sales and Pane 2015) leveraged data from a large-scale randomized control trial (Pane et al. 2014) to find that students who completed more than 27 CT sections greatly outscored equivalent students using a standard curriculum. In that dataset, 27 sections represented the median completion rate, but this number is well below the scope of the full curriculum (approximately 140 sections) or Carnegie Learning's recommendations (75 sections). Still, that level of engagement and success with the software was sufficient to produce favorable pass rates; approximately 72% of those completing more than 27 sections passed the year-end exam, even among many high school students who had previously failed the exam.

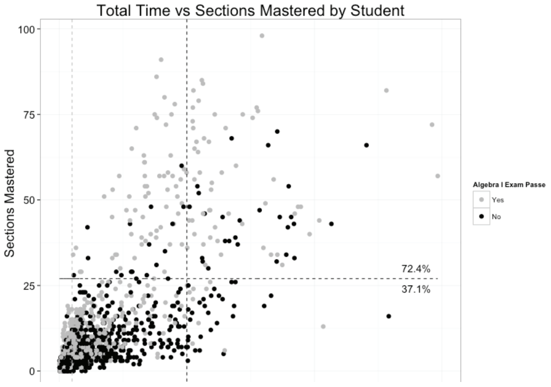


Figure 4. Relationship between time spent on CT, number of sections mastered and Algebra I exam outcome. Each point represents a student. Almost 73% of students completing > 27 sections of the curriculum pass the exam (lighter points), as compared to about 37% of students who completed fewer sections. The vertical line at 50 hours represents Carnegie Learning's recommended usage time. The horizontal line at 27 sections represents the point at which a previous study found strong impact of CT curriculum (Sales and Pane 2015).

Finally, we estimate random and mixed effects regression models to test our intuitions and hypotheses about the relative contribution of school and teacher to accounting for the variability in student time and efficiency. First, we specify and estimate a random effects regression model to predict student time using students' schools and classes each as random effects, for which we estimate random intercepts with fixed means. Schools explain 19.0% of the variability in log-transformed student time while teachers explain 27.6% of this variability in the model including both random effects. This is somewhat surprising, as we expected that school-level policy would play a larger role in determining in time; this result suggests a larger role for teachers. Overall, this model explains 38.6% of the variability in log-transformed student time. This model is preferable to random effects models including each factor individually in that the combined model's AIC and BIC score decreases, indicating an improved goodness of fit without likely over-fitting the data.

Next, we specified and estimated a mixed effects linear regression model to predict sections mastered (or completed) per hour (i.e., student efficiency) with student CT time as a fixed effect and school and class random effects (random intercepts with fixed means). While time alone explains 34.1% of the variance in student efficiency, the teacher random effect explains an additional 11% of variability while school only explains 1.5% of variability. This aligns with our hypothesis that student efficiency is more likely to be driven by teaching practices rather than school-level factors. Nevertheless, the school-level random effect stands in for a number of factors for which data were unavailable, including socio-demographics and other factors that may be associated with learning and efficiency. Model selection again proceeded by noting decreases in AIC and BIC scores; the model including all three effects accounts for 45.7% of variability in sections mastered per hour.

Discussion

The present study considers an important element of teaching practices in classrooms and computer labs in which intelligent learning technologies like ITSs, including the CT, are deployed to improve student learning. We use data from a school district that sought data-driven consulting about implementation factors and implementation fidelity to illustrate differences in how students' time can be used (or not used) to work through material in the CT and have suggested that learner efficiency is driven by whether teachers take an active role in turning engaged time into academic learning time by cognitively, behaviorally, and affectively supporting students and encouraging students to mindfully be "on-task," avoid wasting time and behavior that does not enhance learning.

While ITSs promise to enhance learning by providing tutoring that approaches the effectiveness of one-on-one tutoring, they are almost never intended to fully replace teachers in classrooms and computer labs. Focusing on ways in which we can help teachers, via professional development, data-driven reporting, dashboards and other means, to effectively deploy and implement learning technologies, like ITSs and others, is a long-standing, but perhaps increasingly more prominent, concern in the literature, and more pressingly in practice (e.g., Schofield 1995, Feng and Heffernan 2005, Chronaki and Matos 2014, Cowan 2011, among many others).

In on-going and future research, we are exploring ways in which to use data to manage and optimize "hand-offs" between automated instructional systems like ITSs and human instructors and teachers. This research will help us to better understand when automated systems like ITSs might help improve learning by informing a student that they should seek help from their teacher. It will also help to develop best practices for teachers and instructors to know how and when it is best for them to engage students having difficulties in problem solving versus letting affordances like hints provide necessary support.

References

Aronson, J., Zimmerman, J., Carlos, L. 1998. Improving Student Achievement by Extending School: Is It Just a Matter of Time? Technical Report, WestEd. Retrieved October 21, 2015 from http://www.wested.org/online_pubs/po-98-02.pdf.

Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. 2004. Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System." In *Proceedings of ACM CHI 2004*, 383-390. New York: ACM.

Bloom, B.S. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher* 13(6): 4-16.

Chronaki, A., Matos, A. 2014. Technology Use and Mathematics Teaching: Teacher Change as Discursive Identity Work. *Learning, Media and Technology* 39(1): 107-125.

Corbett, A.T., Anderson, J.R. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User Adapted Interaction* 4: 253-278.

Cortes, K., Goodman, J., Nomi, T. 2013. A Double Dose of Algebra. *Education Next* 13(1): 71-76.

Cowan, P. 2011. The Four I-model for Scaffolding the Professional Development of Experienced Teachers in the Use of Virtual Learning Environments for Classroom Teaching. In *Proceedings of EdMedia: World Conference on Educational Media and Technology 2011* (pp. 130-136). Association for the Advancement of Computing in Education (AACE).

Fancsali, S.E. 2014. Causal Discovery with Models: Behavior, Affect, and Learning in Cognitive Tutor Algebra. In Proceedings of the 7th International Conference on Educational Data Mining, 28−35. International Educational Data Mining Society.

Feng, M., Heffernan, N.T. (2005). Informing Teachers Live about Student Learning: Reporting in the ASSISTment System. In Proceedings of *the 12th Annual Conference on Artificial Intelligence in Education Workshop on Usage Analysis in Learning Systems*.

Feng, M., Roschelle, J., Heffernan, N., Fairman, J., Murphy, R. 2014. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)*, 561-566. Switzerland: Springer International.

Joshi, A., Fancsali, S.E., Ritter, S., Nixon, T. 2014. Generalizing and Extending a Predictive Model for Standardized Test Scores Based On Cognitive Tutor Interactions. In Proceedings of the 7th International Conference on Educational Data Mining, 369-370. International Educational Data Mining Society.

Karweit, N. 1985. Should we lengthen the school term? *Educational Researcher* 14(6): 9-15.

Koedinger, K.R., Anderson, J.R., Hadley, W.H., Mark, M.A. 1997. Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education* 8: 30-43.

Koedinger, K.R., Sueker, E.L.F. 2014. Monitored Design of an Effective Learning Environment for Algebraic Problem Solving, Technical Report, CMU-HCII-14-102, Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA.

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R.S., Hatala, M. 2015. Penetrating the Black Box of Time-on-task Estimation. In *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*, 184-193. New York: ACM.

Miller, W.L., Baker, R.S., Labrum, M.J., Petsche, K., Liu, Y-H., Wagner, A.Z. 2015. Automated Detection of Proactive Remediation by Teachers in Reasoning Mind Classrooms. In *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*, 290-294. New York: ACM.

Pane, J.F., Griffin, B.A., McCaffrey, D.F., Karam, R. 2014. Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis* 36(2): 127-144.

Pane, J.F., McCaffrey, D.F., Steele, J.L., Ikemoto, G.S., Slaughter, M.E. 2010. An Experiment to Evaluate the Efficacy of Cognitive Tutor Geometry. *Journal of Research on Educational Effectiveness* 3(3): 254-281.

Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T. 2007a. Cognitive Tutor: Applied Research in Mathematics Education. *Psychonomic Bulletin and Review* 14(2): 249-255.

Ritter, S., Joshi, A., Fancsali, S.E., Nixon, T. 2013. Predicting Standardized Test Scores from Cognitive Tutor Interactions. In Proceedings of the 6th International Conference on Educational Data Mining, 169−176. International Educational Data Mining Society.

Ritter, S., Kulikowich, J., Lei, P., McGuire, C.L., Morgan, P. 2007b. What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. In Hirashima, T., Hoppe, U., Young, S. S. eds. *Supporting Learning Flow through Integrative Technologies*, 13-20. Amsterdam: IOS Press.

Ritter, S., Yudelson, M., Fancsali, S.E., Berman, S.R. 2016. How Mastery Learning Works at Scale. In Proceedings of the 3rd Annual ACM Conference on Learning @ Scale. New York: ACM.

Sales, A., Pane, J.F. 2015. Exploring Causal Mechanisms in a Randomized Effectiveness Trial of the Cognitive Tutor. In Proceedings of the 8th International Conference on Educational Data Mining. International Educational Data Mining Society.

Schofield, J.W. 1995. *Computers and Classroom Culture*. Cambridge, England: Cambridge University Press.

Taylor, E. 2014. Spending More of the School Day in Math Class: Evidence from a Regression Discontinuity in Middle School. *Journal of Public Economics* 117: 162-181.

Acknowledgments

Ambarish Joshi wrote code on which these analyses rely, for which the authors thank him.