J. EDUCATIONAL COMPUTING RESEARCH, Vol. 43(4) 489-510, 2010

A QUASI-EXPERIMENTAL EVALUATION OF AN ON-LINE FORMATIVE ASSESSMENT AND TUTORING SYSTEM*

KENNETH R. KOEDINGER
ELIZABETH A. McLAUGHLIN
Carnegie Mellon University

NEIL T. HEFFERNAN
Worcester Polytechnic Institute

ABSTRACT

ASSISTments is a web-based math tutor designed to address the need for timely student assessment while simultaneously providing instruction, thereby avoiding lost instruction time that typically occurs during assessment. This article presents a quasi-experiment that evaluates whether ASSISTments use has an effect on improving middle school students' year-end test scores. The data was collected from 1240 seventh graders in three treatment schools and one comparison school. Post-test (7th grade year-end test) results indicate, after adjusting for the pre-test (6th grade year-end test), that students in the treatment schools significantly outperformed students in the comparison school and the difference was especially present for special education students. A usage analysis reveals that greater student use of ASSISTments is associated with greater learning consistent with the hypothesis that it is useful as a tutoring system. We also found evidence consistent with the hypothesis that teachers adapt their whole class instruction based on overall

*This research was funded by grants from the U.S. Department of Education Institute of Education Sciences (Grant #R305K030140 and Grant #R305A07044).

doi: 10.2190/EC.43.4.d
http://baywood.com

student performance in ASSISTments. Namely, increased teacher use (i.e., having more students use the system more often) is associated with greater learning among students with little or no use, suggesting that those students may have benefited from teachers adapting their whole-class instruction based on what they learned from ASSISTments use reports. These results indicate potential for using technology to provide students instruction during assessment and to give teachers fast and continuous feedback on student progress.

High stakes testing has become an ever-present force in American education. Consequently, both formative testing (a diagnostic tool used for immediate remediation and adaptation of teaching practices) and benchmark testing (a measure of how students perform compared to a set of criteria such as state standards) have become increasingly important and are occurring with greater frequency in the classroom. The No Child Left Behind Act of 2001 has exerted accountability pressures on school administrators, teachers, and students. In order to meet the requirements of the NCLB Act, educators are searching for ways to assess student deficiencies, realign curriculums, and alter classroom practices to meet district and state standards. The accountability pressure has led to increased focus on benchmark assessments and practice tests on top of the usual end-of-chapter testing. The hope is that such assessment will help determine what instruction or remediation is needed to raise student achievement and consequently raise their test scores on the high-stakes year-end exam. The ASSISTments program, a web-based mathematics cognitive tutor developed for middle school students, was designed to address the need for assessment while simultaneously providing instruction to students, thereby preventing the loss of instruction time that typically occurs during assessment. This article presents an evaluation of the ASSISTments system, but first we review relevant aspects of prior educational technology evaluations.

EDUCATIONAL TECHNOLOGY GOALS AND BENEFITS

Many early studies of computer use in schools (particularly drill and practice, tutorials, or educational games) reported positive results for the effect of computer technology on student achievement (Guerrero, Walker, & Dugdale, 2004; Honey, Culp, & Carrigg, 2000; Kulik, 2003). Today, advances in technology have created a much richer environment where computers are used for instruction, communication, collaboration, and student research. The improvements in technology, however, do not seem to have led to concomitant advances in achievement, rather the overall results are mixed and the effects of educational technology use are typically small (Angrist & Lavy, 2002; Bielefeldt, 2005; Kulik, 2003; Waxman, Connell, & Gray, 2002).

Researchers have cautioned against thinking of technology as a panacea to the achievement problems in education. Technology, by itself, does not produce nor promote learning (Alspaugh, 1999; Fuchs & Woessmann, 2004; Honey et al., 2000; Schacter & Fagnano, 1999). For technology to impact learning, a number of contextual variables need to be considered, such as the quality of implementation (Bielefeldt, 2005; Wenglinsky, 1998), teacher expertise, knowledge and pedagogical philosophy (Becker & Ravitz, 2001; Fetler, 1999; Odden & Borman, 2004; Vandevoort, Amrein-Beardsley, & Berliner, 2004), teacher support and training (Fetler, 1999), and students' readiness. Readiness to adequately use technology for what and how it is intended, preparedness with regards to content relevance, and more specifically, how adaptive technologies may help to provide more differentiated instruction that better addresses variability in student readiness. Further, technology should be designed with learning principles in mind. In a review by Schacter and Fagnano (1999), conventional instruction was compared to newer technologies including Intelligent Tutoring Systems (ITS), and they concluded that educational technology based on cognitive theory increases student learning and understanding. Anderson, Corbett, Koedinger, and Pelletier (1995), for example, report the success of early cognitive tutors (e.g., ACT Programming and Geometry Proof Tutor) is a result of developing a model that represents student competence as a set of production rules.

Cognitive tutors have been shown to be effective in various domains and under various conditions (Koedinger & Aleven, 2007). Well-defined domains have included algebra (Koedinger, Anderson, Hadley, & Mark, 1997), statistics (Meyer & Lovett, 2002), and programming (Naser, 2009), while ill-defined domains have included legal reasoning (Aleven, 2003) and intercultural competence (Ogan, Walker, Aleven, & Jones, 2008). And some are in widespread use, for instance, the Algebra Cognitive Tutor course is used regularly by about 500,000 students per year.

Bouck and Flanagan (2009) observe that the technology principle presented in the Principles and Standards for School Mathematics (National Council of Teachers of Mathematics [NCTM], 2000) indicates that technology should be available to all students, and technology has the potential to benefit students with disabilities. The diverse range of disabilities (i.e., from physical handicaps to cognitive difficulties to more serious mental disorders) covered by the special education label challenges the educational system to devise curriculums that support all learners while integrating technology into the 21st century classroom. Historically, accessibility was the major problem for special needs students, but assistive technology devices have made technology usable for individuals with special needs (Edyburn, 2001). Computer technology, for example, provides an unthreatening environment where students work individually and at their own pace. Kimmel, Deek, and Frazer (1996) concluded that special education students will be better served if we move away from textbook learning and move toward a more hands-on approach. Woodward and Rieth (1997) described how computer assessment has evolved into more of a formative assessment that models students' cognitive abilities. Thus, formative assessment has become a vehicle for differentiated instruction and should help meet the needs of special education students.

In addition to student learning from computer technology, both teachers and students report more time spent on task, increased motivation, and enhanced confidence (e.g., Schofield, Evans-Rhodes, & Huber, 1990). Furthermore, the potential added value of computer literacy should not be discounted (Kmitta & Davis, 2004). Even with these potential added benefits, one cannot evaluate technology without also considering cost factors.

PRACTICAL CONSIDERATIONS AND COSTS

As accountability measures become more prevalent, the cost of additional testing becomes more transparent (e.g., the loss of instructional time, potential teaching to the test, monetary expenses to grade and report results). The need for a quick turn-around, the increase in volume of assessments, and the per student expense are just a few examples of the problems educators face due to accountability pressures. These problems illustrate the struggle between accountability and instruction. The question arises of how to best achieve accountability without jeopardizing instructional time and especially time spent on deeper conceptual learning and more creative and innovative thinking. Yeh (2009) has suggested that educational technology's greatest contribution will come in the form of rapid formative assessment and the cost-effectiveness gains from doing the work of assessment more efficiently.

The ASSISTments system's claim to efficiency is that it performs the dual tasks of assessing and tutoring at the same time in an online environment. Automated assessment reduces teacher administrative tasks such as grading and can provide immediate results on student and class deficiencies, thus freeing teachers for other activities such as collaboration or supporting struggling students. Besides providing teachers with an assessment management tool, ASSISTments offers a fine granularity of assessment that allows for a more skill specific understanding of student limitations than typical paper-based benchmark assessments. Prior psychometric models of ASSISTments data have indicated that a finer-grained skill model outperforms the courser grained models for assessing student performance (Feng, Heffernan, Mani, & Heffernan, 2006; Pardos, Heffernan, Anderson, & Heffernan, 2006).

Students also benefit from the ASSISTments system by receiving immediate corrective feedback in the form of scaffolded questions and hints. By combining the benefits of automated assessment with tutoring and feedback ASSISTments efficiently and creatively uses technology to improve student academic performance.

Given our concern for the loss of instructional time due to the demands of the NCLB Act and the increase of testing in schools, we reviewed the literature for evaluations of benchmark testing. The Center on Educational Policy, a public education advocacy group, claims testing is the most "defensible" means for making inferences about student learning and believes testing will remain in the forefront of educational assessment. Although the aim of benchmark testing is to improve classroom instruction and ultimately improve student achievement, Henderson, Petrosino, Guckenburg, and Hamilton (2007a, 2007b) found no significant difference on middle school math scores between schools who used quarterly benchmark assessments and comparison schools. According to that report, the findings may be the result of data limitations. However, clearly if benchmark tests are used as summative assessments and no instructional actions are taken in response to a benchmark assessment, achievement gains will not be observed. We next describe the ASSISTments system and how it provides two avenues for instructional response to assessment results: teacher adaptation and student learning. We then describe a quasi-experiment of ASSISTments use designed to evaluate its role as a combined instructional, formative, and benchmark assessment system.

ASSISTment

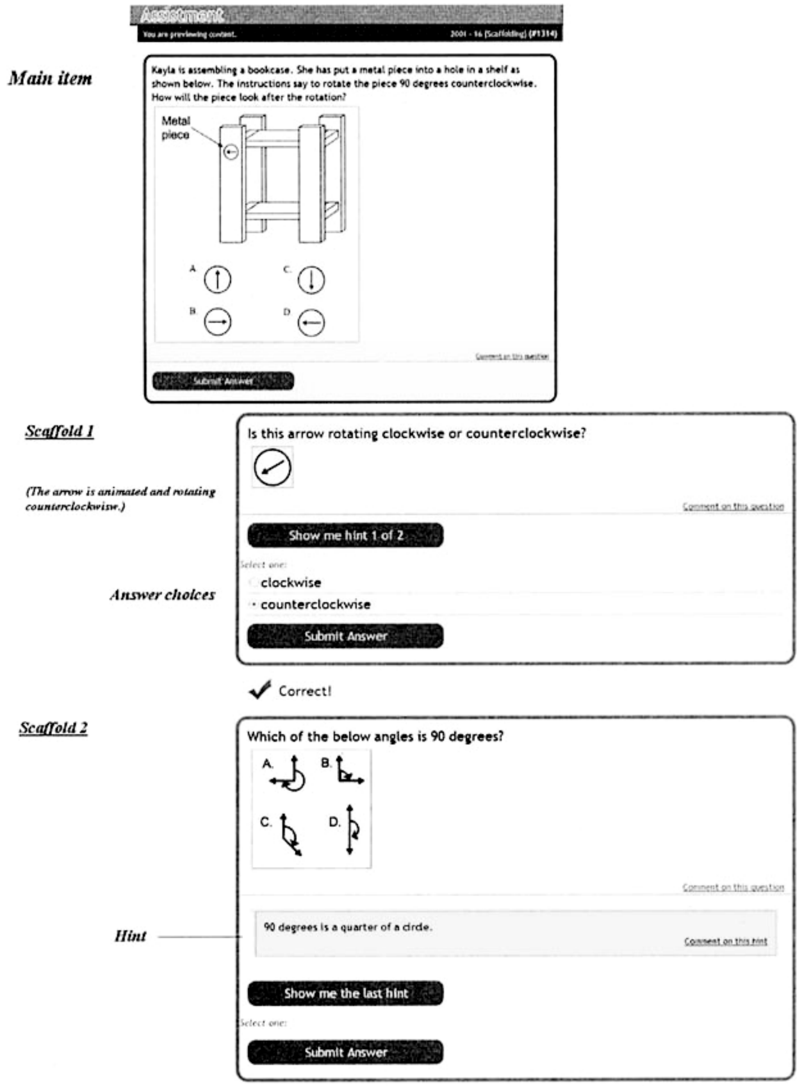
The name ASSISTments was coined by Ken Koedinger to describe a kind of assess*ment* that provides instructional *assist*ance during the test. The system provides timely feedback to assist teachers in making classroom decisions while simultaneously tutoring students as needed. Teachers can learn about particular difficulties their students are having both by directly observing and interacting with students in the computer lab and by looking at the detailed reports that ASSISTments provides. They can respond to what they learn by changing their whole class instructional strategies, for instance, to directly address a difficulty being experienced by many students. The prompt and reliable reporting is a key asset of the ASSISTments system especially since one of the prolific complaints heard about both standardized and benchmark testing is the inopportune delivery of results. It is not possible for teachers to adjust to their current students' needs when test results are not available until the semester is over. Even with quarterly assessments that some assessment services are providing (e.g., 4Sight, Galileo, Pearson Benchmark), teachers can only react to these results three times a year and only for a limited number of test items.

ASSISTments functions as an assessment tool by collecting data on a variety of dynamic metrics that go beyond the typical correct/incorrect measures found in traditional paper and pencil assessment. These metrics include various measures of the assistance needed by a student including the number of attempts made, the number of hints requested, the response time, and the number of opportunities to practice. As students work through items and answer follow-up questions, the system gathers information that is used to ascertain individual student strengths and weaknesses as well as the actions of an entire class. The results are made available to teachers in the form of on-line reports and automated e-mails. Using dynamic metrics to gain information on student performance may turn out to be a better predictor of future student performance on year-end standardized exams, for instance the Massachusetts Comprehensive Assessment System (MCAS), than conventional paper and pencil tests or benchmark assessments. If so, a year's worth of data in a program like ASSISTments may be a viable alternative to a couple hours of high stakes test results. Feng et al. (2006) showed that using the amount of assistance required (i.e., number of attempts and number of hints) made for better predictions on the MCAS exam than correctness alone. In fact, they found that 8th grade ASSISTments use can predict 10th grade achievement as well as the 8th grade MCAS (Feng et al., 2008).

As a tutor, ASSISTments functions by breaking down or scaffolding problems into requisite skills and knowledge components. If a student incorrectly answers the original item or requests help, the first scaffold is automatically presented. Once in the scaffold tutoring, students must complete the series of scaffolds for that item (typically about three scaffolds). The example item in Figure 1 has two scaffolds: the first assesses knowledge of the term clockwise/counterclockwise, and the second focuses on whether a student knows how a 90 degree angle appears graphically. By scaffolding the original item, we get a clearer picture of where a student's difficulty lies if he gets the item wrong. In other words, an incorrect response on the original item does not indicate whether a student doesn't know what counterclockwise means or whether he doesn't know what a 90 degree angle looks like. The scaffolds help supply a more complete picture of individual student and class deficiencies. In addition to scaffolds, students can get assistance by requesting hints. Anytime a student feels confused or is unable to answer, he can ask the system for help. Hints are suggestions on how to proceed and often appear as a definition or question similar to what a human tutor might ask or say. In the example item, Scaffold 1 illustrates counterclockwise in a real world situation by using an animated arrow while the first hint for this scaffold defines clockwise as the direction the hands on a clock move and counterclockwise is the other direction. A noteworthy study by Razzaq and Heffernan (2006) demonstrated the benefit of completing problems with scaffolds as compared to doing problems with hints only. In a more recent study, Razzaq and Heffernan (2009) found that less proficient students benefit the most from scaffolding when controlling for time.

RESEARCH QUESTIONS AND HYPOTHESES

In this study, we were interested in examining whether students learn more from using the ASSISTments system, as measured by the MCAS test, than a group



Kayla is assembling a bookcase. She has put a metal piece into a hole in a shelf as shown below. The instructions say to rotate the piece 90 degrees counterclockwise. How will the piece look after the rotation?

Is this arrow rotating clockwise or counterclockwise?

(The arrow is animated and rotating counterclockwise.)

○ clockwise
● counterclockwise

✓ Correct!

Which of the below angles is 90 degrees?

90 degrees is a quarter of a circle.

Figure 1. Example of an ASSISTment item and its tutoring. If a student answers the main item incorrectly or requests help, he is presented with scaffolds that assess the individual skills from the main item. Students also have the opportunity to request hints as seen in the second scaffold.

of comparison students who did not use ASSISTments. Our analysis factored out students' prior year MCAS test results and used post-intervention MCAS scores as the dependent variable. We hypothesized that students would benefit from the tutoring, feedback, and design of the ASSISTments system and that their progress would be observed by improved MCAS test scores. In addition to a positive effect for students in general, we hypothesized that certain typically disadvantaged sub-groups of students (e.g., special education) may show greater advances in learning as a result of using ASSISTment because their learning needs are less likely to be met in whole class instruction. A few of the advantages offered by educational technology that are not necessarily readily available in traditional classroom instruction include supportive feedback, an interactive and multi-sensory learning environment, and more time for the teacher to provide individual assistance. Teachers have more time to help individual students in the computer lab because they are not running a whole class session and students, other than the one with whom they are interacting, are getting support from the technology.

The benefits of practice with timely feedback (Roediger & Karpicke, 2006) and individualized tutorial assistance (e.g., Corbett & Anderson, 2001; Koedinger & Aleven, 2007) have been documented to support student learning and achievement. We were interested to see if a higher amount of student use of ASSISTments would make a difference on post-test outcomes. We predicted higher usage would result in a greater effect on students MCAS scores.

Informal discussions with teachers have provided testimony of satisfaction with and endorsement of ASSISTments. Some teachers have reported specific changes they made to their classroom instruction. For example, Ms. Metelenis[1] reviewed an ASSISTments report that showed 70% (14) of her students needed help with a word problem that featured the skill decimal multiplication. As a result, the teacher spent an extra 15 minutes of class time to discuss a similar word problem. Of the 14 students who originally needed assistance, 50% benefited from the additional instruction in the sense that they solved a related item correctly the next time they used ASSISTments. We hypothesized that students of high usage teachers would perform better on the MCAS because we expect these teachers are more likely to alter their teaching strategy. In particular, we hypothesized that the students who had little or no ASSISTments use would benefit most from teachers who frequently use the ASSISTments system. This outcome would suggest that teachers are utilizing what they learn about their students' proficiencies and deficiencies and are making appropriate changes to classroom instruction.

[1] Ms. Metelenis is an alias, but the case study report is real. Although the report comes from a subsequent year, it demonstrates how the data from ASSISTments can be used by a teacher to her students' advantage.

METHOD

The hypotheses we wish to evaluate are causal claims, and thus an intervention study is warranted. Interestingly, Robinson, Levin, Thomas, and Vaughn (2007) report an increase in causal statements in teaching-and-learning journals from 1994 to 2004, even though during this same time period there was a decline in intervention studies. While the gold-standard for intervention studies is the controlled randomized experiment, the feasibility of running a true experiment in the field is often prohibited by practical issues of compliance, cost-effectiveness, and the ethics of withholding a potentially positive intervention. Educational researchers (Berliner, 2002; Borman, 2002; Slavin, 2008) have recognized this quandary and note the benefits of a well-designed quasi-experiment. Indeed, Slavin (2008) regards the outcomes from high quality quasi-experiments to be close approximations to those of an experiment. In early development of novel technological innovation, like ASSISTments, the extra costs of a controlled randomized experiment are particularly prohibitive and perhaps not justified. Thus, we pursued a quasi-experiment as an initial evaluation of an ASSISTments intervention.

Participants

For this study, we focused on students from four middle schools in an urban school district in Massachusetts. The full study sample was a group of seventh graders (n = 1,344). The final analysis included only those students with MCAS (Massachusetts Comprehensive Assessment System) scores from both 2006 (6th grade test) and 2007 (7th grade test), and students whose math teacher assignment could be determined either from ASSISTments use or MAP[2] test data. The resultant pool included 1,240 students of which 79% were regular students (n = 985) and 21% were special education students (n = 255). A breakdown of the individual schools shows Treatment school A included 372 students (78% regular), Treatment school B included 322 students (81% regular), Treatment school C included 253 students (77% regular), and Comparison school D included 293 students (81% regular).

This study was a quasi-experiment in that there was no random assignment of students (or classes or schools) to condition. Rather, school D was in the comparison group because they did not have an adequate number of computers at the time the decision to use ASSISTments was made. Due to the lack of computers, school D did not use educational technology as a supplement to their math curriculum. Instead, when treatment students were in the lab, control students would work on traditional text-book activities.

[2] MAP or Measures of Academic Progress are computerized adaptive tests administered three times during the school year.

Both the treatment and comparison groups had similar student characteristics (race, gender, limited English proficiency (LEP), free lunch eligibility, special education students) and teacher/school characteristics (licensed in teacher assignment, percent of core classes taught by "highly qualified teachers," student/teacher ratio). These demographics were compatible with the district, but varied from the state profile. When compared to the state's school demographics, our participants had a higher Hispanic and lower white population, higher LEP, higher English not the first language, and higher percentage of students failing/needing improvement on the MCAS.

Measures

Students' 6th grade MCAS scores were used as a pre-assessment of their incoming knowledge and students' 7th grade adjusted MCAS scores was the dependent variable. Also, both student and teacher usage measures were used to examine the impact of usage on students' 7th grade MCAS scores post intervention. High usage students were defined as those who had completed 60 or more items, low usage students completed less than 60 items, and non-usage students did not use ASSISTments. We decided not to define low vs. high usage using the median (30 items) because to call students "high users" with such a low threshold did not seem appropriate. We chose 60 items as a reasonable threshold as it reflects approximately 2 hours of content, and indeed, the treatment school with the greatest usage and the greatest gain in test scores averaged about 60 items (56.9). Teacher usage was based on student participation because we did not have the means to determine how often teachers read reports nor how they specifically used the information derived from the reports. Teachers were considered high usage if 25% or more of their students completed 60 or more items and low-usage if they had less than 25% of their students complete at least 60 items.

RESULTS

Before discussing the comparison results, it is worth noting that the MCAS test is designed to differentiate students from quite a broad range of potential ability, perhaps 2-3 grade levels above and below the target grade. As a consequence, many of the items on the 7th grade test are either above or below the level of the content addressed in 7th grade and so we might well expect that the 7th grade instruction will yield improvement on only a fraction of the items on the test. We wondered what is a reasonable expectation for an increase in score from a year's worth of instruction. The ideal comparison would be to give some students the same test at the beginning and end of the school year. We do not have such data, but we do have ASSISTment data for students at different grade levels (7th and 8th) solving the same items (selected from the

6th and 8th grade MCAS).[3] For items that were completed by at least 50 7th graders and at least 50 8th graders, performance of 8th graders was 65.5% while 7th graders was 61.3%. This leads to a rough estimate of one year of schooling being associated with about a 4.2% increase in the test score. Even given the argument above about the test containing only a fraction of course-relevant items, we were surprised at how small the measured changes appear to be.

However, a quite similar result was found in a case where a standardized test (the ETS Algebra test) was given to the same students at the beginning and end of a school year. Here the scores of 1404 students improved from 32.3% to 36.5% (Dynarski, Agodini, Heaviside, Novak, Carey, Campuzano, et al., 2007, p. 98); also, coincidentally, only a 4.2% increase from a year's worth of school. We present these figures to put our results on condition differences in context. Namely, we should expect the MCAS test to reveal only relatively small changes as the result of any treatment.

Overall Results

Table 1 shows the percent correct pre-test and post-test means for treatment and control conditions. Note that since the 7th grade test (the post-test) is harder than the 6th grade test (the pre-test) we should not expect an increase in score. We used an ANCOVA to test whether the treatment and control groups differed in their post-test scores (2007, 7th grade MCAS test) after taking into account the pre-test scores (2006, 6th grade MCAS test). A 2 × 2 ANCOVA with condition (treatment vs. control) and student group (regular vs. special education) as factors and pre-test as a covariate revealed main effects for condition, $F(1, 1235) = 12.3$, $p < .001$, and student group, $F(1, 1235) = 119.4$, $p < .001$, and an interaction effect between condition and student group, $F(1, 1235) = 6.6$, $p = .01$. To get a sense for the implications of this difference, we can use the ANCOVA results to compute an adjusted post-test score for both groups that assumes the pre-test scores were equivalent. Adjusted post-test means were computed using a modified version of Searle, Speed, and Milliken's (1980) estimated marginal means as found in the SPSS GLM statistical package.

As shown in the last columns of Table 1, the adjusted post-test means for all (total) students using ASSISTment (M = 51.7, SE = .45) is higher than the adjusted means for the comparison group (M = 48.4, SE = .82) with a difference of 3.3% (51.7-48.4). Relative to the estimate above of about 4.2% gain from a year of schooling, this difference is sizeable, equivalent to a boost of 7 months of schooling (3.3/4.2 * 9 months). The effect size of this difference (.23) is small (using Cohen's

[3] Both the 7th and 8th grade classes completed items from the 6th and 8th grade test that were deemed relevant to their curriculum. At the time of this study, 7th grade MCAS items were not available.

Table 1. Percent Correct of Pre-Test, Post-Test, and Adjusted Post-Test Means

Condition	Student group	Students	Schools	Teachers	Classes	Pre-test means	Post-test means	Unadj. std. err.	Adj. Post-test means	Adj. std. err.
Control	Regular	237	1	4	NA[a]	64.61	60.64	1.34	54.94	0.72
	Sp. Ed.	56	1	NA[b]	NA[a]	46.63	34.26	2.25	41.91	1.48
	Total	293	1	NA[b]	NA[a]	61.19	55.60	1.31	48.43	0.82
Treatment	Regular	748	3	12	42	60.65	58.58	0.77	55.83	0.41
	Sp. Ed.	199	3	NA[b]	NA[a]	36.72	32.57	1.09	47.57	0.83
	Total	947	3	NA[b]	NA[a]	55.63	53.12	0.74	51.70	0.45

[a] Data on assignment of students to class was available only for regular teachers with an ASSISTment account.

[b] Special education students are sometimes assigned a regular teacher and otherwise assigned a special education teacher (each school has at east three of these teachers).

d and the adjusted means; see http://web.uccs.edu/lbecker/Psy590/escalc3.htm). Both regular and special education students in the treatment condition performed better than the comparison students on the post-test when controlling for the pre-test. As shown in Figure 2, this effect was much larger for special education students, a 5.7% gain (47.5% for treatment, 41.9% for control), than for regular students, a 0.9% gain (55.8% for treatment, 54.9% for control). Statistical analysis of the simple main effects indicate that the treatment difference for special education students is statistically significant, $F(1, 1235) = 11.44$, $p < .001$, and the treatment effect is medium ($d = .50$). The difference for regular students does not reach significance, $F(1, 1235) = 1.16$, $p = .28$, $d = .08$.

Additional analyses were run on various sub-group populations including gender, race, free lunch availability and limited English proficiency (LEP). As can be seen in Figure 3, the treatment condition outperforms the control group in all of the sub-groups on the adjusted percent correct means of the 2007 MCAS exam. Individual ANCOVA's were run on each sub-group with a main effect found for free lunch availability, $F(1, 1235) = 6.63$, $p = .010$, non-white (white vs. combined Black and Hispanic, excluding Asian, Native American, and multi-ethnic), $F(1, 1099) = 6.45$, $p = .011$, and student type (regular vs. special education), $F(1, 1235) = 119.39$, $p = .000$. There are no differences between conditions for any of the other subgroups shown in Figure 3.

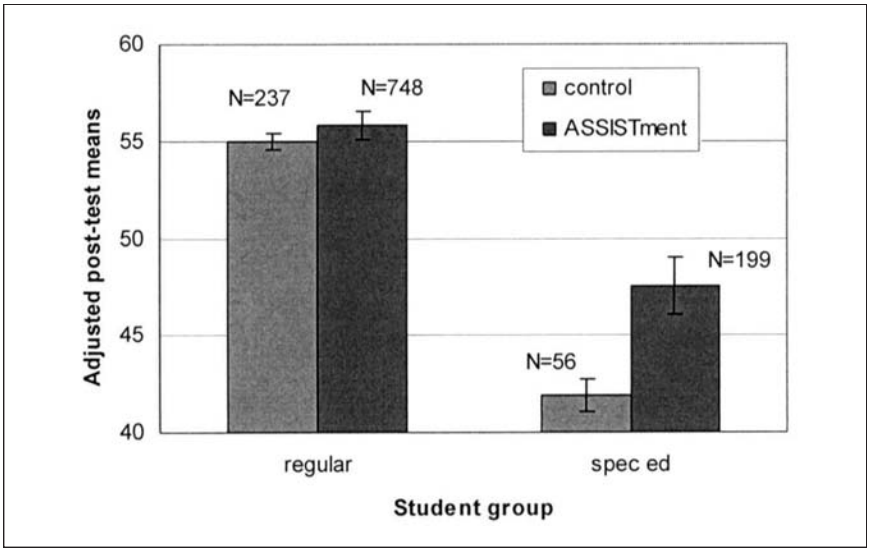


Figure 2. Adjusted post-test means for 7th grade student groups.

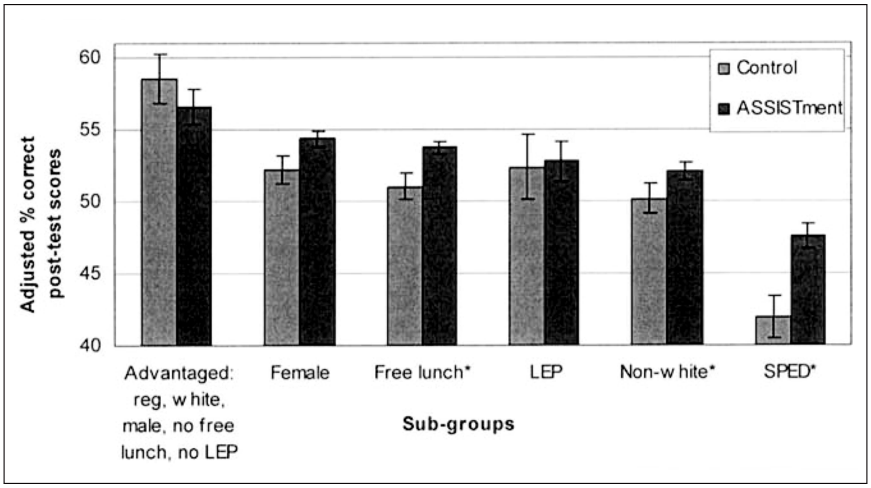


Figure 3. Adjusted percent correct means of sub-groups 2007 MCAS scores with weighted differences from least to greatest.

Usage Analysis: Does amount of use by students and teachers predict greater learning?

We investigated the effect of student usage on performance by analyzing student usage level data. As discussed above, students were labeled high usage if they completed 60 or more items, low usage students completed less than 60 items and non-usage students did not use the program. We first present results for regular students and then for special education students. On average, high usage regular students completed 109.6 items and used the ASSISTment system for 222.5 minutes while low usage regular students completed 22.5 items and used the system for 60.2 minutes. Results of a 3-level ANCOVA (high vs. low vs. no usage) show a main effect for regular student usage, $F(2, 744) = 15.05$, $p = .000$, with high usage students (adj M = 61.65, SE = .75) outperforming both the low usage students (adj M = 58.06, SE = .52) and non-usage students (adj M = 54.99, SE = .99) on the 2007 MCAS exam.

We also were interested to see whether we might find evidence that teachers used the advice given by ASSISTment to adjust their teaching which in turn would help their students learn more. If so, we might see that students with low (< 60 items) or no usage would nevertheless gain more if they are in the class of a high usage teacher than in a class of a low usage teacher. As discussed above, high-usage teachers had 25% or more of their students completing 60 or more items and low-usage teachers had less than 25% of their students completing 60 or more items. The means shown in Figure 4 are consistent with this hypothesis.

Having a high usage teacher appears to benefit low- and no-usage students, but not high usage students. To test whether this interaction is statistically reliable, we performed a 2 × 3 ANCOVA with teacher usage (high vs. low) and student usage (high vs. low vs. none) as factors and pre-test as a covariate. Consistent with the simple analysis presented above, the more a student uses ASSISTments the greater his performance on the MCAS, *F*(2, 734) = 11.52, *p* = .000. There is not a main effect of teacher usage, *F*(1, 734) = .395, *p* = .53. But, the interaction between teacher usage and student usage is significant, *F*(2, 734) = 3.67, *p* = .03. This interaction can be seen in Figure 4, where the low users and non-users of high usage teachers have higher adjusted means than their counterparts of the low usage teachers. These results are consistent with the idea that high usage teachers benefit from the system. They learn and can provide better instruction than low usage teachers.

A simple main effects analysis reveals the following. Students who were non-users performed better with high usage teachers (M = 58.27, SE = 1.86) than with low usage teachers (M = 53.93, SE = 1.22), *F*(1, 734) = 3.81, *p* = .05. For the low usage students, mean performance was also greater with high usage teachers (M = 59.22, SE = 1.18) than with low usage teachers (M = 58.10, SE =.576), but this difference was not significant, *F*(1, 734) = .734, *p* = .39. In contrast, high usage students did not appear to benefit from high usage teachers and, if anything, appeared to do worse (M = 61.31, SE = .821) than those with

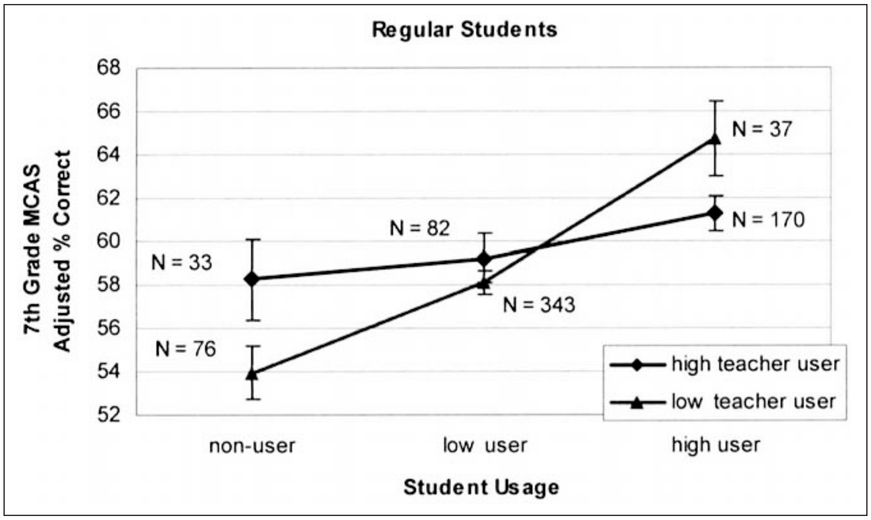


Figure 4. Adjusted percent correct means on seventh grade 2007 MCAS exam for regular students.

low usage teachers (M = 64.73, SE = 1.77). This difference is marginally significant, $F(1, 734) = 3.12$, $p = .08$.

These results are subject to selection effects and thus should be interpreted with caution. For instance, high student users may be better with a low teacher user (if the marginal effect is to be believed) not because of the teacher per se, but perhaps because to be high users in a low usage class such students must be particularly self-motivated to do more math with ASSISTments than their classmates. Similarly, it is possible that no-usage and low-usage students are not better because of their better-informed high-usage teacher, but for some other reason.

Special education students were not included in the 3 × 2 analysis because of different sample characteristics (special education students are not all associated with a regular teacher) and the small number of students in the six usage categories (e.g., there are only six non-usage special education students with a high usage teacher). Given the small numbers, it is not surprising that we do not find any significant effects when we run the 3 × 2 for special education students only ($p = .18$ for special education student usage, $p = .27$ for teacher usage, and $p = .20$ for the interaction effect).

However, we do have enough data to do main effect analyses of student and teacher usage. On average, special education student usage was 91.6 items completed and 231.2 minutes for high users and 19.9 items completed and 56.6 minutes for low users. A 3-level ANCOVA with pre-test as a covariate and special education student usage (no usage vs. low vs. high) as a between subjects factor did not reveal an effect of special education student usage, $F(2, 195) = .481$, $p = .62$ (no usage adj M = 31.5, SE = 1.3, low usage adj M = 33.2, SE = 1.0, high usage adj M = 33.3, SE = 2.8). And, a 3-level ANCOVA with teacher usage (no, low, high) as a between-subjects factor also did not reveal a significant effect ($F(2, 195) = 1.9$, $p = .16$) for teacher usage for special education students. The adjusted post-test scores of special education students were 30% with no usage teachers, 34% with low usage teachers, and 34% with high usage teachers. The difference between special education students of no usage teachers and some usage teachers (either low or high) is statistically reliable ($F(1, 196) = 3.76$, $p = .05$). It is reassuring that the adjusted post-test means of the treatment special education students who do not use ASSISTments are the same as the control groups adjusted means (30.2, SE = 1.4 to 30.1, SE = 1.6, respectively).

On further analysis, we discovered a significant association for special education students between the ASSISTment treatment and whether they are "immersed" in a regular classroom (i.e., they attend a math class with regular students as opposed to being separated with a special education teacher). In the treatment schools, 70% of special education students (140 of 199) are immersed, but only 43% of special education students (24 of 56) are immersed in the control school, a significant association ($\chi^2 = 14.39$, $p = .000$). All 125 special education students using ASSISTment are immersed and only 15 immersed students in the treatment schools are non-users. It seems that the higher level of immersion in the treatment schools has been facilitated by ASSISTment use. The technology may make immersion easier to implement. We observed with Cognitive Tutors that special education teachers were pleased to bring (and join) their special education students in the computer lab as these students could both be a part of regular class but at the same time get more individual attention, from the software and the teacher, than they can in a regular classroom (Koedinger, 2001). Another possibility, which we cannot completely exclude given the quasi experimental nature of this study, is that the treatment schools may be more proactive in their implementation of both ASSISTments and immersion and it may be that this proactive character of these schools is behind the better performance.

DISCUSSION

This study evaluates the ASSISTments system, an innovative educational technology tool that provides formative assessment during instruction. This was a large-scale, long-term study where data was collected during the entire 2006-2007 school year. While there have been a number of evaluations of educational technology use in school settings, results are not consistent and more studies are needed. The primary goal of the ASSISTment project is to address the assessment dilemma. While assessments provide useful feedback, they take time away from instruction and frequent hand grading of paper-based assessment takes teacher time away from preparation. The ASSISTment system proposes a practical solution whereby learning opportunities continue to exist during assessment. By using technology for assessment and feedback, ASSISTment enables teachers to make data-driven decisions about classroom strategy while at the same time providing students with intelligent tutoring. Hence, teacher workload is not increased and instruction time is not lost.

In this study we considered three research questions:

1. What is the effect of ASSISTments on learning after one year's usage as measured by a year-end state exam?
2. Is there a usage effect for students and/or teachers?
3. Is there any evidence teachers are using ASSISTment as a formative assessment aid?

Roediger and Karpicke (2006) reported the benefit of "testing" for enhanced learning—that is, when tests are used for instruction or learning, not for assessment. They found students had greater long-term retention and performed better when *tested* on content as compared to *studying* content. Based on these results, it is reasonable to expect ASSISTment users to demonstrate improved learning. Thus, according to Roediger and Karpicke's definition of "testing," students who use ASSISTments are being "tested" with each problem they work on and our results show the "tested" students perform better on the year-end state exam than those students who do not use ASSISTments and don't receive the benefits of being "tested." We found special education students, in particular, benefit from using ASSISTments.

In the ASSISTment system, both teachers and students receive feedback and we anticipated greater usage yielding greater performance. Our findings support such a usage effect for students, but we did not find an overall teacher usage effect. Most interesting is the significant interaction effect found between teacher and student usage and the potential impact it may have on student learning. Consistent with our hypothesis that ASSISTment provides teachers with useful formative assessment information, our results indicate low and non-using students are benefiting from what their teachers learn from observing students using the ASSISTment system or from inspecting its diagnostic assessment reports. In a review of three formative assessment systems, Militello and Heffernan (2009) found only ASSISTment is used by teachers "in a real time, cognitively diagnostic manner" (p. 5). Further study is needed to uncover what teachers need from a system like ASSISTments, and how they can best utilize the information to enhance student learning.

Effective educational technology incorporates learning theory and principles into its design (Anderson et al., 1995; Schacter & Fagnano, 1999) and should consider the cost of implementation (Yeh, 2009). We thought ASSISTments would be effective as a formative assessment tool because:

1. teacher workload is not increased;
2. the structure of the assessment data is fine-grained;
3. the results are timely;
4. it mimics a human tutor by scaffolding problems into individual skills and knowledge components; and
5. students are provided individually adapted feedback.

Although the results indicate a significant and positive learning difference, we cannot be certain that the results are caused by ASSISTments due to the nature of quasi-experimentation and potential selection bias.

This research supports the blending of technology and traditional instruction, in that ASSISTments is designed to be used in conjunction with classroom instruction, not in place of it. It highlights the value of formative testing while elucidating the problem of lost instructional time. Furthermore, it recognizes the possibility of using technology to predict student success (e.g., analyzing a school year worth of dynamic data to determine student competency). It uncovers the need to determine what metrics are required to best evaluate student achievement and what information teachers and administrators need to make informed educational decisions. As a community, we ought to have a better understanding of how teachers can use formative assessment to produce greater achievement gains.

CONCLUSIONS

Seventh grade students in three treatment schools showed a significantly higher gain in their MCAS scores from 2006 to 2007 than the students in the comparison school. Although the difference is statistically reliable, given the lack of random assignment, we cannot be certain the difference was due to the ASSISTment system and not due to other factors at the schools. We also found that regular students who were higher users of the ASSISTment system had higher MCAS scores than those students who used it less. Again, while this association between high and low users is encouraging, we cannot make firm claims about whether it was the extra time spent using ASSISTment that led to the learning gains. Nevertheless, the results are promising and suggest further research is warranted.

ACKNOWLEDGMENTS

We would like to thank the members of the ASSISTment team especially Hui Cheng for her programming expertise.

REFERENCES

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence, 150,* 183-237.

Alspaugh, J. W. (1999). The relationship between the number of students per computer and educational outcomes. *Journal of Educational Research, 21*(2), 141-150.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4,* 167-207.

Angrist, J., & Lavy, V. (2002). New evidence on classroom computers and pupil learning. *The Economic Journal, 112,* 735-765.

Becker, H., & Ravitz, J. (2001). *Computer use by teachers: Are Cuban's predictions correct?* Paper presented at the 2001 Annual Meeting of the American Educational Research Association, Seattle.

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18-20.

Bielefeldt, T. (2005). Computers and student learning: Interpreting the multivariate analysis of PISA 2000. *Journal of Research on Technology in Education, 37*(4), 339-347.

Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education, 77*(4), 7-27.

Bouck, E. C., & Flanagan, S. M. (2009). Assistive technology and mathematics: What is there and where can we go in special education. *Journal of Special Education Technology, 24*(2), 17-30.

Corbett, A., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In J. Jacko, A. Sears, M. Beaudouin-Lafon, & R. Jacob (Eds.), *Proceedings of ACM CHI'2001 Conference on Human Factors in Computing Systems,* 245-252. New York: ACM Press.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., et al. (2007). *Effectiveness of reading and mathematics software products: Findings from the First Student Cohort,* Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Edyburn, D. L. (2001). Critical issues in special education technology research: What do we know? What do we need to know? In M. Mastropieri & T. Scruggs (Eds.), *Advances in Learning and Behavioral Disabilities* (Vol. 15, pp. 95-118). New York: JAI Press.

Feng, M., Beck, J., Heffernan, N., & Koedinger, K. (2008). Can an intelligent tutoring system predict math proficiency as well as a standardized test? In R. Baker & J. Beck (Eds.), *Proceedings of the 1st International Conference on Education Data Mining*, 107-116. Montreal, Canada: Education Data Mining.

Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). Predicting state test scores better with intelligent tutoring systems: Developing metrics to measure assistance required. In M. Ikeda, K. Ashley, & T. W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 31-40). Berlin: Springer-Verlag.

Feng, M., Heffernan, N., Mani, M., & Heffernan C. (2006). Using mixed-effects modeling to compare different grain-sized skill models. In J. Beck, E. Aimeur, & T. Barnes (Eds.), *Educational data mining: Papers from the AAAI Workshop* (pp. 57-66, Technical Report WS-06-05). Menlo Park, CA: AAAI Press.

Fetler, M. (1999). High school staff characteristics and mathematics test results. *Education Policy Analysis Archives, 7*(8). Retrieved from http://epaa.asu.edu/epaa/v7n9.html

Fuchs, T., & Woessmann, L. (2004). Computers and student learning: Bivariate and multivariate evidence on the availability and use of computers at home and at school. *CESifo Working Paper 1321.* Munich: CESifo.

Guerrero, S., Walker, N., & Dugdale, S. (2004). Technology in support of middle grade mathematics: What have we learned? *Journal of Computers in Mathematics and Science Teaching, 23*(1), 5-20.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007a). *Measuring how benchmark assessments affect student achievement* (Issues & Answers Report, REL 2007–No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007b). *REL Technical Brief—A second follow-up year for "Measuring how benchmark assessments affect student achievement"* (REL Technical Brief, REL Northeast and Islands 2007–No. 002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Honey, M., Culp, K. M., & Carrigg, F. (2000). Perspectives on technology and education research: Lessons from the past and present. *Journal of Educational Computing Research, 23*(1), 5-14.

Kimmel, H., Deek, F. P., & Frazer, L. (1996). Technology and hands-on strategies for teaching science and mathematics to the special education population. *Information Technology and Disabilities Journal, 3*(2/3). Available: http://bubl.ac.uk/journals/lis/com/itad/v03n0296/

Kmitta, D., & Davis, J. (2004). Why PT3? An analysis of the impact of educational technology. *Contemporary Issues in Technology and Teacher Education, 4*(3), 323-344.

Koedinger, K. R. (2001). Cognitive tutors as modeling tool and instructional model. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 145-168). Menlo Park, CA: AAAI/MIT Press.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*(3), 239-264.

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.

Kulik, J. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say.* Arlington, VA: SRI International.

Meyer, O., & Lovett, M. (2002). Implementing a computerized tutor in a statistical reasoning course: Getting the big picture. In B. Phillips (Ed.), *Proceedings of the sixth international conference on teaching statistics: Developing a statistically literate society*. Voorburg, The Netherlands.

Militello, M., & Heffernan, N. (2009). Which one is "just right"? What educators should know about formative assessment systems. *International Journal of Educational Leadership Preparation, 4*(3), 1-8.

Naser, S. (2009). Evaluating the effectiveness of the CPP-Tutor an intelligent tutoring system for students learning to program in C++. *Journal of Applied Sciences Research, 5*(1), 109-114.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Odden, A., & Borman, G. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4-32.

Ogan, A., Walker, E., Aleven, V., & Jones, C. (2008). Toward supporting collaborative discussion in an ill-defined domain. In E. Aimeur & B. Woolf (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS 2008)* (pp. 825-827). Berlin: Springer-Verlag.

Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan, C. (2006). *Using fine-grained skill models to fit student performance with Bayesian networks.* Workshop in Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems. Taiwan.

Razzaq, L., & Heffernan, N. T. (2006). Scaffolding vs. hints in the Assistment System. In M. Ikeda, K. Ashley & T. W. Chan (Eds.), *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems* (pp. 635-644). Berlin: Springer-Verlag.

Razzaq, L., & Heffernan, N. T. (2009) To tutor or not to tutor: That is the question. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the Conference on Artificial Intelligence in Education* (pp. 457-464). Amsterdam: IOS Press.

Robinson, D. H., Levin, J. R., Thomas, G. D., & Vaughn, S. (2007). The incidence of "causal" statements in teaching-and-learning research journals. *American Educational Research Journal, 44*(2), 400-413.

Roediger, H., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention, *Psychological Science, 17*(3), 249-255.

Schacter, J., & Fagnano, C. (1999). Does computer technology improve student learning and achievement? How, when, and under what conditions? *Journal of Educational Computing Research, 20*(4), 329-343.

Schofield, J. W., Evans-Rhodes, D., & Huber, B. R. (1990). Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students. *Social Science Computer Review, 8*(1), 24-41.

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician* (American Statistical Association), *34*(4), 216-221.

Slavin, R. E. (2008). Perspectives on evidence-based research in education. What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004, September 8). National board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46). Retrieved April 1, 2009 at http://epaa.asu.edu/epaa/v12n46/

Waxman, H. C., Connell, M. L., & Gray, J. (2002). *A quantitative synthesis of recent research on the effects of teaching and learning with technology on student outcomes.* Naperville, IL: North Central Regional Educational Laboratory.

Wenglinsky, H. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics.* Princeton, NJ: Educational Testing Service Policy Information Center.

Woodward, J., & Rieth, H. (1997). A historical view of technology research in special education. *Review of Educational Research, 67*(4), 503-536.

Yeh, S. S. (2009). Class size reduction or rapid formative assessment? A comparison of cost-effectiveness. *Educational Research Review, 4,* 7-15.

Direct reprint requests to:

Dr. Kenneth Koedinger
Human Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
e-mail: koedinger@cmu.edu