Working Paper

Effectiveness of Cognitive Tutor Algebra I at Scale

John F. Pane, Beth Ann Griffin, Daniel F. McCaffrey and Rita Karam

RAND Education

WR-984-DEIES
March 2013
Prepared for Institute of Education Sciences, U.S. Department of Education

Effectiveness of Cognitive Tutor Algebra I at Scale

John F. Pane, RAND Corporation

Beth Ann Griffin, RAND Corporation

Daniel F. McCaffrey, Educational Testing Service

Rita Karam, RAND Corporation

Abstract

This article examines the effectiveness of a technology-based algebra curriculum in a wide variety of middle schools and high schools in seven states. Participating schools were matched into similar pairs and randomly assigned to either continue with the current algebra curriculum for two years or to adopt Cognitive Tutor Algebra I (CTAI), which uses a personalized, mastery-learning, blended-learning approach. Schools assigned to implement CTAI did so under conditions similar to schools that independently adopt it. Analysis of posttest outcomes on an algebra proficiency exam finds no effects in the first year of implementation, but strong evidence in support of a positive effect in the second year. The estimated effect is statistically significant for high schools but not for middle schools; in both cases, the magnitude is sufficient to improve the average student's performance by approximately eight percentile points.

Introduction

Cognitive Tutor Algebra I (CTAI), published by Carnegie Learning, Inc., is a first-year algebra course designed for students at a variety of ability and grade levels. The curriculum includes traditional textbook and workbook materials along with innovative automated tutoring software that provides self-paced individualized instruction and attempts to bring students to mastery of a topic before progressing to more advanced topics. On the basis of prior evidence of this curriculum's efficacy in some isolated contexts, we conducted a large-scale randomized controlled trial (RCT) evaluation to estimate its effectiveness when implemented in a wide variety of natural school settings, in conditions similar to those of schools that independently adopt the curriculum.

This article reports CTAI's effects on student mathematics achievement and student confidence and attitudes about mathematics. As one of the few rigorous large-scale evaluations to date of interventions that use a personalized, mastery-learning, blended-learning (partly online and partly classroom-based) approach, the study contributes to the evidence base on this increasingly popular approach to incorporating technology into instruction.

Background and context

Mathematics proficiency rates of students in the United States continue to be a concern for educators and policymakers. Although scores and proficiency rates on the National Assessment of Educational Progress have been on an upward trend since 1990, in 2011 only 35 percent of 8th-grade students performed at a level of proficient or higher (National Center for Education Statistics, 2011). Similarly, even though 8th-grade U.S. students have improved in international

comparisons between 1995 and 2011, they continue to lag well behind the top-scoring countries (Mullis et al, 2012). Moreover, low percentages of high school graduates demonstrate preparedness for higher education (ACT, 2012).

Educators are undertaking a variety of efforts to address these concerns, and many of these efforts are focused on algebra because it is considered a gateway course to higher-level mathematics and science courses. At the same time, widespread availability of computers and network connections inside and outside of schools has drawn greater attention to technology-based course materials. There is some empirical evidence suggesting that technology-based curricula can help personalize students' learning experiences and facilitate the development of mathematical skills (Koedinger et al., 2000; Ritter et al., 2007b; Schacter, 1999; Wenglinsky, 1998). However, a meta-analysis conducted by the U.S. Department of Education (2010) concluded that evidence of the effects of online learning was very weak, with few rigorous controlled studies that enabled computing effect sizes. Nonetheless, that meta-analysis estimated that interventions combining online and face-to-face instruction in a blended-learning approach produced larger positive effects than either online or face-to-face instruction alone. This meta-analysis appears to have been influential in spurring widespread development and adoption of blended learning approaches in recent years.

Stakeholders see a number of potential advantages to online or blended-learning courses (U.S. Department of Education, 2012). Such courses can provide access to high-quality instruction outside of normal school times and places. They can afford more efficient use of teaching resources, by providing teachers with frequent and detailed information about the progress and

struggles of each student, and enabling them to provide focused attention to some students while others work online. Perhaps most importantly, they are seen as having the potential to improve student achievement by providing more engaging and personalized instruction and more immediate feedback. Moreover, many educators have called for curriculum and teaching methods that emphasize active learning, build on prior knowledge, and apply mathematical concepts to real-world problems (National Council of Teachers of Mathematics, 2000).

In summary, the field is in the early stages of confirming whether these approaches produce positive effects. Where rigorous positive evidence does exist, it primarily comes from efficacy trials, which evaluate the intervention under optimal conditions in a limited context. Large-scale tests of effectiveness in diverse real-world school contexts, without any extraordinary effort to optimize implementation, like this study, have been rare.

The intervention

CTAI is a technology-based mathematics curriculum designed to promote student understanding of algebraic concepts and principles, to develop students' problem solving skills, and to enable them to master higher-order mathematical concepts (Ritter et al., 2007a). It is part of a broader set of curricula covering a number of secondary mathematics courses. In addition to textbook materials, each course includes an automated computer-based Cognitive Tutor (Anderson, et al., 1995) that provides individualized instruction to address students' specific needs. The individualization is built into the software, and is facilitated by detailed computational models of student thinking in a domain. Through the tutor, students work on challenging problems that reflect real-world situations and provide opportunities for students to progress from concrete to

abstract thinking. The company recommends that students spend two days per week of their class time using the computer-based, individualized one-on-one tutorial provided by the software while the teacher works with individual students as needed, and three days on classroom activities that are student-centered and involve group work and problem solving, guided by the teacher and the textbook but not using the software. The software is also available for students to use at other times during the school day, and outside of school from public libraries or homes, although this study did not collect data on how much this outside-of-class use occurred.

The CTAI software, as delivered, is aligned to National Council of Teachers of Mathematics (NCTM) (2000) standards, and schools or districts can customize it to align with state or local standards. It utilizes multiple representations, including diagrams, equations, and text, and concepts are often contextualized in real-world problem scenarios. Lessons address topics such as solving linear equations and systems of linear equations, mathematical modeling with linear and quadratic expressions, problem solving using proportional reasoning, and analyzing data and making predictions. During the lessons, students complete worksheets and other activities in which they record answers to questions posed in the problem scenarios, and are encouraged to engage in a variety of problem solving strategies such as breaking an unfamiliar problem down into simpler problems.

As part of implementation, teachers receive four days of training. During a three day session prior to the start of the school year teachers are introduced to the curriculum materials, tutor software and teacher tools, and given suggestions for how to implement the recommended practices. The fourth day is held during the school year, when professional development staff

observe classrooms, offer recommendations, and help teachers address any problems they are having with implementation. Teachers also receive a set of training materials, an implementation guide, and a book of resources and assessments.

Prior research on the curriculum

The most rigorous evidence of the efficacy of CTAI comes from a randomized controlled trial of the curriculum in the Moore Independent School District in Moore, Oklahoma (Morgan & Ritter, 2002). The study randomly assigned students to classrooms using either CTAI or the district's existing traditional Algebra I curriculum, and controlled for possible teacher effects by having some teachers teach both types of classes. The students using CTAI scored significantly higher on the Educational Testing Service's Algebra End-of-Course Assessment, received higher grades, and demonstrated more positive attitudes toward math than their counterparts using the traditional curriculum. The study measured achievement effect sizes of 0.29 at the student level and 0.78 at the classroom level.

Two other experiments testing the efficacy of CTAI found negative, though not significant, treatment effects for the curriculum. These were an experiment testing CTAI in five schools in Hawaii (Cabalo et al., 2007), and a two-year multi-site experiment sponsored by the U.S. Department of Education that evaluated CTAI along with nine other reading and mathematics software products (Campuzano et al., 2009). The latter study did find a significant positive effect for experienced teachers using algebra software in the study's second year, although the result was for a composite of products that included CTAI. In addition, an experiment testing the

efficacy of the related Cognitive Tutor Geometry curriculum found a significant negative effect on student achievement (Pane et al., 2010).

Design of the effectiveness trial

This study seeks to implement the CTAI curriculum under conditions similar to those that exist when schools independently decide to obtain the curriculum from Carnegie Learning. Enabling the study schools to implement the curriculum authentically was influential in choices of research design and required us to consider several challenging pragmatic issues in order to retain experimental control.

The first major challenge was how to design the study so that it minimized the disruption to normal school operations such as the assignment of students and teachers to curriculum and classes, and the ability of teachers to interact freely with colleagues in their school. To meet this challenge we chose to randomize assignment at the school level, as this would not require any modifications to the activities within schools other than the implementation of the CTAI curriculum. Moreover, randomization of schools makes it easier to prevent crossover between the treatment and control conditions.

A second consideration was the necessity for the study to administer an algebra posttest, because assessment programs vary across states and across districts or schools within states. Combining the results from many different assessments would have been problematic; moreover, not all high school students are tested in mathematics every year and the amount of algebra content on existing assessments might be small, making it more difficult to detect intervention effects.

However, administering an algebra assessment to all students in the school, regardless of grade level or the level of mathematics attained, would have seemed illogical and would have been disruptive, imposing a recruiting challenge because few schools would have been willing to participate in the study. Schoolwide testing would also have been cost-infeasible given the available funding for this study. Thus, we determined that the study would administer algebra proficiency posttests to only study participants, ruling out some analytic approaches that require measuring outcomes for all students in the school.

Third, and perhaps most challenging, was defining and holding constant the study population for the duration of the experiment, given that students who enroll in algebra come from a range of grade levels and schools typically do not have firm rules about algebra course taking. Generally, the majority of students take algebra in 9th grade, more advanced students take algebra in 8th grade or earlier, and lower achieving students might enroll in 10th grade or later. Schools also sometimes spread the curriculum over two years for lower-performing students. The definitions of these groups such as lower achieving, mainstream, or advanced/honors/gifted are not precisely specified and may drift over time, and the placement of individuals into these groups is often partly based on the discretion of families, teachers, principals, or guidance counselors. Moreover, at the time of the study, some states and districts were establishing new policies to encourage more students to enroll in algebra in 8th grade, and some of the schools participating in the study were subject to these policies. For these reasons, authentic implementation would not enable the research team to dictate exactly which students would take algebra or in what grade, and thus the study design avoided doing so, allowing schools to assign students to algebra classes according to their normal routines.

A fourth consideration was the need to allow teachers time to prepare to teach the new curriculum under a typical timeline. Meeting this goal required that randomization of schools would occur well before that start of the school year in order that treatment teachers could receive curriculum training, and schools could install the curriculum software along with any hardware necessary to support it. Moreover, the study intended to conduct the experiment for two years in each school, in order to be able to capture any improvement in implementation in the second year. This meant that second-year classes would begin more than a year after randomization. Thus, because of natural student mobility and the necessity for students to meet prerequisites, it was not considered authentic to require school officials to define, prior to randomization, the precise set of students to take algebra during the two years of the study. Instead, we sought to allow schools to retain discretion regarding when to enroll students in algebra and, as a consequence, the resulting population of students taking algebra at any particular time. Such discretion, left unchecked, would enable schools to change algebra enrollment patterns in response to their randomly assigned experimental condition, that is, after the treatment is assigned. For instance, treatment schools might view the adoption of the new algebra curriculum as an opportunity to change algebra enrollment patterns. Such enrollment changes in response to assigned experimental condition subvert the control that the researcher is trying to establish through randomization. Randomization is intended to ensure that the pre-existing characteristics of the students in the treatment and control groups are unrelated to experimental conditions, but the authentic discretion held by schools in this study could lead to systematic differences between groups.

Ultimately, we settled on an approach where, prior to randomization, schools specified *schema* for selecting the students who would participate for both years. The schema identified a set of criteria schools would apply to determine the students who would participate each year. Examples of schema include enrolling all algebra classes or the enrolling or excluding of specific classes by ability level or teacher. Thus, schools specified exact rules for selecting the types of students who would participate even though they did not specify the specific individuals. We monitored, and to the extent possible enforced, adherence to the schema throughout the study. We judged that this method achieved an appropriate balance between granting schools authentic discretion over algebra enrollment and providing sufficient control over experimental assignment to treatment, while avoiding serious feasibility issues. Moreover, we could include the schema as one of the characteristics used in blocking schools for randomization, helping to ensure that similar types of students would be participating in the treatment and control schools.

Methods

The project conducted two parallel experiments, one in middle schools and one in high schools. The study was designed and powered to examine these groups separately because the populations of students taking algebra in middle schools (grade 8 or earlier) is generally higher-achieving than the population of students taking algebra in high schools (grades 9-12) and the curriculum might have different effects with these student populations or their schools.

Study Setting: The study was conducted in 73 high schools and 74 middle schools in 52 school districts in seven states. Participating schools include urban, suburban, and rural public schools, and some Catholic Diocese parochial schools. The sites include city districts in Texas,

Connecticut, New Jersey, and Alabama, suburban districts near Detroit, MI, generally rural districts in Kentucky, and districts throughout Louisiana. Each school participated for two years. All sites participated in both the middle school and high school arms of the study except Alabama (middle school only).

Study Population: We define the study population as all students present at the time of the pretest as well as any additional students who entered the study later and remained for the posttest. Nearly 18,700 students in grades 9 through 12 participated in the high school study, with 89% of the participants in 9^{th} grade. Nearly 6,800 students in grades 6 through 8 participated in the middle school study, with more than 99% of them in 8^{th} grade.

Research Design: The study used a pair-matched cluster randomized design to assign schools to study condition. Schools within each state were matched into pairs on a number of criteria, including school-level variables and the achievement profile of participating students, as specified by the schema that schools prepared as part of enrollment in the study. Schools were randomized in the spring prior to their first year of implementation.

Schools randomized to the treatment group implemented the CTAI curriculum and those assigned to the control group continued to use their existing algebra I curriculum. Those were published by Prentice Hall, Glencoe and McDougal Littell. Assignments to treatment or control groups continued for two academic years in each school.

Data Collection: The study administered pretests and posttests from the CTB/McGraw-Hill Acuity series. About three weeks after the start of the algebra course, the study administered the Algebra Readiness Exam as a pretest. This is a 40-item multiple-choice assessment designed for students who have completed grades 6 through 11 to assess their preparation in the skills necessary for successful performance in algebra. At the end of the course, the study administered the Algebra Proficiency Exam as a posttest. This is a 32-item multiple-choice assessment designed to measure mastery of Algebra I content knowledge at the end of the course. The exams were scored using a three-parameter IRT model, and posttest scores were standardized within the population analyzed to have a mean of zero and standard deviation of 1, enabling regression coefficients to be read as standardized effect sizes.

We collected additional administrative data from district or state sources. This consisted of socio-demographic information, including race/ethnicity; gender; socio-economic status, as indicated by eligibility for the federal free or reduced-price meal program (FRL); whether the student is an English-language learner (ELL); and whether the student is in special education or gifted programs. The administrative data also included state test scores from the two years prior to enrollment in the study for each student.

At the end of the algebra course, the study also administered a 17-item student survey that measured student opinions about mathematics, computers, the algebra course they just completed, and their future schooling plans. The survey was derived from Fennema & Sherman (1976) and a similar survey that RAND previously developed and administered in an efficacy study of Cognitive Tutor Geometry (Pane et al., 2010). From this survey, we derived scales on

mathematics confidence (α=0.84[1]), utility of mathematics for the future (α=0.77), technology confidence and enjoyment (α=0.73), and opinion about the course (α=0.86). Two remaining items, not part of scales, asked opinions of the utility of computers in learning math, and future schooling plans.

Statistical Analyses: To assess success of the random assignment blocked by matched schools, we examined pretreatment group balance at the school level, using a permutation test (Efron and Tibshirani, 1993). The permutation test is an approach for calculating the probability of obtaining the observed differences between the intervention and control schools by chance that does not rely on model assumptions like a Wald test of whether the intervention and control schools are significantly different from each other. To calculate the p-values in a permutation test, we randomly re-assigned the treatment assignment indicators to schools following the randomization design (i.e., after schools were matched into pairs), to simulate the group differences with alternative realizations of the randomization. We repeated this process 10,000 times, and for each re-sampled dataset, we calculated the group differences (mean differences for continuous measures and McNemar's test statistic for binary measures). The p-value is the proportion of times the difference from the re-sampled data equals or exceeds the observed difference between the two groups from the randomization that actually occurred in our study.

Later, after the sample was defined, we assessed pre-treatment group balance at the student level, using the same hierarchical model as described below (Model 1) for measuring posttest

[1] We report Cronbach's alpha, a measure of internal consistency reliability.

outcomes, by substituting the pretest score for the outcome and determining if the treatment indicator was significantly associated with the pretest. Because this analysis revealed pre-treatment differences between the treatment and control groups, described below in "Results", we also fit covariate-adjusted models and models using prognostic scores (Hansen, 2008) that attempt to restore group balance, described below.

Including error-prone covariates in a regression model to control for systematic group differences without a correction for measurement error will generally result in biased estimates of all model parameters including the treatment indicator (Greene, 2003; Lockwood and McCaffrey, forthcoming). To avoid this potential bias given the apparent purposive selection of students for CTAI classes, we used regression calibration (Carroll et al., 2006) to correct for measurement error in the pretests. We implemented regression calibration by replacing each error-prone test score with a random variable drawn from the conditional distribution of the corresponding error-free test score given the error-prone test score and all other model covariates. The required conditional distribution was constructed assuming a linear relationship between the error-free scores and the other model covariates and used the conditional standard error of measure of each test score. To support standard error estimation, we imputed twenty values of the error-free test score for each student. An additional benefit of this approach was that we were able to impute missing pretest values for approximately 18% of the high school sample and 7% of the middle school sample who were absent for the study-administered pretest.

We also used as covariates students' achievement test scores from the two years prior to their participation in the study. Because tests differ across states, our models include interaction terms

between the prior test scores and state indicators, allowing for different relationships between prior achievement and the posttest by state. Within states, prior tests additionally differ across grades, and students have different numbers of prior tests due to whether the state administers tests in a particular grade, student absences and transfers, and students who were retained in or skipped grades. Consequently, our models allow the coefficients on prior tests to depend not only on state but also on the grade and year in which a prior test was completed and on the number of prior tests available for the student. We did not use the prior test scores available for students with uncommon test taking patterns; instead these students were treated as having missing prior test scores. Overall, about 92 percent of the available scores were used in our models.

To enable models that control for any potential imbalance remaining after covariate adjustment, we estimated students' expected posttest outcomes using prognostic scores (Hansen, 2008). Prognostic scores, which are defined as the predicted value of an outcome conditional on individuals being in the control condition, have been shown to be a useful tool for handling covariate imbalance in cases when multivariate adjustment is not sufficient (Miettinen, 1976; Hansen, 2008; Arbogast & Ray, 2011). Prognostic scores collapse the covariates of a study into a single measure which summarizes the covariates' association with potential responses (here, potential posttest scores) had each study participant been in the control condition. They are particularly useful in settings like our study where information about the outcome's relationship with covariates in the control condition is readily available. To estimate prognostic scores for each student in our study, we fit linear regression models to posttest scores in the control group sample, controlling for imputed pretest scores, prior state test scores, and student socio-demographic measures (race/ethnicity, gender, ESL, FRL, gifted status, and grade level) along

with the appropriate missingness indicators. These models were then used to calculate predicted posttest values for all students in both the treatment and control groups. Twenty prognostic scores were calculated for each student, one corresponding to each of the imputed pretest scores.

To estimate the impact of the treatment on student mathematics achievement and student confidence and attitudes about mathematics, we compare the performance of the experimental (CTAI) and control (standard algebra) groups on the posttest scores and survey items. Specifically, we fit the following hierarchical linear models (Raudenbush and Bryk, 2002) for student posttest scores. For each model, let y_{ijk1} denote the score on the outcome for the kth student in classroom j =1 to J_i for school i=1 to I where J_i denotes the number of classrooms in school i and I denotes the total number of schools in the analysis. Let y_{ijk0} denote a student's score on one imputed version of the IRT pretest, centered to have mean 0. At the first level, the student level, we model the student's score at the end of the course using four different specifications. In Model 1, we only include the overall classroom mean (μ_{ij}) and a student-specific residual error term (ε_{ijk}) at level 1:

$$y_{ijk1} = \mu_{ij} + \varepsilon_{ijk} \qquad \textit{(Model 1)}$$

where the ε_{ijk} are independent N(0, σ^2) random variables. The mean and the slope parameters are classroom specific and specified via the classroom level model: $\mu_{ij} = \gamma_{0i} + \eta_{ij0}$ with the term, η_{ij0}, representing random classroom effects that are assumed to be normal with mean zero and an unknown variance. The term γ_{0i} is specified in a school level model:

$$\gamma_{0i} = \omega_{00} + \omega_{01} T_i + \omega_{02[i]} + \zeta_{i0} \qquad \textit{(School-level model)}$$

where T_i indicates the school's treatment assignment (0 for traditional curriculum and 1 for the CTAI curriculum), $\omega_{02[i]}$ denote the fixed effects for matched pairs corresponding the pair for

school i and ξ_{i0} is normally distributed random school error terms each with mean zero and variance τ^2. The random school and classroom effects allow for the inherent clustering of outcomes at these levels of the hierarchical sample. The models were fit and parameter testing was doing using the lme command in R.

The effect of the CTAI curriculum on student achievement is tested by testing the null hypothesis that $\omega_{01} = 0$ with two-tailed test and a 0.05 level of significance. Because the pretest scores are centered to have mean zero, ω_{01} estimates the average effect of the intervention on students in the study.

Additional models use the same structure but extend the level 1 model in order to increasingly control for variables that potentially confound the effect of the intervention and to increase precision of the estimated treatment effects. In Model 2, students' imputed pretest scores (y_{ijk0}) are included as covariates, such that:

$$y_{ijk1} = \mu_{ij} + \beta y_{ijk0} + \varepsilon_{ijk.} \qquad \textit{(Model 2)}$$

In Model 3, we add more student covariates, including prior state test scores and sociodemographic measures along with appropriate missingness indicators. Namely, we have:

$$y_{ijk1} = \mu_{ij} + \beta y_{ijk0} + \alpha' x_{ijk0} + \varepsilon_{ijk.} \qquad \textit{(Model 3)}$$

where x_{ijk0} is the vector of the student variables and α is the corresponding vector of their regression coefficients.

Finally, Model 4 additionally includes four dummy indicators of prognostic score quintiles within each matched pair in the level 1 model

$$y_{ijk1} = \mu_{ij} + \beta y_{ijk0} + \alpha' x_{ijk0} + \sum_{ijm} \gamma_{mi} p_{mijk} + \varepsilon_{ijk.} \qquad \textit{(Model 4)}$$

where p_{mijk} denotes the indicator for whether the student fell into the m^{th} quintile ($m = 1,\ldots,4$) and γ_{mi} denotes the regression coefficient for each pair by quintile dummy.

Finally, in order to examine treatment effects by student ability level, we supplemented Model 4 to include interaction terms between the prognostic score quintiles and the treatment indicator.

Our analysis plan specified that we would analyze results separately by cohort (the first or second year of implementation in the school), to allow for the possibility that implementation might be better the second year due to teachers gaining experience with the curriculum. In order to assess the performance of our prognostic scores at creating balance between the treatment and control students, we examined how the regression coefficient for the treatment indicator changed between models that did and did not additionally control for the variables used to estimate the prognostic scores (imputed pretests, sociodemographics, and prior state test scores).

Additional analyses examined whether treatment was associated with a number of secondary outcomes measuring student attitudes and confidence towards mathematics and technology. Specifically, models like Model 3 for were fit to each of the following survey scales: mathematics confidence, utility of mathematics for the future, technology confidence and enjoyment, utility of computers in learning math, opinion about the course, and future schooling plans.

Finally, additional analyses explored whether there was any evidence of interactions with treatment by site or, for cohort 2 only, whether the teacher was new to the study or was in the study the previous year. The latter analysis explores whether implementation of CTAI might be more effective the second year teachers use it, by comparing with teachers in the control group who were also present both years.

Results

Table 1 summarizes balance on school-level variables between the treatment and control groups. Randomization did not achieve perfect balance on all variables. In the high school study, the overall student bodies of schools assigned to treatment had a greater proportion of black students and fewer white students; additionally, the percentage of students classified as proficient was lower the previous three years in mathematics, and two of the previous three years in reading, although the proficiency differences were not significant. In the middle school study, the overall student bodies of schools assigned to treatment had significantly fewer students classified as proficient in reading in 2004 and both reading and mathematics in 2005.

Table 2 summarizes information about student participants in the high school sample. The final sample included 13,445 students after approximately 28% attrition from both treatment and control groups, with the treatment group scoring 0.116 standard deviation units lower than the control group on the pretest (p=0.19). Similarly, Table 3 summarizes information about student participants in the middle school sample, which included 5,940 students after attrition of approximately 11% in the treatment group and 14% in the control group. In this study, the

treatment group scored 0.328 standard deviation units lower than the control group on the pretest ($p<.01$). Similar group differences are also apparent on students' prior state test scores (not shown in tables).

Table 4 summarizes the results for the high school study. Models consistently estimated negative treatment effects for cohort 1, ranging from 0.10 to 0.19 standard deviation units and not significant. In contrast, models for cohort 2 consistently estimate positive treatment effects, ranging from 0.14 to 0.21 standard deviation units; results for Models 2 through 4 were all below the 0.05 level of significance.

Similarly, Table 5 summarizes the results for the middle school study. Here again, models estimated treatment effects for cohort 1 that are not significant. These estimates are near zero in all the models that adjust for pretest scores. For cohort 2, the unadjusted estimate is negative, and the estimates become positive with covariate adjustment. Although these estimates for middle school cohort 2 are not significant, the estimated treatment effects are similar in magnitude to those found in the high school study.

Figures 1 and 2 show the estimated treatment effects for each of the prognostic score quintiles from our regression models that included interaction terms between the quintile indicators and treatment. For cohort 2 in both the middle schools and high schools, effects are stable across the prognostic score quintiles and the interactions between the quintiles and the treatment indicator are not significant (joint Wald test p-values = 0.51 and 1.00, respectively). Conversely, there was highly significant evidence of moderation by the prognostic score quintiles in middle school

cohort 1 (joint Wald test p-value <0.001; see Figure 1) which indicated that there were potentially moderately large positive treatment effects in the lowest quintile and small negative effects of treatment in the highest two quintiles. Nonetheless, it is important to note that all of the quintile-specific treatment effect estimates for middle school cohort 1 had confidence intervals that contained 0. Similar negative effects of treatment in the highest two quintiles were found for high school cohort 1 students (see Figure 2); however, the joint Wald test for this cohort was not significant (p=0.30).

Analyses involving secondary outcomes on student attitudes and confidence in mathematics and technology only revealed one significant relationship. Across all cohorts, students in the treatment condition reported significantly higher mean scores on the item that asked about the utility of computers in learning math.

Finally, we found no significant interactions between the treatment and treatment sites nor treatment and the indicator of whether or not a teacher was previously in the study.

Discussion

It is necessary to be cautious in interpreting these results because mean student pretest scores were lower in the treatment group than the control group. This pretreatment difference, at -0.12 standard deviation units across the two cohorts in the high school study, is within the limit of -0.25 set by the What Works Clearinghouse (WWC) (U.S. Department of Education, 2011) for acceptable pretreatment differences. However, the difference was much larger in the middle

school study. At -0.33 standard deviation units, it exceeds the WWC guideline and raises concern about the validity of the middle school experiment.

Our assessment of post-randomization balance on school-level variables found imbalance on some variables suggesting the treatment groups in both studies may have been disadvantaged through the luck of random assignment. Such differences are not unexpected when randomizing with such small sample sizes (at the school level), however this cannot fully explain the magnitude of pretreatment differences on student pretests in the middle school study. Schools may have differentially exercised discretion over the population of students taking algebra and participating in the study depending on whether they were assigned to the treatment or control groups. As discussed above, the study was designed to allow for this discretion within limits; schools specified schema to describe the population of students in the study, but not the precise set of students. We monitored adherence to these schema and did not detect serious non-compliance. Nonetheless, schools may have worked within the constraints of the schema to shape the population of students participating in the study after becoming aware of treatment assignment. This may have occurred if, for example, treatment schools believed the new curriculum would have positive effects, causing them to encourage more low-performing students to enroll in CTAI classes. Conversely, they may have believed the curriculum was better for lower-performing students and thus might have enrolled higher-performing students in algebra classes that were not using CTAI and thus not part of the study. We explored these or other potential explanations with school officials and they uniformly expressed no awareness of such deliberate inclusion or exclusion of students.

For whatever reason, in both years CTAI students underperformed control group students on the pretest by a modest and non-significant amount in the high school study and a greater, significant amount in the middle school study. This could potentially raise concerns of bias in the impact estimates for both studies even though pretreatment group differences in the high school study are well within levels generally considered acceptable (e.g., U.S. Department of Education, 2011). Models 2 through 4 attempt to address the bias and improve precision through covariate and prognostic score adjustments.

The first thing to note is that for both studies and both years, the treatment effect estimates from Model 1, which has no covariate adjustments, are substantially lower than for the models with adjustments, as would be expected if pretreatment differences were biasing the estimate. Second, in year 2 of the high school study, all three of the adjusted estimates are positive, of similar magnitude, and significant at the 0.05 level. Results from Models 2 through 4 for year 2 of the middle school study are also all positive and of similar magnitude to each other and to the effects for the year 2 high school cohort. However, the middle school results are not significant because of the small sample size; the middle school cohort of students is only about one-third the size of the high school cohort.

Together the results provide strong evidence in support of a positive effect for CTAI in year 2: when we control for the selection of lower achieving students in the treatment group, apparent on the basis of their prior achievement, we find positive effects that are robust to model specification and replicate in both samples. To help interpret the importance of these second-year effects, consider a student who would score at the 50th percentile of the posttest distribution if

they were in the control group; an effect size of 0.20 is equivalent to having that student score at the 58th percentile if they were in the treatment group.

Examination of prognostic score quintiles suggests that the positive effect for high school cohort 2 was relatively uniform for students of all ability levels. Estimates for all five quintiles were in the range of 0.20 to 0.23. For middle score cohort 2, the effect estimates are 0.23 for the lowest-performing quintile, and decrease for students of greater ability to 0.12 for the highest quintile. However, results from the two studies are not directly comparable because, as is evident in mean pretest scores in Tables 2 and 3, middle school students in the study are much higher achieving than their high school counterparts.

Finally, although the positive results for cohort 2 varied from site to site, lack of statistical significance does not support any attempt to interpret this variation.

In contrast to the positive results for the second year of implementation, treatment effect estimates are not significant the first year. The estimates are negative in the high school study and near zero in the middle school study. Examination of the effects by prognostic score quintiles suggest that the poor results in the first year may have been concentrated among higher-performing students in both middle schools and high schools.

It is quite interesting that significant positive effects emerge in high school cohort 2 after the negative (though not significant) results the prior year. One potential explanation is that teachers improved their implementation of CTAI after a year of experience using it. We explored this by

dividing the cohort 2 teachers into two groups by whether they were also in the study the prior year, and testing if treatment effects were more positive among those present the prior year. This analysis produced mixed, non-significant results that thus do not lend support to this hypothesis. This question can be further informed by examining implementation data the study collected through teacher surveys and site visits.

Karam et al. (submitted) explores a variety of implementation questions, including the extent to which teachers report implementing the curriculum as specified by the CTAI developers, the relationship between implementation variables and outcomes, and how implementation changed over time. That article finds that treatment and control group teachers reported greater contrast in instructional practices the first year than they did the second year. Relative to control group teachers' reports, treatment group teachers reported less implementation of traditional practices such as lecturing with students taking notes and greater implementation of more progressive practices such as facilitating student work or assigning students to work in groups and give presentations. These differences are aligned with the practices recommended by Carnegie Learning for implementing CTAI, suggesting they were induced by the curriculum. The second year, treatment group teachers also reported practices aligned with the recommendations, although the contrast with the control group decreased. This suggests the teachers reverted somewhat back toward the more traditional practices reported by control group teachers. This result is consistent with adaptation that may have been responsive to poor results the prior year.

We found no meaningful effects of CTAI on secondary outcomes such as student attitudes and confidence in any of the cohorts.

Daugherty et al. (2012) examines the costs of CTAI relative to the curricula in use in the control group schools in this study. While that analysis finds that CTAI is substantially more expensive, the cost must be weighed alongside the benefits reported herein. Educators may judge that the positive effects are large enough to warrant the additional cost.

Conclusions

This large-scale effectiveness trial of Cognitive Tutor Algebra I finds a significant positive effect in high schools in the second year of implementation, relative to similar schools that continued to implement a variety of existing textbook-based algebra curricula. The effect size of approximately 0.20 is educationally meaningful – equivalent to moving an algebra I student from the 50^{th} to the 58^{th} percentile. This positive result is important for educators and policy makers who are seeking interventions to improve algebra I achievement, and is particularly notable because it was obtained in an effectiveness trial, where broad variety of schools implemented the curriculum without extraordinary support. The results may also be of broader potential interest because this curriculum uses technology to enable a personalized, blended-learning approach. As one of the first large-scale effectiveness trials of this type of intervention, the results help to inform researchers and practitioners whether this may be a productive way to employ technology to improve student achievement in mathematics or other subjects.

Acknowledgements

The authors would like to express their gratitude to the many individuals who helped to make this study possible. Current and former RAND researchers Andrea Phillips, Abby Robyn, Nidhi

Kalra, Regan Main, and J.R. Lockwood made extraordinary contributions, as did administrative assistants Crystal Baksis and Melanie Rote. Also playing important research or support roles were Scott Ashwood, Diane Bronowicz, Richard Bowman, Jaime Connors, Lindsay Daugherty, Maria Edelen, Amy Haas, Ann Haas, Laura Hamilton, Mark Hanson, Gina Ikemoto, Brian McInnis, Scott Naftel, Lawrence Painter, Louis Ramirez, Mary Ellen Slaughter, Anisah Waite, and Deborah Wesley. JoAnn Arcement, Michelle Auster, John Dilegghio, Barbara Dilegghio, Gayle Glusman, Kathy Hughes, Gary Kubina, and Diana Perez provided essential coordination and support at the seven research sites. We offer our special thanks to the state and district administrators who helped with recruiting, support, and data access, including: David Akridge, Gary Asmus, Luellen Bledsoe, Rebecca Feola, Debbie Ferry, Michael Henderson, Kevin Hill, Monica Kendall, Gayle Kirwan, Brian McCarty, Karen Mohr, Ricardo Rosa, Marianne Srock, Liz Storey, and Kelly Trlica. We thank the teachers and principals in the participating schools, without whose participation this research would not have been possible. We also thank Carnegie Learning personnel who supported the study, including Sandy Bartle, Steve Fancsali, Joseph Goins, Christy McGuire, Tristan Nixon, Steve Ritter, and Sean Sykes, as well as the company's field support and training staff. Finally, we thank Paco Martorell, who reviewed and provided helpful feedback on an earlier draft of this article. The first author assumes full responsibility for any omissions from this acknowledgement, and offers his apologies and appreciation to those persons as well. The research reported in this article was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A070185 to the RAND Corporation. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

ACT. (2012). The Reality of College Readiness: National. Iowa City, IA: ACT, Inc.

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

Arbogast, P.G & Ray, W.A. (2011). Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders. *American Journal of Epidemiology*. 174(5): 613-620.

Cabalo, J. V., Jaciw, A., & Vu, M.-T. (2007). Comparative effectiveness of Carnegie Learning's Cognitive Tutor Algebra I curriculum: A report of a randomized experiment in the Maui School District. Palo Alto, CA: Empirical Education, Inc.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts* (NCEE 2009-4041). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (Second ed.). London: Chapman and Hall.

Daugherty, L., Phillips, A., Pane, J. F., & Karam, R. (2012). Analysis of Costs in an Algebra I Curriculum Effectiveness Study: RAND Corporation.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Fennema, E., & Sherman, J. A. (1976). Fennema-Sherman Mathematics Attitude Scales: Instruments Designed to Measure Attitudes Toward the Learning of Mathematics by Males and Females. JSAS Catalog of Selected Documents of Psychology, 6(31).

Greene, W. H. (2003). *Econometric Analysis* (Fifth ed.). Upper Saddle River, NJ: Prentice Hall.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95 (2) 481-88.

Karam, R., Pane, J. F., Griffin, B. A., & Slaughter, M. E. (submitted). Evaluating Cognitive Tutor Algebra 1 Curricula At Scale: Focus on Implementation.

Koedinger, K. R., Corbett, A. T., Ritter, S., & Shapiro, L. J. (2000). *Carnegie Learning's Cognitive Tutor: Summary research results*. Pittsburgh, PA: Carnegie Learning.

Lockwood, J. R., & McCaffrey, D. F. (forthcoming). Should nonlinear functions of test scores be used as covariates in a regression model? In R. Lissetz (Ed.), Informing the Practice of Teaching Using Formative and Interim Assessment: A Systems Approach. Charlotte, NC: Information Age Publishing, Inc.

Miettinen, O.S. (1976). Stratification by a multivariate confounder score. *American Journal of Epidemiology*. 104(6):609–620.

Morgan, P., & Ritter, S. (2002). An experimental study of the effects of Cognitive Tutor Algebra I on student knowledge and attitude. Pittsburgh, PA: Carnegie Learning, Inc.

Mullis, I. V. S., Martin, M. O., Foy, P., & Alka, A. (2012). TIMSS 2011 International Results in Mathematics. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

National Center for Education Statistics. (2011). The Nation's Report Card: Mathematics 2011(NCES 2012–458). Washington, D.C.: Institute of Education Sciences, U.S. Department of Education.

National Council of Teachers of Mathematics. (2000). Principles and Standards for School Mathematics. Reston, VA.

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

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models. Thousand Oaks, CA: Sage Publications.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). The Cognitive Tutor: Applied research in mathematics education. *Psychonomics Bulletin & Review*, 14(2), 249-255.

Ritter, S., Kulikowich, J., Lei, P., McGuire, C. L., & Morgan, P. (2007b). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. *Supporting Learning Flow through Integrative Technologies, 162*, 13-20.

Sarkis, H. (2004). *Cognitive Tutor Algebra I Program Evaluation*. Miami-Dade: Miami-Dade County Public Schools, Reliability Group.

Schacter, J. (1999). The impact of education technology on student achievement: What the most current research has to say. Santa Monica, CA: Milken Exchange on Educational Technology, Milken Family Foundation.

U.S. Department of Education. (2010). Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Washington, D.C.: Office of Planning, Evaluation, and Policy Development.

U.S. Department of Education. (2011). What Works Clearinghouse: Procedures and Standards Handbook (Version 2.1): Institute of Education Sciences.

U.S. Department of Education. (2012). Understanding the Implications of Online Learning for Educational Productivity. Washington, D.C.: Office of Educational Technology.

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

Tables and Figures

Table 1: School level balance after randomization

		High school study		*Middle school study*	
		Control group	Treatment group	Control group	Treatment group
	Number of schools	37	36	37	37
	School size	852	825	836	780
Race/ethnicity	American Indian / Alaskan Native	0.00	0.00	0.00	0.00
	Asian / Pacific Islander	0.02	0.01	0.03	0.02
	Black non-Hispanic	0.34	0.41 *	0.30	0.30
	Hispanic	0.12	0.13	0.24	0.27
	White non-Hispanic	0.52	0.45 *	0.44	0.41
Student demographics	Eligible for free or reduced-price lunch	0.38	0.44	0.63	0.69
	Classified as English-learner	0.09	0.09	0.10	0.10
Proficiency rates on state assessments	Reading 2004	52.37	49.98	69.50	66.18 *
	Reading 2005	55.00	51.27	74.98	71.64 *
	Reading 2006	53.82	54.97	75.46	74.71
	Mathematics 2004	47.03	42.39	51.02	51.53
	Mathematics 2005	47.40	43.36	54.12	49.60 *
	Mathematics 2006	44.32	42.34	58.11	57.48
Intended study population as defined by schema	All classes	0.55	0.61	0.68	0.65
	Identified teachers only	0.42	0.39	0.39	0.42
	Includes low-achieving students	0.61	0.55	0.03	0.00
	Includes average-achieving students	0.52	0.48	0.16	0.19
	Includes high-achieving students	0.33	0.35	0.52	0.48
	Number of classes projected to participate	7.04	6.56	2.30	2.03
	Number of students projected to participate	161	153	56	48

Note: * indicates $p<.05$ for group differences, calculated with a permutation test (see text; all $p>.01$).

Table 2: High School Study Attrition and Group Balance

		Treatment group		Control group		Group difference^	p-value
		N	*Pretest mean*	*N*	*Pretest mean*		
Cohort 1	Eligible sample	4,541	-0.468	5,014	-0.347	-0.194	0.033
	Attrition	1,330	-0.673	1,723	-0.485	-0.155	0.091
	Final sample	3,211	-0.365	3,291	-0.258	-0.144	0.187
	Attrition rate		29.3%		34.4%		
Cohort 2	Eligible sample	3,990	-0.390	5,146	-0.359	-0.099	0.284
	Attrition	1,058	-0.591	1,135	-0.619	-0.067	0.529
	Final sample	2,932	-0.302	4,011	-0.270	-0.111	0.276
	Attrition rate		26.5%		22.1%		
Both cohorts	Eligible sample	8,531	-0.432	10,160	-0.353	-0.139	0.082
	Attrition	2,388	-0.637	2,858	-0.538	-0.106	0.205
	Final sample	6,143	-0.335	7,302	-0.265	-0.116	0.188
	Attrition rate		28.0%		28.1%		

Notes: ^Model-adjusted standardized mean difference in pretest scores between treatment and control groups (negative indicates treatment scored lower than control). Eligible sample is defined as students present at pretest or entering the study after pretest. Attrition is defined as the portion of the eligible sample that did not take the posttest.

Table 3: Middle School Study Attrition and Group Balance

		Treatment group		Control group		Group difference^	p-value
		N	*Pretest mean*	*N*	*Pretest mean*		
Cohort 1	Eligible sample	1,681	0.265	1,743	0.654	-0.296	0.021
	Attrition	169	-0.067	212	0.318	-0.295	0.135
	Final sample	1,512	0.306	1,531	0.706	-0.312	0.016
	Attrition rate		10.1%		12.2%		
Cohort 2	Eligible sample	1,534	0.400	1,828	0.738	-0.422	0.003
	Attrition	170	0.015	295	0.733	-0.392	0.012
	Final sample	1,364	0.449	1,533	0.739	-0.347	0.016
	Attrition rate		11.1%		16.1%		
Both cohorts	Eligible sample	3,215	0.331	3,571	0.698	-0.366	0.003
	Attrition	339	-0.026	507	0.560	-0.385	0.007
	Final sample	2,876	0.376	3,064	0.722	-0.328	0.007
	Attrition rate		10.5%		14.2%		

Notes: ^Model-adjusted standardized mean difference in pretest scores between treatment and control groups (negative indicates treatment scored lower than control). Eligible sample is defined as students present at pretest or entering the study after pretest. Attrition is defined as the portion of the eligible sample that did not take the posttest.

Table 4: High School Study Treatment Effect Estimates

Model	*Cohort 1*				*Cohort 2*			
	Estimate	*Std. Error*	*t-value*	*p-value*	*Estimate*	*Std. Error*	*t-value*	*p-value*
1	-0.19	0.12	-1.68	0.10	0.14	0.12	1.20	0.24
2	-0.12	0.10	-1.20	0.24	0.19	0.09	2.05	0.05^
3	-0.10	0.10	-0.97	0.34	0.22	0.09	2.33	0.03
4	-0.10	0.10	-1.02	0.31	0.21	0.10	2.23	0.03

Notes: ^ value is less than 0.05 before rounding. Model 1 estimates group differences without any covariates; Model 2 includes regression-calibrated pretest scores; Model 3 also includes additional student covariates; and Model 4 includes all covariates as well as prognostic score quintiles.

Table 5: Middle School Study Treatment Effect Estimates

Model	*Cohort 1*				*Cohort 2*			
	Estimate	*Std. Error*	*t-value*	*p-value*	*Estimate*	*Std. Error*	*t-value*	*p-value*
1	-0.20	0.15	-1.34	0.19	-0.07	0.17	-0.41	0.69
2	0.00	0.10	0.01	0.99	0.17	0.13	1.30	0.21
3	-0.03	0.11	-0.24	0.81	0.19	0.13	1.44	0.16
4	0.01	0.11	0.11	0.91	0.19	0.14	1.38	0.17

Note: Model 1 estimates group differences without any covariates; Model 2 includes regression-calibrated pretest scores; Model 3 also includes additional student covariates; and Model 4 includes all covariates as well as prognostic score quintiles.

Figure 1. Estimated treatment effects within the five prognostic score quintiles (1=lowest and 5=highest) for the Middle School study cohorts. MS = Middle School

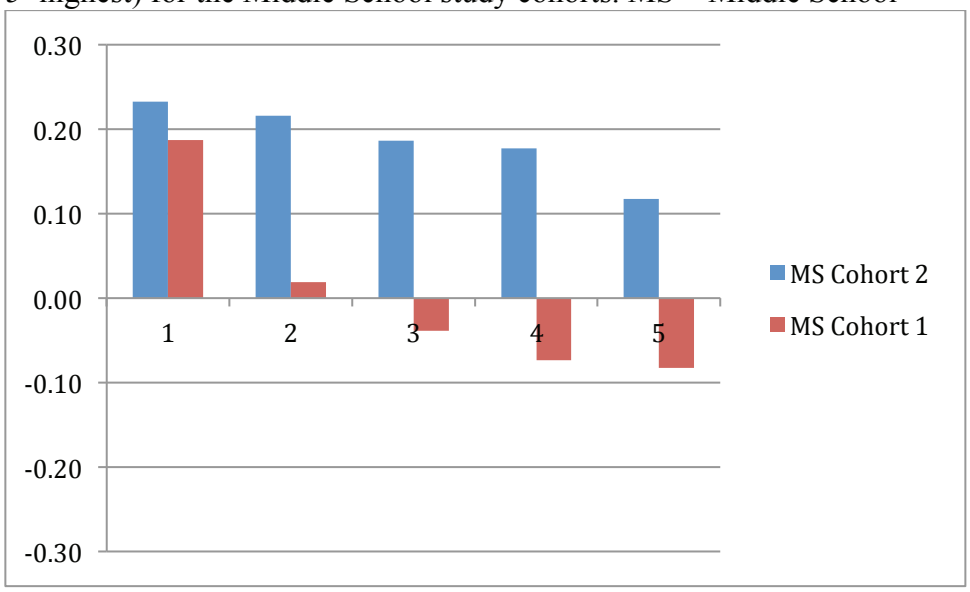

Figure 2. Estimated treatment effects within the five prognostic score quintiles (1=lowest and 5=highest) for the High School study cohorts. HS = High School