



RESEARCH REPORT

Comparative Effectiveness of Carnegie Learning's *Cognitive Tutor* Algebra I Curriculum:

A Report of a Randomized Experiment
in the Maui School District

Andrew P. Jaciw
Jessica Villaruz Cabalo
Minh-Thien Vu
Empirical Education Inc.

May 2007

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734

Acknowledgements

We are grateful to the people at the Maui Department of Education and Maui Community College for their interest and assistance in conducting this research. The research was funded by a grant (#R305E040031) to Empirical Education Inc. from the US Department of Education. The purpose of this grant is to improve our ability to conduct small scale experiments to assist local decision-makers. The US Department of Education is not responsible for the content of this report.

This report was presented as a paper discussion at the Annual American Education Research Association conference in April 2007 (Division C-Learning and Instruction / Section 7: Technology Research).

About Empirical Education Inc.

Empirical Education Inc. was founded to help K–12 school districts, publishers, and the educational R&D community assess new or proposed instructional programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

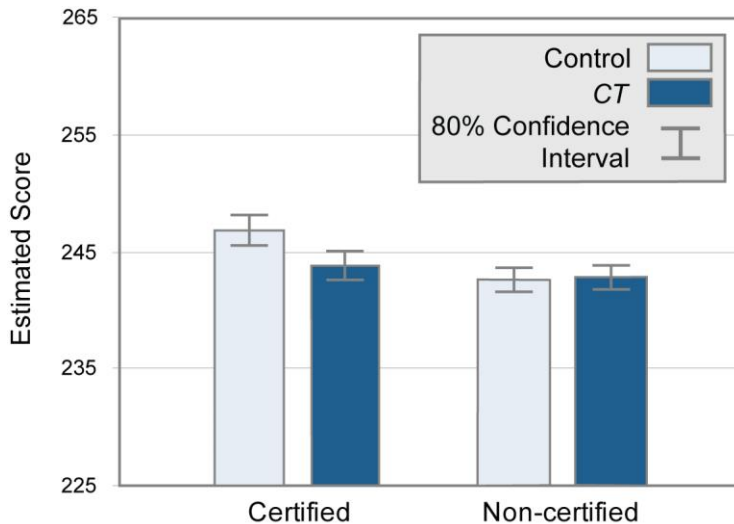
© 2007 by Empirical Education Inc. All rights reserved.

Executive Summary

Introduction. The Maui Hawaii Educational consortium (the Maui School District and Maui Community College) sought scientifically based evidence for the effectiveness of the *Cognitive Tutor (CT) Algebra I Curriculum* to inform adoption decisions. Decision makers were particularly interested in whether the use of the *CT* program affects achievement of their students and is a useful tool for teachers. The participating teachers used the program in their randomly selected *CT* classes for six months during the 2005-2006 school year while teaching their other classes with their standard materials. The experiment involved nine teachers, 22 classes, and 541 students.

Findings. Our experiment for the most part was unable to detect any impact of *Cognitive Tutor* on student achievement. Of the four sub-strands, the analysis found negative results for two of them: Quadratic Equations and Algebraic Operations. The results must be interpreted in the context of the particular resources for the implementation of the program on Maui, which were not favorable. Nonetheless, the size of the negative impact of the two sub-strands was considerable. The effect sizes were -0.33 and -0.25. These are comparable in magnitude to the positive effects found in other studies of this product.

One explanation for the negative outcome is shown in the analysis of teacher certification. While the certified teachers out performed the uncertified teachers with their control class, they performed the same as the uncertified teachers with their *CT* classes. This is shown in the following figure of the moderating effect of teacher certification on NWEA overall math outcomes.



Perhaps, because both groups were equally unfamiliar with *CT* their results were the same. With the existing program, however, the more experienced teachers were far more familiar with the methods and content of the program than the uncertified teachers (who were equally unfamiliar with both programs). It is possible that over time, with more familiarity with *CT*, both groups may improve. Tracking these teachers into their second year with *CT* may provide some indication as to whether the experienced teachers can overcome the initial deficit.

While teachers gave a generally positive view of *CT*, they reported difficulties with implementation—particularly access to computers. However, our exploratory analysis of student usage data did not support lack of computer access as an explanation for the poor results. Aspects of the implementation other than computer lab time may have influenced the results—the late delivery of the product and lack of familiarity, for example.

We conducted an additional exploratory analysis of the data provided by the *CT* intervention on units completed. This indicated that there is an overall positive relationship between the number of units completed and student outcomes on the NWEA test. This also suggested that the NWEA test was a good measure of what was covered in the *Cognitive Tutor* curriculum. We must caution, however that the nature of the outcome measure (the paper version of the NWEA end-of-course Algebra I test) makes us less confident in our conclusions about lower achieving students. First, it appears that lower achieving students were less likely to take or complete the posttest. Second, the test itself is not sensitive to students scoring at the low end and appears to distort their scores upward due to a floor effect in the paper-version of the NWEA test.

If the implementation and start-up problems were part of the reason for the poor performance, the results suggest that school districts should avoid undertaking this program without adequate resources

and preparation time. Our recommendation is to continue the experiment into an additional year, providing adequate resources for implementation, before undertaking wider deployment.

This small study illustrates a general caution in interpreting findings from isolated experiments. Previous research has indicated that the *Cognitive Tutor* program is effective (Morgan & Ritter, 2002). This experiment demonstrates the importance of conducting multiple replication trials of any application in varying contexts and conditions.

Setting. The district and community college are located in a mixed suburban and rural community. The community is ethnically diverse and the average ethnic breakdown for the participating schools includes approximately 32% Filipino, 28% Part-Hawaiian, 11% White, 8% Japanese, 5% Hawaiian, 3% Hispanic, and 14% Other. The district's existing math program consisted of a variety of Algebra I textbooks from several publishers.

The intervention. *Cognitive Tutor* is a full Algebra I curriculum, combining software-based, individualized computer lessons with collaborative, real-world problem-solving activities. The publisher, Carnegie Learning, describes *CT* as a research-based approach to improving student understanding of mathematical concepts that includes a simple and straightforward design, research-based pedagogy, multiple representations of word problems, just-in-time feedback, a skillometer, and a blended curriculum of 40% computer lab and 60% classroom activities. *CT* also uses verbal, numerical, algebraic, and graphical representations as different modalities to ease problem-solving. A Carnegie Learning consultant led all eight *CT* teachers in three days of training, after which they could use of the materials as best suited the needs of their *CT* classes and students.

Research design. We conducted a randomized experiment which compared outcomes for groups of students taught using the *CT* program and students taught using the regular Algebra I curriculum. We randomly assigned the Algebra I classes of each participating teacher to the *CT* or control group. Randomization of participating teachers' classes was stratified according to class size and achievement level. Using the tests obtained from the Northwest Evaluation Association, we collected pretest measures of general math and posttest measures of algebra achievement. The posttest provided sub-straunds for Algebra Operations, Linear Equations, Quadratic Equations, and Problem-Solving. We also gathered demographic information on students and teachers. We collected three types of qualitative data: 1) classroom observations, to document implementation; 2) phone interviews, to gauge teachers' attitudes and opinions about the *CT* program as well as the kinds of challenges and difficulties encountered; and 3) web-based surveys, to learn about factors that may influence the results. In addition, as a feature of the *CT* Teacher's Toolkit, we were able to gather information on student time spent on the software and the number of problems and units each student completed.

Analysis. The basic question for the statistical analysis was whether, following the intervention, students in the *CT* classrooms had higher math scores than those in control classrooms. We used multi-level models that account for the clustering of students in classes, providing a more accurate, and often more conservative, assessment of the confidence we should have in the findings. To increase the precision of our estimate, we included students' pretest scores in the analysis and examined the interaction between this covariate and the experimental condition. Finally, to better understand unexpected results, we use other data such as that from the *CT* Teacher's Toolkit in exploratory analyses.

Comparative Effectiveness of Carnegie Learning's *Cognitive Tutor* Algebra I Curriculum:

A Report of a Randomized Experiment in the Maui School District

Table of Contents

INTRODUCTION	1
METHODS	1
RESEARCH DESIGN	1
MATERIALS	2
SITE DESCRIPTIONS	2
Maui County	2
Maui District Schools	2
Existing Math Program.....	2
SAMPLE AND RANDOMIZATION	3
DATA COLLECTION	3
Observations	3
Interviews	3
<i>Table 1. Survey Response Rates</i>	4
Surveys	4
Carnegie Learning Data.....	4
Achievement Test Scores	4
STATISTICAL ANALYSIS AND REPORTING	5
RESULTS	6
FORMATION OF THE EXPERIMENTAL GROUPS	6
<i>Table 2. Distribution of the CT and Control Groups by Schools, Teachers, Grades, and Counts of Students</i>	6
Post Randomization Composition of the Experimental Groups.....	6
Ethnic Composition of Student Population	6
<i>Table 3. Ethnicity for CT and Control Groups</i>	6
Teacher Certification.....	7
<i>Table 4. Teacher Certification: Comparison among Students</i>	7
Achievement Pretests.....	7
<i>Table 5. Independent t Test of the Difference between Students in CT and Control Groups for the NWEA Math Goals Survey 6+ ALT Pretest</i>	7
Attrition	7
<i>Table 6. Counts of Students Missing Test Score Data</i>	8
<i>Table 7. Independent t Test of the Difference in Pretest Scores between Students Who Had Pretest Scores Only and Students with Both Pre- and Posttest Scores</i>	8
IMPLEMENTATION RESULTS	8
Classroom Observations.....	9
Teacher Responses	9
Positive Experiences	9
Challenges	10
IMPACT OF COGNITIVE TUTOR	11

Overall Score on the NWEA End-of-Course Algebra I Test	11
<i>Table 8. Overview of Sample and Impact of CT on the Overall Score: NWEA End-of-Course Algebra I Test</i>	<i>11</i>
<i>Figure 1. Impact on the Overall Score of the NWEA Algebra I Test: Unadjusted Pre- and Posttest Means for Control and CT (Left); Adjusted Means for Control and CT (Right)</i>	<i>12</i>
<i>Table 9. The Impact of CT on Student Performance on the Overall Score of the NWEA End-of-Course Algebra I Test.....</i>	<i>13</i>
<i>Figure 2. Overall Score—Scatterplot for All the Students</i>	<i>14</i>
Quadratic Equations	14
<i>Table 10. Overview of Sample and Impact of CT on the Quadratic Equations Sub-strand.....</i>	<i>15</i>
<i>Figure 3. Impact on the Quadratic Equations Sub-strand: Unadjusted Pre- and Posttest Means for Control and CT Left); Adjusted Means for Control and CT (Right)</i>	<i>15</i>
<i>Table 11. The Impact of CT on Student Performance on the Quadratic Equations Sub-strand of the End-of-Course Algebra I Test</i>	<i>16</i>
<i>Figure 4. Quadratic Equations—Bar Graph Showing the Difference between the CT and Control Groups</i>	<i>16</i>
Algebraic Operations	17
<i>Table 12. Overview of Sample and Impact of CT on the Algebraic Operations Sub-strand.....</i>	<i>17</i>
<i>Figure 5. Impact on Algebraic Operations: Unadjusted Pre- and Posttest Means for Control and CT (Left); Adjusted Means for Control and CT (Right)</i>	<i>18</i>
<i>Table 13. The Impact of CT on Student Performance on the Algebraic Operations Sub-strand of the End-of-Course Algebra I Test.....</i>	<i>18</i>
<i>Figure 6. Algebraic Operations Score—Difference between CT and Control Groups Showing the Values for the Median Student at Each Quartile of the Pretest</i>	<i>19</i>
<i>Figure 7. Algebraic Operations—Bar Graph Showing the Difference between CT and Control Groups for the Median Student in the Fourth Quartile on the Pretest.....</i>	<i>20</i>
Linear Equations	20
<i>Table 14. Overview of Sample and Impact of CT on the Linear Equations Sub-strand</i>	<i>20</i>
<i>Figure 8. Impact on Linear Equations: Unadjusted Pre- and Posttest Means for Control and CT (Left); Adjusted Means for Control and CT (Right)</i>	<i>21</i>
<i>Table 15. The Impact of CT on Student Performance on the Linear Equation Sub-strand of the End-of-Course Algebra I Test</i>	<i>21</i>
Problem Solving.....	22
<i>Table 16. Overview of Sample and Impact of CT on the Problem Solving Sub-strand</i>	<i>22</i>
<i>Figure 9. Impact on the Problem Solving Sub-strand: Unadjusted Pre- and Posttest Means for Control and CT (Left); Adjusted Means for Control and CT (Right)</i>	<i>23</i>
<i>Table 17. The Impact of CT on Student Performance on Problem Solving Sub-strand of the End-of-Course Algebra I Test.....</i>	<i>23</i>
MODERATING EFFECT OF TEACHER CERTIFICATION ON STUDENT OUTCOMES	24
NWEA Overall End-of-Course Algebra I Test.....	24

<i>Table 18. Moderating Effect of Teacher Certification on NWEA Overall End-of-Course Algebra I Test Outcomes</i>	24
<i>Figure 10. Moderating Effect of Teacher Certification on NWEA Overall Math Outcomes</i>	25
NWEA Algebraic Operations	25
<i>Table 19. Moderating Effect of Teacher Certification on NWEA Algebraic Operations Outcomes</i>	25
<i>Figure 11. Moderating Effect of Teacher Certification on NWEA Algebraic Operations Outcomes</i>	26
RELATIONSHIP BETWEEN THE OUTCOMES AND THE AMOUNT OF CT USAGE	26
<i>Table 20. Regressions of Posttest Sub-strands against Units Completed</i>	27
<i>Table 21. Results Summary for the Regressions of Posttest Sub-strands against Number of Hours Spent</i>	27
DISCUSSION	27
REFERENCES	29

Introduction

Under the *Math Science Partnership Grant*, the Maui Hawaii Educational Consortium sought scientifically based evidence for the effectiveness of the *Cognitive Tutor® (CT)* program, published by Carnegie Learning, as part of the selection process for Algebra I programs to be considered for adoption. The US Department of Education's research funds supported Empirical Education's efforts in the research. A measure of the impact of the program could provide useful evidence to support district decisions about which math program to adopt. The question being addressed specifically by the research is whether students in classes that use *CT* materials achieve higher scores on the standardized math assessment, as measured by the Northwest Evaluation Association (NWEA) math test, than they would if they had been in a control classroom using the Algebra I curricula the Maui schools currently have in place. We conducted an experiment at the Maui Community College and five schools in the Maui School District. For the nine participating teachers, we randomly assigned each of their Algebra I classes to either the group using the new program (the *CT* group) or the group continuing to use the currently adopted textbook program (the control group). The *CT* group teachers used *CT* in their classes for six months during the 2005-2006 school year until the NWEA posttest in math was administered in May 2006.

The district was particularly interested in whether the use of the *CT* program 1) has an impact on student achievement and 2) is a teacher-friendly tool that could be used feasibly in this setting.

The choice of the Carnegie Learning *CT* program was motivated in part by previous research that showed positive results. For example, an experiment reported by the publisher showed that the impact of *CT* was quite substantial (Morgan & Ritter 2002). This research was conducted in a 19,000 student school district in Oklahoma with an ethnic mix including 66.7% White and 17.6% American Indian. Overall, the size of *CT*'s impact in this experiment was 0.29 of a standard deviation. In the K-12 education context, an effect size of 0.29 is considered substantial. This metric for effect size gives us a way to standardize across studies that use different outcome measures.

The design of our experiment in Maui was similar to the Oklahoma study, in that Algebra I classes were randomly assigned to *CT* or to the control condition. This experimental design reflects the requirements of the No Child Left Behind Act, which directs schools to consult reports of rigorous research to guide their adoptions of instructional programs. Random assignment is the best way to avoid potential sources of bias in the result. We are cautious from the outset to emphasize that this study was designed to provide useful information to support a local decision in Maui but not, by itself, to generate broadly generalizable results. The results should not be considered to apply to school districts with practices and populations different from those found in Maui. In addition, because of the small number of teachers involved, the local decision-makers must consider carefully whether those teachers are a good representation of their staff as a whole.

Methods

Research Design

Our study consists of a randomized experiment at the Maui School District and Maui Community College, which compared the outcomes for groups of students taught using the *CT* program (the *CT* group) and students taught using the regular Algebra I curricula (the control group). The design uses a randomization process in which we assigned participating teachers' classes to one group or the other randomly. Each teacher had one or more Algebra I classes, all of which were designated to be in either the *CT* or control group. The primary outcome measure was the overall score on the NWEA end-of-course Algebra test. This design is commonly called a group randomized trial (Bloom, Bos, & Lee, 1999; Raudenbush, 1997). As with any randomized experiment or randomized control trial (RCT), the design has the advantage that it yields unbiased estimates of the treatment impact. This means that we can measure the impact of the intervention without being concerned that systematic differences between the treatment and control groups are due to something other than treatment (e.g., variation in home experiences).

The experiment started mid-Fall of the 2005-2006 school year. We based our analysis on nine teachers, 22 classes, and 541 students.

Materials

As described by Carnegie Learning (2006), *Cognitive Tutor* is a research-based approach to improving student understanding of mathematical concepts. According to the publisher, the program is characterized as having six unique aspects, including a simple and straightforward design, research-based pedagogy, multiple representations of word problems, just-in-time feedback, a skillometer¹, and a blended curriculum of computer lab and classroom activities that complement each other. In practice, students spend about 40% of their class time using software for individualized lessons and the balance of their time engaged in teacher facilitated collaborative, real-world problem-solving activities. The design of the program emphasizes the use of verbal, numerical, algebraic, and graphical representations to solve problems.

CT teachers attended three days of professional development led by a Carnegie Learning consultant and received their CT materials after the initial meeting. Beyond the initial training, teachers were free to make use of the materials as best suited the needs of their classroom and students.

Site Descriptions

Maui County

Maui County, Hawaii is a mixture of a suburban and rural community located on one of the seven islands of Hawaii. According to the US Census Bureau, the total population in 2005 was 138,433. Of the adult population, 86.6% have a high school diploma and 22.4% have a Bachelor's degree. The median age in Maui County is 34 and the median family income in 2005 was \$67,156. For pupils reporting to have one ethnicity, the ethnic breakdown includes 37.4% White, 0.1% Black or African American, 0.4% American Indian or Alaska Native, 31.9% Asian, 9.3% Native Hawaiian or other Pacific Islander, 1.6% Other, and 8.7% Hispanic or Latino. The remaining 19.3% of the population reported to be of two or more ethnicities.

Maui District Schools

The Maui School District is part of the Hawaii Department of Education. The district covers the Molokai and Lanai school systems. It is the second largest school complex, with 20 elementary schools (K-5) and seven middle schools (one K-8 and six 6-8). The School Status and Improvement Report for each of the schools in the district provided information about their student populations for the 2004-2005 school year. The average ethnic breakdown for the participating schools includes approximately 32% Filipino, 28% Part-Hawaiian, 11% White, 8 percent Japanese, 5% Hawaiian, 3% Hispanic, and 14% Other. An average of 26.7% of students participated in the National School Lunch Program, while 15.6% were in Special Education and 5.76% were designated as Limited English Proficient. The Maui Community College had a similar representation of student ethnicity.

Existing Math Program

For their control classes, teachers used a variety of Algebra I textbooks from the following publishers: McDougal Littell, Bittinger and Beecher, Addison-Wesley, Holt, Prentice-Hall, and Merrill. During the study, the control group classes continued using these materials as usual.

¹ Skill bars show students what skills they have mastered, and where they need to improve, motivating the student to take responsibility for his or her own learning.

Sample and Randomization

The initial meeting for the experiment was conducted on October 12, 2005 and attended by six teachers and several administrators from the district and community college. The meeting included an explanation of the *CT* study, a presentation of NWEA testing, and a discussion about the planned research procedures. After a question-and-answer period, those who decided to participate in the study filled out a teacher background information form, which later revealed that among the nine teachers who eventually volunteered, four were first-time math teachers and the remaining five had taught math in the range of 8 to 23 years. Six of the nine teachers were math majors in college and the remaining three had taken some math courses.

We realized at this point that we did not have as large a sample as was called for by our initial design. Because the importance of the information warranted gathering the available data even if the results ultimately proved inconclusive, the district and community college representatives in consultation with the researchers decided to move forward with the experiment. We conducted the randomization by class such that each teacher had both *CT* and control classes. By randomizing in this way rather than randomizing the teachers, we maximized the number of units in the analysis. Since access to *CT* was controlled by the limited number of software licenses, the contamination of the control classes by the teacher methods used in the *CT* classes was expected to be minimal. The classes were further placed into pairs. Teachers paired their most similar classes based on factors such as class size and achievement level. Once the pairs were established, the researchers tossed a coin between each class pair to determine which class would be assigned the *CT* group and which would be assigned to the control group. Where there was an uneven number of classes, the coin was tossed separately for the unpaired member.

Data Collection

The research for this experiment encompasses a multiple methods approach. We collected pre- and posttest math scores from NWEA, class rosters and demographic information on students and teachers from the district, and student usage data from Carnegie Learning. We also collected qualitative data through the use of classroom observations, phone interviews, and web-based surveys from all participating teachers in both groups. We integrated all the information from these multiple sources into a standard data warehouse for the study. Through these methods, we measured and documented implementation factors and student and teacher interactions with the *CT* materials.

Observations

Classroom observations occurred in January 2006. Their purpose was to help us understand and document 1) student and teacher interactions with the *CT* materials (and/or existing math program) materials, 2) the kinds of resources teachers had available for their use, 3) the type of support provided by Carnegie Learning, and 4) the extent to which the *CT* program was being implemented.

Interviews

Structured phone interviews were conducted with the teachers and with a representative from Carnegie Learning in March 2006. Each interview lasted between 20 to 30 minutes. The purpose of these interviews was to gain an understanding of teachers' attitudes and opinions about the *CT* program as well as the kinds of challenges and difficulties they may have encountered with the program. Results from these interviews helped drive subsequent survey questions (described in the following section).

Surveys

Nine web-based surveys were administered to all participating teachers (in both groups) on a bi-weekly schedule from January to April of 2006. The content of these surveys covered any factors that might have influenced the results. Examples of these include preparation time with *CT* supplemental material, outside resources, student engagement, access to computers, and technical difficulties. A final survey addressed questions about teachers' overall experience with the *CT* program as well as the specific units covered throughout the study.

Carnegie Learning Data

We also collected data from Carnegie Learning on individual students' interactions with the *CT* software. As a feature of the *CT* software, the

Teacher's Toolkit automatically recorded student time spent using the software and the number of questions and units each student completed and answered correctly in the different topic areas. Maui School District gave us access to this information and it was used to give a descriptive account of the relationship between achievement and different modes of use.

Achievement Test Scores

The primary outcome measures are student-level test scores on the Northwest Evaluation Association (NWEA) test in math. In fall, the NWEA Math Goals Survey 6+ Achievement Level Test (ALT) was administered to the students in the Maui schools. As a posttest measure, the Algebra end-of-course ALT was administered in spring. The Maui Community College administered the computer-based versions of these tests called Measures of Academic Progress (MAP) at the start and end of the semester. The NWEA Math Goals Survey 6+ test and the Algebra end-of-course test are adaptive and comprehensive tests that reflect the instructional level of each student and measure growth over time. These tests are scored on a Rasch unit (RIT) scale, a measurement scale developed to simplify the interpretation of test scores. This scale is used to measure student achievement and student growth on an equal-interval scale so that a change of one unit indicates the same change in growth, regardless of the actual numerical values. RIT scores range from about 150 to 300 and indicate a student's current achievement level along a curriculum scale for a particular subject.

Table 1. Survey Response Rates

Survey number	Date	Topic	Response rate
Survey 1	Jan. 16-20	Setting the Stage	55.56%
Survey 2	Jan. 23-27	Resources: Books, Computers and other People	100.00%
Survey 3	Feb. 6-10	Assessments	100.00%
Survey 4	Feb. 20-24	Interactions with Materials	100.00%
Survey 5	Mar. 6-10	Student Engagement	100.00%
Survey 6	Mar. 20-24	Planning and Professional Development	100.00%
Survey 7	Apr. 10-14	More Interactions with Materials	77.78%
Survey 8	Apr. 24-28	Program Progress	100.00%
Survey 9	May 22-26	Final Survey: Content Covered and Opinions	77.78%

Statistical Analysis and Reporting

The basic question for the statistical analysis was whether, following the intervention, students in the CT classrooms had higher math scores than those in the control classrooms. The mean impact is estimated using multi-level models that account for the clustering of students in classes, which provides a more accurate, and often more conservative, assessment of the confidence we should have in the findings. We use SAS PROC MIXED (from SAS Institute Inc.) as the primary tool for this work. To increase the precision of our estimate, we include students' ethnicity and pretest scores in the analysis. In our experience, these are good predictors of achievement; including them as covariates in the impact analysis reduces the error variance, which makes it easier to discern the treatment impact.

In addition to the basic analysis of the mean impact, the plan for the study identifies the teacher- and student-level covariates that we expect (through theory or prior research) to make a difference in the effectiveness of the program being tested. The analysis tests for the interactions between these covariates and the experimental condition.

In addition to examining impacts and interactions where we anticipate effects, to better understand unexpected results, we use other demographics, teacher characteristics, and supplementary observational data in exploratory analyses to generate additional hypotheses about which factors potentially moderate the treatment impact.

Our analyses produce several results: among them are the coefficients for fixed effects, effect sizes, and p values. These are found in all the tables where we report the results of the statistical models.

Coefficients. The coefficient can be thought of as the estimated size of an effect. Specifically, it is how much we would predict the outcome to change for a one-unit increase in the corresponding variable. We are often most interested in the coefficient associated with the condition, which is the estimated difference in the outcome between those in the treatment and control groups (holding the values of the other covariates constant and assuming no interactions).

Effect sizes. We also translate the difference between treatment and control into a standardized effect size by dividing the difference by the standard deviation of the outcome, which is a measure of how variable the outcome is. This allows us to compare the results with results from other studies that use different measurement scales. In studies involving student achievement, effect sizes as small as 0.1 (one tenth of a standard deviation) are sometimes found to be important educationally. When possible we also report the effect size of the difference after adjusting for pretest, since that provides a more precise estimate of the effect (i.e., in theory, with many replications, we would expect the adjusted effect size on average to be closer to the true value).

p values. The p value is very important because it gives us a gauge of how confident we can be that the result we are seeing is not due simply to chance. Specifically, it tells us what the probability is that we would get a result with an absolute value as large as – or larger than – the one observed when in fact there is no effect. Roughly speaking, it tells us the risk of concluding that the treatment has had an effect, when in fact it hasn't. Thus a p value of .1 gives us a 10% probability of that happening. We can also think of it as the level of confidence, or the level of belief we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting p values:

1. We have a high level of confidence when $p \leq .05$. (This is the level of confidence conventionally referred to as "statistical significance.")
2. We have some confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.
4. We have no confidence when $p > .20$.

Results

Formation of the Experimental Groups

The randomizing process does not guarantee that an experiment's groups will be perfectly matched. It simply guarantees that there is no intentional selection bias. It is important to inspect the two groups to determine whether any significant differences occurred that might affect the results. The following tables address the nature of the groups in each of the school sites. Table 2 shows the counts of schools, teachers, classes, grades, and students between *CT* and control conditions. This is the full number of students in the experiment at the time the experiment began in October 2005.

Table 2. Distribution of the *CT* and Control Groups by Schools, Teachers, Grades, and Counts of Students

	No. of schools	No. of teachers	No. of classes	Number of students						Total students
				Grade 8	Grade 9	Grade 10	Grade 11	Grade 12	Grade 13	
<i>CT</i>	6	7	11	56	71	97	25	11	13	281
Control	6	7	11	47	83	66	32	9	23	260
Totals	12	14	22	103	154	163	57	20	36	541

Post Randomization Composition of the Experimental Groups

In checking for balance in the composition of the experimental groups, we examine student ethnic background first, followed by teacher certification and achievement pretest outcomes.

Ethnic Composition of Student Population

Table 3 displays the ethnic make-up of the study participants. We observe that that the majority of students in the sample are Asians. The available dataset does not distinguish among the various Asian ethnicities, nor does it provide a breakdown of native Hawaiians, Samoans, Portuguese, Koreans, Filipinos, Japanese, Chinese, etc. or mixed ethnicities.

Table 3. Ethnicity for *CT* and Control Groups

Condition	Ethnicity						Totals
	Asian	Hispanic	Native American	Multi-racial	Black	White	
<i>CT</i>	160	8	2	69	2	40	281
Control	138	16	2	69	1	34	260
Totals	298	24	4	138	3	74	541

Teacher Certification

We observe in Table 4 that teacher certification was not distributed evenly between the conditions even after randomization. There are proportionally more students in classrooms of certified teachers in the *CT* group than in the control group. Chi-square tests confirm that this characteristic was not balanced between conditions. The imbalance may lead the estimate of the impact to depart from its true value.

Table 4. Teacher Certification: Comparison among Students

Condition	Number of students		
	Certified	Uncertified	Totals
<i>CT</i>	182	99	260
Control	153	107	260
Totals	335	206	541
Statistics	DF	Value	<i>p</i> value
Chi-square test	1	2.01	.16

Achievement Pretests

With randomization, we expect the pretest scores to be equally distributed

between *CT* and control groups, but in any single randomization there may be discrepancies between the distributions due to chance. This was the case here: the *CT* and control groups had slightly different average pretest scores on the Fall assessment, as shown in Table 5. However, when we accounted for the fact that outcomes for students of the same teacher tend to be dependent by modeling these dependencies, the discrepancy became less discernable. In the analyses that follow, we add the pretest covariate in order to increase the precision of the impact estimate. (But we recognize that, with or without this covariate, the impact estimate is unbiased as a result of the randomization.)

Table 5. Independent *t* Test of the Difference between Students in *CT* and Control Groups for the NWEA Math Goals Survey 6+ ALT Pretest

Descriptive statistics: Pretest scores	Raw group means	Standard deviation	Number of students	Standard error
<i>CT</i>	226.46	13.14	253	0.83
Control	230.18	12.46	234	0.82
<i>t</i> test for difference between independent means	Difference	DF	<i>t</i> value	<i>p</i> value
Condition (<i>CT</i> -control)	-3.72	485	3.2	<.01 ^a

^aWhen we took clustering of students within teachers into account, the *p* value increased to .16. We expect this initial dependency to exist because students are not randomly assigned to teachers.

Attrition

A high percentage of students did not take the posttest. Out of a total enrollment of 541 based on fall class rosters, 54 students (or 10%) did not have pretest scores. Of these remaining 487 students, posttest scores are missing for 142 or 29%. Table 6 shows the breakdown by the *CT* and control groups. A Chi-square test indicates no relationship between attrition and experimental condition. This is important because it means that the attrition does not bias the comparison between the two groups. Data used in the table also reflect that 54 students did not have a pretest score.

Table 6. Counts of Students Missing Test Score Data

Condition	Categories of missing data		Totals
	Having both pre- and posttest scores	Missing posttest scores	
<i>CT</i>	182	71	253
Control	163	71	234
Totals	345	142	487
Chi-square statistics	DF	Value	<i>p</i> value
	1	.31	.58

Considering all categories of missing data, 36% of enrolled students could not be used in the analysis. Part of the explanation for this high rate of missing data lies in the methods by which the NWEA test was scored and reported. There were 63 known absentees on the day of testing: 24 were *CT* students; 39, control students. The remaining 79 students without a test score may not have completed a sufficient number of items to be given a score and were not distinguished in NWEA's report as having

started the test. In this situation, there is a concern that the non-completers may tend to be students who had difficulty with the test and would have received low scores if they had been able to complete it. Table 7 shows that students with no score for the posttest scored lower on the pretest. This is a large difference and we can be very confident, given the very low *p* value, that it is not due to chance.

Table 7. Independent *t* Test of the Difference in Pretest Scores between Students Who Had Pretest Scores Only and Students with Both Pre- and Posttest Scores

Descriptive statistics: Pretest scores	Raw group means	Standard deviation	Number of students	Standard error	Effect size
Have pretest scores only	224.30	15.13	142	0.62	-0.41
Have both pre- and posttest scores	229.87	11.56	345	1.27	
<i>t</i> test for difference between independent means	Difference		DF	<i>t</i> value	<i>p</i> value
(Missing posttest) – (Have posttest)	-5.57		485	4.40	<.01

Regardless of the explanation for the high attrition rate, it impacts the way the results can be interpreted. While there was no bias in favor of *CT* or control groups, there is a bias toward including higher scoring students. Thus, we can be less confident of the applicability of the findings for lower scoring students.

Implementation Results

Data from three qualitative sources provided information on how the teachers felt about the *CT* program and helped us understand the implementation process. Classroom observations, phone interviews, and surveys were processed and analyzed as separate data sources. These qualitative data were minimally coded and used as descriptive information only.

Classroom Observations

During the time of classroom observations, all observed classes were visited by a group consisting of the project Point of Contact (POC) from the Maui School District, a representative from Carnegie Learning, and a research manager from Empirical Education. Principals from four of the five participating Maui schools greeted and guided this group to the appropriate classrooms. Overall, this group visited all eight of the teachers with *CT* classes in each of the schools (the ninth teacher had only control classes). They observed six *CT* classes and two control classes. Of the six observed *CT* classes, three were using the *CT* textbook only. Two *CT* classes were on a computer-textbook rotation, where half of the class used the computers and the other half used the *CT* textbook. One of the observed *CT* classes had students with full use of computers.

The Carnegie Learning representative spent five to ten minutes with each teacher prior to or at the beginning of each class to troubleshoot any difficulties teachers were having with the *CT* program, assist in technical setup, and provide support to the teachers and students. In every classroom, the representative engaged the students in a ten-minute discussion at the end of the class period about the pros and cons of the *CT* program. In all six of the observed *CT* classes, students reported that they enjoy using the *CT* program overall and that they like the *CT* software better than the *CT* textbook. They also gave suggestions about how to improve the program (e.g., a glossary of terms would be helpful).

In one of the schools, a parent night was scheduled for any of the families who needed assistance with the *CT* program. The principal, POC, and Carnegie Learning representative were present, but no families attended.

In one of the two observed control classrooms, the students followed along in their textbooks, while the teacher used a lecture-style structure. In the other control class, students were working in small groups on an activity while the teacher circulated the room to offer help when needed.

These observational data helped us formulate questions for the phone interviews and web-based surveys.

Teacher Responses

Data collected from the ongoing web-based surveys and one-time phone interviews revealed that all of the teachers with *CT* classes expressed both positive and negative comments about the program.

Positive Experiences

In surveys, teachers reported overall general ease of use of the program as well as positive interactions with the materials. For their control classes, teachers relied on the use of their existing math program or textbook as well as supplemental material that they sought out themselves through the Internet and/or other sources. Teachers responding to the survey reported that they had to spend an average of 1.54 hours utilizing outside resources to enhance their effectiveness teaching their existing math program versus 0.88 hours for their *CT* classes. On a 5-point Likert scale (5 = fully engaged, 1=not engaged), teachers reported a 3.5 rating of student engagement in their control classes versus 4.0 in their *CT* classes.

The final survey focused on overall teacher experience and satisfaction with the *CT* program. One teacher commented that “the program gave students the practice that they need to enhance their class work.” Another teacher said, “Students found the process stimulating. The nature of the self-pacing assisted in differentiating instruction for the varying levels of algebraic ability in class.” Some things that the teachers found difficult with the *CT* program were computer-related issues, the lack of teacher directions and drill work in the textbook, and the level of difficulty in the software.

On a 4-point Likert scale (4 =very satisfied, 1=very dissatisfied), teachers reported a 3.5 overall satisfaction with the *CT* software, 3.2 with the *CT* textbook, and 3.2 with the *CT* Teacher’s Toolkit. All teachers said that they would recommend the *CT* program to other math teachers.

On a similar scale, teachers rated three statements about *CT* materials. They reported an agreement of 2.5 that the *CT* materials helped students learn the content outlined by Hawaii state and district standards, a 3.3 rating of agreement that the *CT* materials helped students learn that math is everywhere, and a 3.17 rating of agreement that the *CT* materials helped students learn that math is a problem-solving approach more than a series of computational steps.

In an interview, one teacher stated, “From a teacher’s standpoint, it makes teaching more interesting and cuts back on prep for material. The materials are really strong.”

Another teacher said, “CogTutor [*CT*] is excellent—it’s exactly what the average student needs today. In other classes, they’re not paying attention. During lecture in my control classes, students start gossiping. I like the computer—it has the utmost individual attention of students. Students want to get on the computer.”

All teachers reported a noticeable difference between their *CT* classes and their control classes. The students in the *CT* classes are more interested in working on the *CT* software than any textbook. They enjoy the lab aspect of the program and appreciate the non-traditional classroom setting. A few teachers added that the students like the math problems with real-life situations because they can relate to them.

Challenges

Some teachers also described their negative experiences and difficulties with implementation and the program itself. In an interview, one teacher shared, “The program is pretty good except for the fact that we started late. I really like the program. My classes are further than my traditional algebra classes.” S/he noted experiencing a lack of computers, which necessitates a computer-textbook rotation. S/he reiterated that “it’s not the program itself, but the implementation could be improved.”

Teachers expressed frustration about not being able to use *CT* as it is designed to be used. Most teachers reported insufficient time for professional development, planning, or collaboration with others. What quickly became apparent during observations, phone interviews, and examination of the initial survey responses were two key factors: delayed program implementation and lack of access to computers. Teachers started using the *CT* textbook in mid-Fall and the *CT* software in January after the technical set-up of the program was complete. By the first week of February, teachers reported going to the computer lab an average of 1.55 days out of the week for an average of 40.63 minutes per lab visit. Teachers continually mentioned these factors in the comment section of the periodic surveys, and again in a survey that focused solely on these issues. Six out of seven teachers who responded to the single-focus survey reported that their *CT* classes were behind schedule due to lack of computers (the one remaining teacher reported that his/her *CT* classes were on schedule). Five of these teachers also commented on the delayed start of the program and/or limited implementation time due to the delayed start of the program. One teacher reported that 40% to 50% of his/her *CT* students were experiencing a lack of motivation due to personal responsibilities outside of school.

For some classes, computer time was limited by the need for students to rotate during a single period. Computer lab days were sometimes inconsistent as teachers had to struggle to secure a time slot. On occasion, when classes had access to computers, they experienced technical difficulties with the software, the school server, or the computer itself. Thus teachers believed that the program was not fully implemented because of this lack of technical capacity.

One teacher reported that his/her students liked the program, but that it was too repetitive. S/he also did not like that the *CT* program did not include tests.

Overall, these qualitative data sources revealed that teachers reported support of the *CT* program, increased student interest and engagement compared to classes without the program, and frustration with the delayed start of the program and lack of access to computers.

Impact of Cognitive Tutor

Our overall outcome measure was the score on the NWEA end-of-course Algebra I test. This score was broken down into four sub-strands: Problem Solving, Quadratic Equations, Linear Equations, and Algebraic Operations. We addressed each of these outcomes in a separate analysis. Across these outcomes, the basic question for the statistical analysis was whether, following the intervention, students in *CT* classrooms had higher scores than those in control classrooms.

For each sub-strand of the NWEA Algebra I test, we first estimate the average impact of *CT* on student performance. These results are presented in terms of effect sizes. We then show the results of mixed model analyses where we estimate whether the impact of the intervention depends on the level of certain moderator variables. For instance, we show the results of a model that tests whether there is a differential impact across the prior score scale. We also model the potential moderating effects of teacher certification. In particular, we were interested in whether the condition's (*CT* versus control) effect varies among classrooms of certified versus uncertified teachers. We provide a separate table of results for each of these moderator analyses. The fixed factor part of each table provides estimates of the factors of interest. For instance, in the case where we look at the moderating effect of a student's prior score, we show whether being in a *CT* or a control class makes a difference for the average student. We also show whether the impact of the intervention varies across the prior score scale. At the bottom of the table we give results for technical review – these often consist of random effects estimates which are added to the analysis to account for the fact that the individual results that come from a common upper-level unit (e.g., class or teacher) tend to be similar (i.e., the observations are dependent). In some cases, to account for these dependencies, we model fixed rather than random effects but do not present the individual fixed effects estimates. Modeling the dependencies results in a more conservative estimate of the treatment impact. We report interactions for which we have at least limited confidence that the observed effect was not due to chance.

Overall Score on the NWEA End-of-Course Algebra I Test

Table 8 provides a summary of the sample we used in the analysis and the results for the comparison of NWEA scores for students in *CT* and control groups. The “Unadjusted” row gives

Table 8. Overview of Sample and Impact of *CT* on the Overall Score: NWEA End-of-Course Algebra I Test

	Condition	Means	Standard deviations ^a	No. of students	No. of classes	No. of teachers	Effect size	<i>p</i> value ^b	Percentile standing
Un-adjusted	<i>CT</i>	242.92	7.67	193	11	8	-0.24	.12	40.5%
	Control	244.56	7.37	171	11	9			
Adjusted	<i>CT</i>	246.60	7.67	182	11	8	0.03	.83	51.14%
	Control	246.38	7.46	163	11	9			

^a The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated from the scores of the students in the sample for that row.

^b The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes both the pretest and, where relevant, indicators for upper-level units within which the units of randomization are nested.

^c Modeling separate intercepts for upper-level units leads to estimates of performance, in the absence of treatment, which are specific to those units. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant across upper-level units, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

information about all the students in the original sample for whom we have a posttest score. This shows the means and standard deviations as well as a count of the number of students, classes, and teachers in that group. The last two columns provide the effect size, which is the size of the difference between the means for *CT* and control in standard deviation units. Also provided is the *p* value, indicating the probability of arriving at a difference as large as, or larger than, the absolute value of the one observed when there truly is no difference. The “Unadjusted” row is based on all students with a posttest and the estimated effect size takes into consideration the clustering of students in upper-level units (i.e., that students are grouped within teachers). The “Adjusted” row is based on the students who have both pretest and posttest scores. This is the sample that we use in the analyses on which we base our results reported in Table 9. The means, and therefore the effect size, are adjusted to take into account the student pretest scores; hence, these statistics are adjusted for any chance imbalance in the two randomized groups.² They also take into consideration the effect of students being grouped within teachers.³

Figure 1 provides a visual representation of specific information in Table 8. The bar graphs represent average performance using the metric of the NWEA Algebra I test.

The panel on the left shows average pre- and posttest scores for the control and *CT* groups. The pre- and posttest bars show that both the *CT* and control groups on average grew in their math achievement during the year.

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students’ pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled ‘adjusted’ in Table 8.) The overall impact on math as an effect size (standard deviation units) is 0.29 which is equivalent to a gain of about 11 percentile points if the median *CT* student were placed in the control group. However, the high *p* value for the treatment effect (.83) indicates that we should have no confidence that the actual difference is different from zero. We added 80% confidence intervals to the tops of the bars. The overlap in these intervals further indicates that any difference we see is easily due to chance.

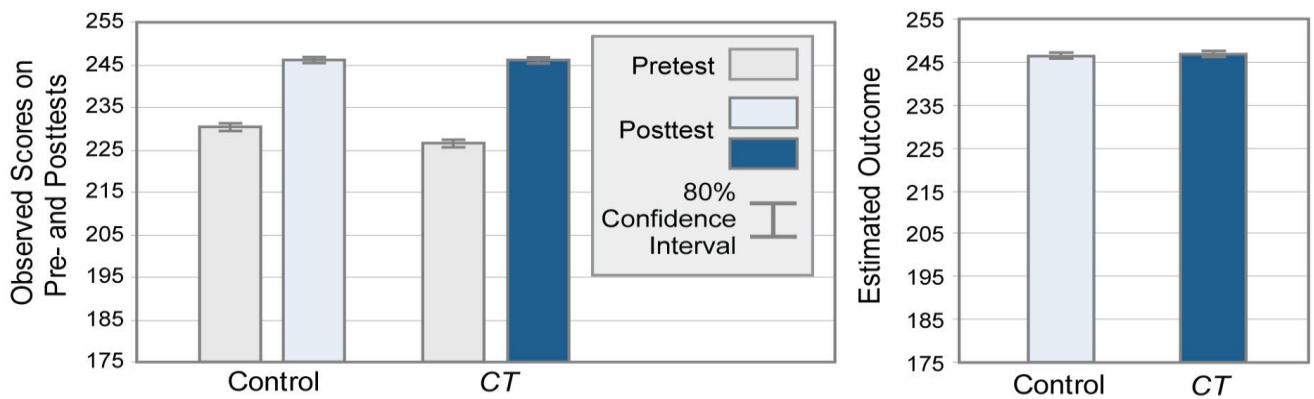


Figure 1. Impact on the Overall Score of the NWEA Algebra I Test: Unadjusted Pre- and Posttest Means for Control and *CT* (Left); Adjusted Means for Control and *CT* (Right)

² We also include any fixed effects used to estimate differences among upper level units.

² We note that the number of cases used to calculate the effect size will often be larger than the number used in the mixed model analysis because to be included in the latter analysis a student must have both a pretest and a posttest score.

Table 9 shows the estimated impact of *CT* on students' performance on the overall score on the NWEA end-of-course Algebra I test. In our analysis we included students' pretest scores as a covariate in order to increase the precision of our estimate of the treatment effect. We accounted for the dependencies among observations within classes by modeling random effects for classes. Also, we modeled teacher fixed effects to reflect our design where we blocked by teacher.

The bottom rows of Table 9 contain the details about random effects that are needed for technical review. The row in the table labeled "Effect of *CT* for a student with an average pretest" gives us information about whether *CT* made an overall difference in test scores. The estimate associated

Table 9. The Impact of *CT* on Student Performance on the Overall Score of the NWEA End-of-Course Algebra I Test

Mixed model: Fixed factors related to overall outcome	Estimate of coefficient	Standard error	DF	t value	p value
Outcome for the control student with an average pretest	243.86	1.59	12	153.87	<.01
Change in outcome for each unit-increase of the pretest	0.40	0.03	313	14.99	<.01
Effect of <i>CT</i>	-1.34	1.05	12	-1.28	.23
Mixed model: Random components	Estimate of variance component	Standard error		z value	p value
Class mean achievement	3.96	2.25		1.76	.04
Within class variation	21.25	1.70		12.51	<.01

Notes. We can refer to the student with an "average pretest" because the pretest score is centered at the mean. In this analysis we blocked by teachers, which resulted in fixed effects estimates for teachers that we exclude from the table.

with the treatment is -1.34, which is the estimated difference between students in the *CT* and control conditions, adjusting for variation in pretest scores. This shows a small negative difference associated with *CT*. The *p* value of .23 gives us no confidence that the true impact is different from zero. In other words, the result could easily reflect a chance difference.

We see a discrepancy in the sizes of the estimated effects between Table 8 and Table 9; in Table 8 we describe an average effect whereas here we describe an effect for the average student, and these are not equivalent. However, neither can be distinguished statistically from zero.

As a visual representation of the result described in Table 9, we present a scatterplot in Figure 2, which graphs student growth over the school year in terms of overall math achievement as measured by the NWEA test. This graph shows where each student fell in terms of his or her starting point (horizontal x-axis) and his or her outcome score (vertical y-axis). Each point plots one student's post-intervention score against his or her pre-intervention score. The darker points represent *CT* students; the lighter points, control students. The shaded area in the lower right of the graph is the area of negative change (i.e., where students lost ground). Nearly all students, regardless of condition, improved on the overall math scale used in the NWEA tests. Our analysis is unable to discern a difference between the two conditions on the overall score.

The scatterplot displays a weakness in the NWEA end-of-course Algebra I test used as the outcome measure. The paper version of this test, ALT, does not have different forms for students at different levels and has a distinct floor that is evident just above 230 points. (The computer version of this same test, MAP, does not have this weakness since it is an adaptive test providing easier or harder questions depending on the student performance on earlier questions.) This may distort the results—especially for students who, in an adaptive test, may have scored lower than 230. In order to test the impact of this distortion, we ran the same analysis as in Table 9 but removing students with scores below 233. The outcome was substantially the same but with a slightly less low coefficient and a slightly lower p value. In the sub-strands presented below, the floor is not evident. For the overall score, and possibly for the sub-strands, we have to be cautious in the application of the results to lower scoring students.

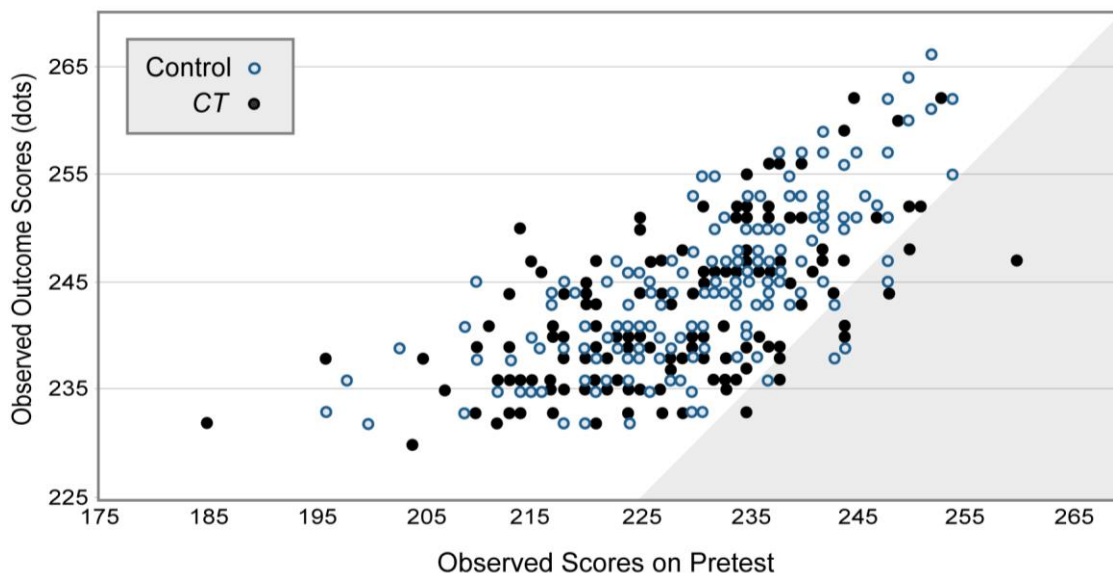


Figure 2. Overall Score—Scatterplot for All the Students

Because there was no discernible difference between the *CT* and control groups on the overall score on the end-of-course Algebra I test, we conducted further analyses to determine whether the impact of *CT* on students' performance varied by sub-strand, which included Quadratic Equations, Algebraic Operations, Linear Equations, and Problem Solving.

Quadratic Equations

Our next set of analyses addresses Algebra achievement as measured by the Quadratic Equations sub-strand. Table 10 provides a summary of the sample we used in the analyses and the results for the comparison of the *CT* and control groups. The interpretation of this table is the same as for Table 8. The information for the adjusted effect size is based on the sample that we use in the analyses on which we base our results reported in Table 11. The means, and therefore the effect size, are adjusted to take into account the student pretest scores.

Table 10. Overview of Sample and Impact of CT on the Quadratic Equations Sub-strand

	Condition	Means	Standard deviations ^a	No. of students	No. of classes	No. of teachers	Effect size	<i>p</i> value ^b	Percentile standing
Un-adjusted	CT	238.78	11.52	193	11	8	-0.34	<.01	36.69%
	Control	241.54	11.28	171	11	9			
Adjusted	CT	238.55	11.66	182	11	8	-0.33	<.01	37.07%
	Control	242.40	11.42	163	11	9			

^a The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated from the scores of the students in the sample for that row.

^b The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes both the pretest and, where relevant, indicators for upper-level units within which the units of randomization are nested.

^c Modeling separate intercepts for upper-level units leads to estimates of performance, in the absence of treatment, which are specific to those units. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant across upper-level units, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Figure 3 provides a visual representation of specific information in Table 10. The bar graphs represent average performance using the metric of the Quadratic Equations sub-strand.

The panel on the left shows average pre- and posttest scores for the control and CT groups. The pre- and posttest bars show that both the CT and control groups on average improved their Quadratic Equations scores.

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled "Adjusted" in Table 10). The overall impact on the Quadratic Equations sub-strand as an effect size (standard deviation units) is -0.33, which is equivalent to a loss of about 13 percentile points if the median control student were placed in the CT group. The *p* value of <.01 gives us high confidence that the actual effect is different from zero. In other words, students in the control condition achieve higher Quadratic Equations scores than CT students.

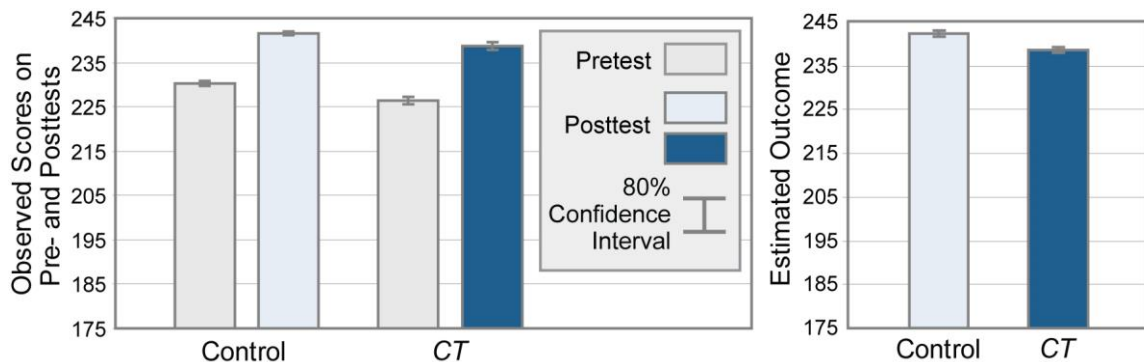


Figure 3. Impact on the Quadratic Equations Sub-strand: Unadjusted Pre- and Posttest Means for Control and CT (Left); Adjusted Means for Control and CT (Right)

Table 11 shows the estimated impact of *CT* on students' performance on the Quadratic Equations sub-strand. The coefficient associated with the *CT* condition is -3.44, which is the estimated difference between students in the *CT* and control conditions, adjusting for variation in the pretest score. That is, based on pretest performance, we estimated that students would score more than three points less on the outcome measure if they are in a *CT* classroom instead of a control classroom. This effect has a *p* value of .02, which means that there is a probability of only 2% that an impact with an absolute value this large or larger would happen by chance. We can be very confident that this result is not due to chance.

Table 11. The Impact of *CT* on Student Performance on the Quadratic Equations Sub-strand of the End-of-Course Algebra I Test

Mixed model: Fixed factors related to the Quadratic Equations outcome	Estimate of coefficient	Standard error	DF	<i>t</i> value	<i>p</i> value
Outcome for the control student with an average pretest	240.87	2.01	12	119.79	<.01
Change in outcome for each unit-increase of the pretest	0.32	0.05	315	6.58	<.01
Effect of <i>CT</i>	-3.44	1.22	12	-2.83	.02

Mixed model: Random components	Estimate of variance component	Standard error	<i>z</i> value	<i>p</i> value
Class mean achievement	83.22	6.62	12.58	<.01
Within class variation	1.70	2.87	0.59	.28

Note. We can refer to the student with an “average pretest” because the pretest score is centered at the mean. In this analysis we blocked by teacher, which resulted in fixed effects estimates for teachers that we exclude from the table.

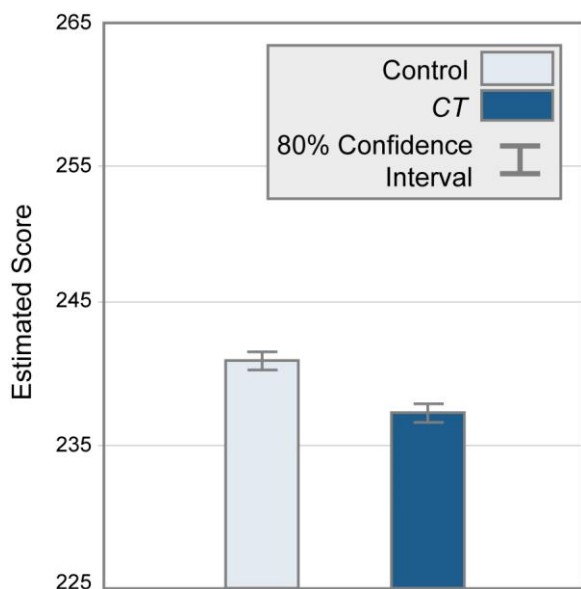


Figure 4. Quadratic Equations—Bar Graph Showing the Difference between the *CT* and Control Groups

To help visualize this result, we present in Figure 4 a bar graph showing the predicted difference on the posttest between the *CT* and control conditions for a student who has an average score on the pretest. We are 80% sure that the true difference between conditions would place the top of the bars within the confidence interval markers. We can see that the impact for *CT* was in the negative direction.

We conducted a further exploratory analysis of these results based on information about the specific units covered by most of the teachers in their *CT* and control classes. If the teachers covered this topic more fully in their control classes, this could account for the negative results for Quadratic Equations. We identified

two teachers who covered Quadratic Equations in their *CT* classes, but not in their control classes. We then conducted two additional exploratory analyses. First, we looked at the results just for the teachers who covered Quadratic Equations in their *CT* but not their control classes. Second, we looked at the results for the teachers who covered Quadratic Equations in both their classes. In both cases, the control classes still outperformed the *CT* classes. The results for these exploratory analyses support our initial findings of a negative impact.

Algebraic Operations

Our next set of analyses addresses Algebra achievement as measured by the Algebraic Operations sub-strand. Table 12 provides a summary of the sample we used in the analyses and the results for the comparison of the *CT* and control groups. The interpretation of this table is the same as for Table 8. The information for the adjusted effect size is based on the sample that we use in the analyses on which we base our results reported in Table 13. The means, and therefore the effect size, are adjusted to take into account the student pretest scores.

Table 12. Overview of Sample and Impact of *CT* on the Algebraic Operations Sub-strand

	Condition	Means	Standard deviations ^a	No. of students	No. of classes	No. of teachers	Effect size	<i>p</i> value ^b	Percentile standing
Un-adjusted	<i>CT</i>	240.27	10.18	193	11	8	-0.29	.10	38.59%
	Control	243.37	10.03	171	11	9			
Adjusted	<i>CT</i>	241.05	9.99	182	11	8	-0.25	<.01	40.13%
	Control	243.53	10.18	163	11	9			

^a The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated from the scores of the students in the sample for that row.

^b The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes both the pretest and, where relevant, indicators for upper-level units within which the units of randomization are nested.

^c Modeling separate intercepts for upper-level units leads to estimates of performance, in the absence of treatment, which are specific to those units. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant across upper-level units, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Figure 5 provides a visual representation of specific information in Table 12. The bar graphs represent average performance using the metric of the Algebraic Operations sub-strand.

The panel on the left shows average pre- and posttest scores for the control and *CT* groups. The pre- and posttest bars show that both the *CT* and control groups on average improved their Algebraic Operations scores.

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled "Adjusted" in Table 12). The overall impact on the Algebraic Operations sub-strand as an effect size (standard deviation units) is -0.25 which is equivalent to a loss of about 10 percentile points if the median control student were placed in the *CT* group. We have strong confidence that this effect is not due to chance.

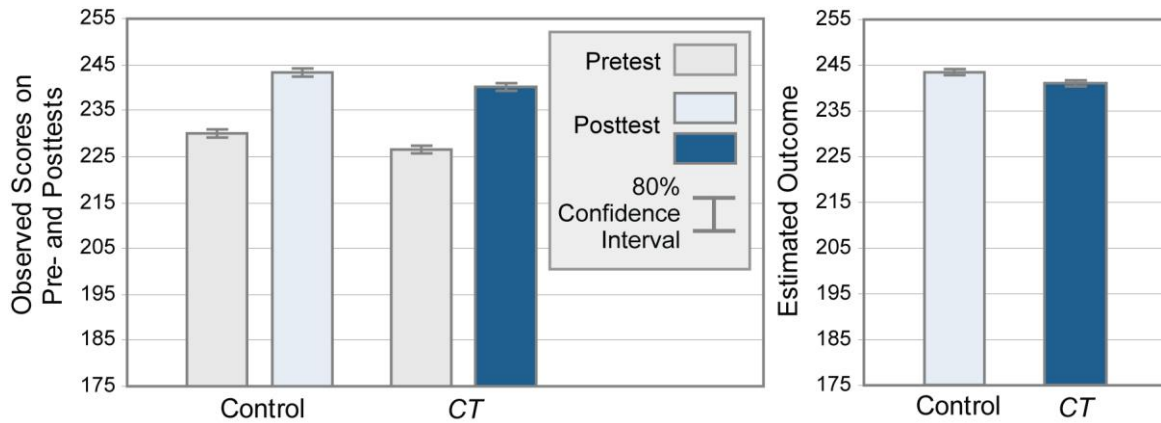


Figure 5. Impact on Algebraic Operations: Unadjusted Pre- and Posttest Means for Control and CT (Left); Adjusted Means for Control and CT (Right)

Table 13 shows the estimated impact of CT on students' performance on the Algebraic Operations sub-strand. As with Quadratic Equations, we find the control students scoring higher on this sub-strand than the CT group. For a student with an average score on the pretest, there is roughly a

Table 13. The Impact of CT on Student Performance on the Algebraic Operations Sub-strand of the End-of-Course Algebra I Test

Mixed model: Fixed factors related to the Algebraic operations outcome	Estimate of coefficient	Standard error	DF	t value	p value
Outcome for the control student with an average pretest	242.71	2.39	12	101.41	<.01
Change in outcome for the control student for each unit-increase on the pretest	0.55	0.06	317	8.88	<.01
Effect of CT for a student with an average pretest	-2.60	1.55	12	-1.68	.12
Change in the effect of CT for each unit-increase on the pretest	-0.15	0.08	317	-1.82	.07
Mixed model: Random components	Estimate of variance component	Standard error		z value	p value
Class mean achievement	7.89	5.05		1.56	.06
Within class variation	59.02	4.69		12.58	<.01

Note. We can refer to the student with an "average pretest" because the pretest score is centered at the mean. The effect size is computed using the 364 students who had posttest data. The analysis that uses the pretest covariate involved 345 students who had both pretest and posttest scores. In this analysis we blocked by teachers, which resulted in fixed effects estimates for teachers which we exclude from the table.

2.6-point disadvantage to being in the *CT* group. Considered within the model presented in Table 13, the low p value of .12 gives us some confidence that the estimated impact for a student with an average score on the pretest is not due to chance. More importantly for this model, the relatively low p value (.07) for the interaction between the pretest score and *CT* points out that the value of *CT* cannot be understood without considering how *CT* and the pretest score work together.

This interaction is most readily interpreted through inspection of graphs. As a visual representation of this result, Figure 6 shows the predicted difference between the *CT* and control groups. This graph is a representation of this separation as a difference, that is, the predicted outcome for a *CT* student minus the predicted outcome for a control student. Around the difference line, we provide gradated bands representing confidence intervals. These confidence intervals are an alternative way of expressing uncertainty in the result. The band with the darkest shading surrounding the dark line is the “50-50” area, where the difference is considered equally likely to lie within the band as not. The region within the outermost shaded boundary is the 95% confidence interval—we are 95% sure that the true difference lies within these extremes. Between the 50% and 95% confidence intervals we also show the 80% and 90% confidence intervals. Consistent with the results in Table 13, there is evidence of a negative impact for higher performing students, and little or no impact at the lower end of the pretest scale.

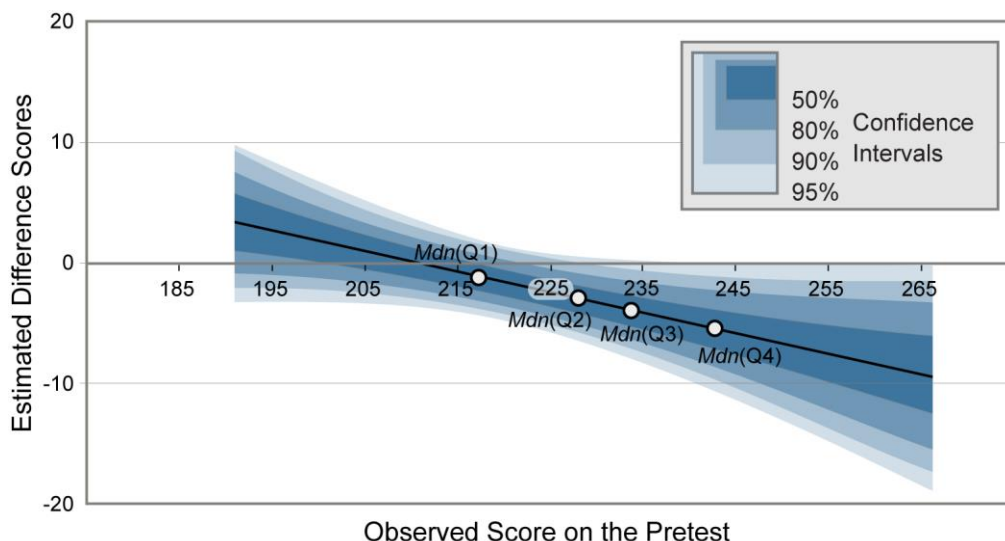


Figure 6. Algebraic Operations Score—Difference between *CT* and Control Groups Showing the Values for the Median Student at Each Quartile of the Pretest

An alternative way of understanding the information in Figure 6 is to represent the result using a bar graph specifically for the student at the median of the top quartile of the pretest. Figure 7 presents the estimated difference between *CT* and control for the median student in the fourth quartile on the pretest. Figure 7 indicates that the median student in the top quartile performs substantially better using the control materials than the *CT* materials.

Linear Equations

Our next set of analyses addresses Algebra achievement as measured by the Linear Equations sub-strand. Table 14 provides a summary of the sample we used in the analyses and the results for the comparison of the *CT* and control groups. The interpretation of this table is the same as for Table 8. The information for the adjusted effect size is based on the sample that we use in the analyses on which we base our results reported in Table 15. The means, and therefore the effect size, are adjusted to take into account the student pretest scores.

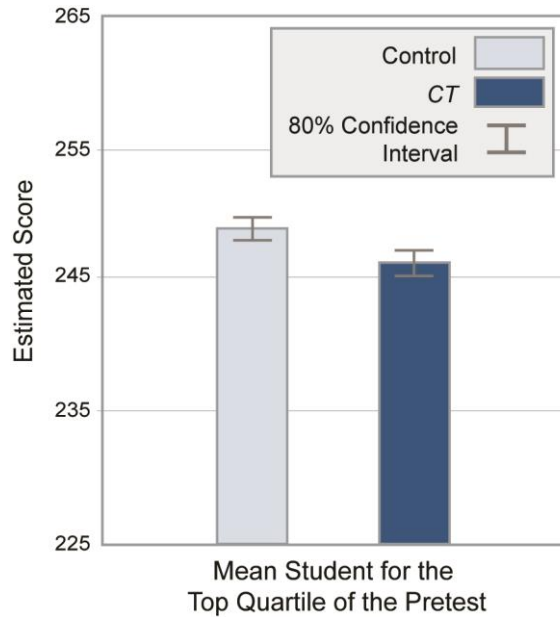


Figure 7. Algebraic Operations—Bar Graph Showing the Difference between *CT* and Control Groups for the Median Student in the Fourth Quartile on the Pretest

Table 14. Overview of Sample and Impact of *CT* on the Linear Equations Sub-strand

	Condition	Means	Standard deviations ^a	No. of students	No. of classes	No. of teachers	Effect size	<i>p</i> value ^b	Percentile standing
Un-adjusted	<i>CT</i>	244.63	9.71	193	11	8	-0.14	.28	44.43%
	Control	245.76	9.69	171	11	9			
Adjusted	<i>CT</i>	244.81	9.57	182	11	8	-0.04	0.65	48.40%
	Control	245.24	9.84	163	11	9			

^a The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated from the scores of the students in the sample for that row.

^b The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes both the pretest and, where relevant, indicators for upper-level units within which the units of randomization are nested.

^c Modeling separate intercepts for upper-level units leads to estimates of performance, in the absence of treatment, which are specific to those units. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant across upper-level units, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Figure 3 provides a visual representation of specific information in Table 14. The bar graphs represent average performance using the metric of the Linear Equations sub-strand.

The panel on the left shows average pre- and posttest scores for the control and *CT* groups. The pre- and posttest bars show that both the *CT* and control groups on average improved their Linear Equations scores. The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled "Adjusted" in Table 14). The overall impact on the Linear Equations sub-strand as an effect size (standard deviation units) is -0.04 which is equivalent to a loss of about one and a half percentile points if the median control student were placed in the *CT* group. However, the *p* value indicates that there is a .65 probability that the difference is due to chance.

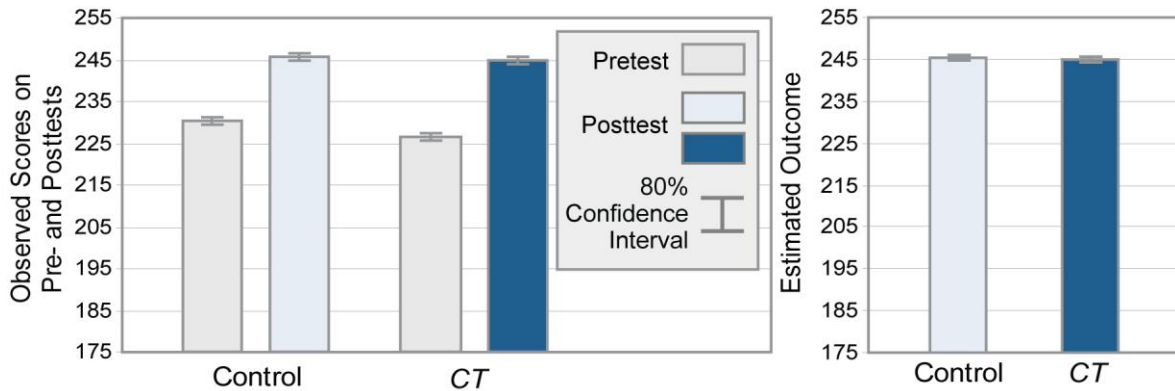


Figure 8. Impact on Linear Equations: Unadjusted Pre- and Posttest Means for Control and *CT* (Left); Adjusted Means for Control and *CT* (Right)

Table 15 shows the estimated impact of *CT* on students' performance on the Linear Equations sub-strand. The coefficient associated with the effect of *CT* is -0.34, which is the difference between students in the two groups after we adjust for the pretest score. This is a small difference and the *p* value, which indicates the probability of finding a difference with an absolute value of 0.34 or larger, gives us no confidence that this result is not just due to chance. In other words, we could easily observe an effect of this size when in fact there is no difference.

Table 15. The Impact of *CT* on Student Performance on the Linear Equation Sub-strand of the End-of-Course Algebra I Test

Mixed model: Fixed factors related to the Linear equations outcome	Estimate of coefficient	Standard error	DF	<i>t</i> value	<i>p</i> value
Outcome for the control student with an average pretest	246.36	1.99	12	123.26	<.01
Change in outcome for each unit-increase on the pretest	0.32	0.04	308	7.98	<.01
Effect of <i>CT</i>	-0.34	1.29	12	-0.26	.80
Mixed model: Random components	Estimate of variance component	Standard error		<i>z</i> value	<i>p</i> value
Class mean achievement	4.45	3.24		1.37	.09
Within class variation	53.24	4.28		12.43	<.01

Note. We can refer to the student with an "average pretest" because the pretest score is centered at the mean. The effect size is computed using the 364 students who had posttest data. The analysis that uses the pretest covariate involved 345 students who had both pretest and posttest scores. In this analysis we blocked by teacher, which resulted in fixed effects estimates for teachers that we exclude from the table.

Problem Solving

Our next set of analyses addresses Algebra achievement as measured by the Problem Solving sub-strand. Table 16 provides a summary of the sample we used in the analyses and the results for the comparison of the *CT* and control groups. Again, the interpretation of this table is the same as for Table 8. The information for the adjusted effect size is based on the sample that we use in the analyses on which we base our results reported Table 17. The means, and therefore the effect size, are adjusted to take into account the student pretest scores.

Table 16. Overview of Sample and Impact of *CT* on the Problem Solving Sub-strand

	Condition	Means	Standard deviations ^a	No. of students	No. of classes	No. of teachers	Effect size	<i>p</i> value ^b	Percentile standing
Un-adjusted	<i>CT</i>	246.20	12.02	193	11	8	-0.04	.78	48.40%
	Control	246.25	11.12	171	11	9			
Adjusted	<i>CT</i>	246.60	12.30	182	11	8	0.02	.83	50.80%
	Control	246.38	11.13	163	11	9			

^a The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated from the scores of the students in the sample for that row.

^b The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes both the pretest and, where relevant, indicators for upper-level units within which the units of randomization are nested.

^c Modeling separate intercepts for upper-level units leads to estimates of performance, in the absence of treatment, which are specific to those units. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant across upper-level units, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Figure 9 provides a visual representation of specific information in Table 16. The bar graphs represent average performance using the metric of the Problem Solving sub-strand.

The panel on the left shows average pre- and posttest scores for the control and *CT* groups. The pre- and posttest bars show that both the *CT* and control groups on average improved their Problem Solving scores.

The panel on the right shows estimated performance on the posttest for the two groups based on a model that adjusts for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled "Adjusted" in Table 16). The overall impact on the Problem Solving sub-strand as an effect size (standard deviation units) is 0.02 which is equivalent to a gain of less than one percentile point if the median control student were placed in the *CT* group. The high *p* value indicates that there is a .83 probability that the difference is due to chance.

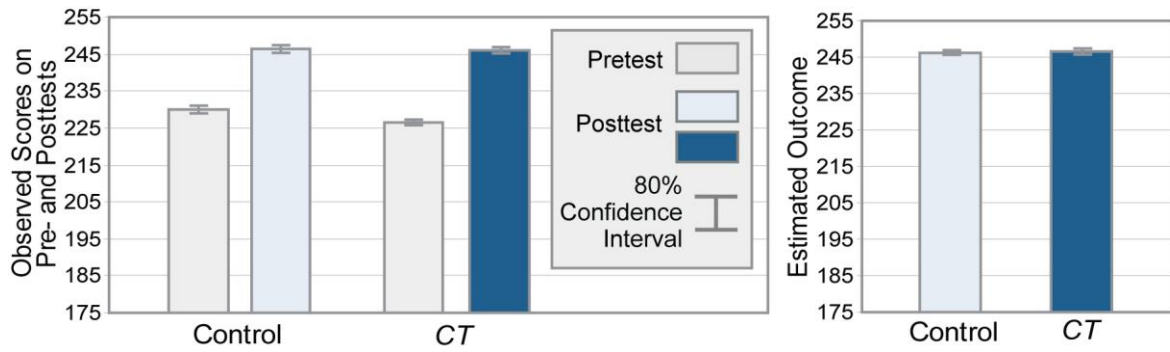


Figure 9. Impact on the Problem Solving Sub-strand: Unadjusted Pre- and Posttest Means for Control and CT (Left); Adjusted Means for Control and CT (Right)

Table 17 shows the estimated impact of CT on students' performance on the Problem Solving sub-strand. There was a very small difference associated with CT (0.29). Although in this case the estimate is in the positive direction, this amounts to a very small difference and the high *p* value indicates that we cannot distinguish this difference from zero.

Table 17. The Impact of CT on Student Performance on Problem Solving Sub-strand of the End-of-Course Algebra I Test

Mixed model: Fixed factors related to the Problem solving outcome	Estimate of coefficient	Standard error	DF	<i>t</i> value	<i>p</i> value
Outcome for a control student with an average pretest	242.36	2.51	12	96.75	<.01
Change in outcome for each unit-increase on the pretest	0.47	0.05	315	10.07	<.01
Effect of CT	0.29	1.62	12	0.18	.86
Mixed model: Random components	Estimate of variance component	Standard error		<i>z</i> value	<i>p</i> value
Class mean achievement	8.08	5.25		1.54	.06
Within class variation	72.52	5.77		12.57	<.01

Note. We can refer to the student with an “average pretest” because the pretest score is centered at the mean. In this analysis we blocked by teacher, which resulted in fixed effects estimates for teachers that we exclude from the table.

Moderating Effect of Teacher Certification on Student Outcomes

NWEA Overall End-of-Course Algebra I Test

Following a suggestion by the district’s MSP consultant, we also considered whether the treatment impact is differentially effective for students who had certified teachers versus those with uncertified teachers. The suggestion was that the negative impact of *CT* may have been a result of implementation issues such as the late delivery of the product. This may have impacted the less experienced teachers to a greater extent because they may have been less able to compensate for the lack of materials available for those classes. Table 18 shows the moderating effect of teacher certification on students’ performance on the NWEA Overall End-of-Course Algebra I Test outcomes in math.

Table 18. Moderating Effect of Teacher Certification on NWEA Overall End-of-Course Algebra I Test Outcomes

Fixed effects ^a	Estimate	Standard error	DF	t value	p value
Outcome for the uncertified teacher’s control student with an average pretest	242.67	1.45	12	167.12	<.01
Change in outcome for each unit-increase on the pretest	0.37	0.03	317	14.76	<.01
Difference (students of certified – students of uncertified teachers) in control outcome	2.71	2.21	12	1.22	.24
Effect of <i>CT</i> for uncertified teachers’ students	0.11	1.13	12	0.10	.93
Difference (students of certified – students of uncertified teachers) in the effect of <i>CT</i>	-4.13	1.79	12	-2.30	.04
Random effects	Estimate	Standard error		z value	p value
Class mean achievement	2.30	1.68		1.37	.09
Within-class variation	22.24	1.76		12.60	<.01

Notes. We can refer to the student with an “average pretest” because the pretest score is centered at the mean.

As a visual representation of the result described in Table 18, Figure 10 shows the estimated difference between *CT* and control for an average student with a certified teacher versus a non-certified teacher. The overlap in confidence intervals indicates no difference between *CT* and control for students with a non-certified teacher. However, we found that students in the control condition outperformed those in the *CT* condition when taught by an uncertified teacher.

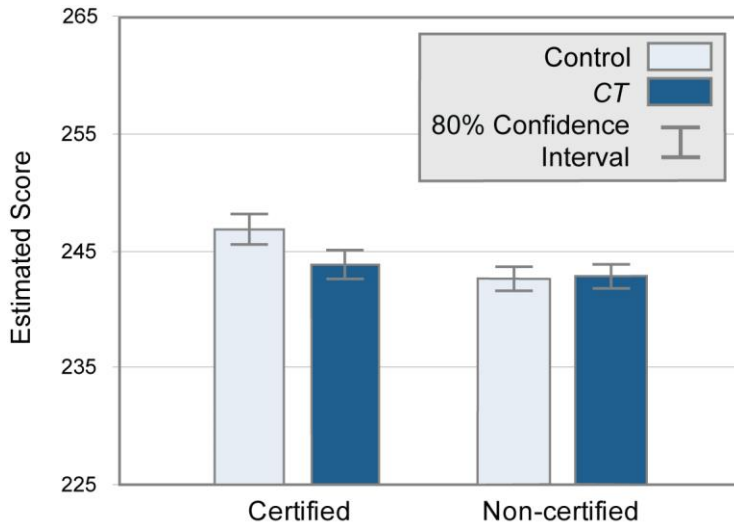


Figure 10. Moderating Effect of Teacher Certification on NWEA Overall Math Outcomes

condition outperformed those in the *CT* condition when taught by a certified teacher.

NWEA Algebraic Operations

Table 19 shows the same moderator analysis but using the Algebraic Operations subscale as the outcome. We chose Algebraic Operations for this exploratory analysis because this sub-strand has the largest negative effect on student outcomes, and because the program under study is focused on Algebra.

As a visual representation of the result described in Table 19, Figure 11 shows the estimated difference between *CT* and control for an average student with a certified teacher versus a non-certified teacher. The overlap in confidence intervals indicates no difference between *CT* and control for students with a non-certified teacher.

However, students in the control

Table 19. Moderating Effect of Teacher Certification on NWEA Algebraic Operations Outcomes

Fixed effects	Estimate	Standard error	DF	<i>t</i> value	<i>p</i> value
Outcome for the uncertified teacher's control student with an average pretest	241.12	2.12	12	113.60	<.01
Change in outcome for each unit-increase on the pretest	0.45	0.04	317	10.78	<.01
Difference (students of certified – students of uncertified teachers) in control outcome	4.29	3.21	12	1.34	.21
Effect of <i>CT</i> for uncertified teachers' students	0.16	1.63	12	0.10	.92
Difference (students of certified – students of uncertified teachers) in the effect of <i>CT</i>	-7.18	2.58	12	-2.79	.02
Random effects	Estimate	Standard error		<i>z</i> value	<i>p</i> value
Class mean achievement	3.82	3.58		1.07	.14
Within-class variation	59.43	4.72		12.59	<.01

For both of the scales we analyzed, we found the same pattern. Certification made a significant difference but not in the expected way. *CT* appears to have the effect of lowering the performance of the more experienced teachers. Students in the certified teachers' *CT* classes performed at the lower level of the students in the uncertified teachers classes.

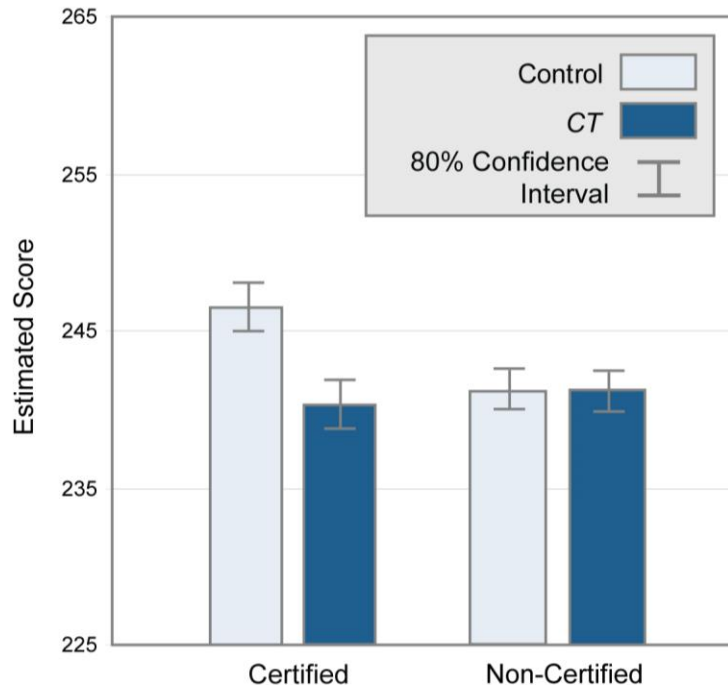


Figure 11. Moderating Effect of Teacher Certification on NWEA Algebraic Operations Outcomes

Relationship between the Outcomes and the Amount of *CT* Usage

Given that there were negative results for the outcomes for Algebraic Operations and Quadratic Equations, and no discernible differences between *CT* and control for the overall score, Linear Equations, and Problem Solving, we investigated whether our observations and data from other sources might provide some insight. Because our qualitative data sources revealed that a consistent barrier to implementation was the lack of computer time, we extracted student usage data collected from the *CT* Teacher's Toolkit. We specifically looked at the measures of time to determine whether there was a relationship between time with *CT* software and achievement for the *CT* group. We reasoned that if lack of access to computers was responsible for the poor results, we should see a relationship between the amount of time that the student spends with *CT* and his or her scores on the posttest. The *CT* Teacher's Toolkit provided detailed information on student usage of the *CT* software (including individual student time spent per problem, per unit). We looked at the relationships between posttest performance on the different sub-strands and the amount of time students spent with *CT*. We also looked at the relationships between these outcomes and the number of units students completed while using *CT*. For both analyses we adjusted for pretest performance. The results are summarized in Table 20 and Table 21.

We start with Units Completed because it is, according to Carnegie Learning, the measure mostly likely to correlate with positive outcomes.

Table 20 shows the relationship between the number of units completed and the score on each of the outcomes, controlling for performance on the pretest. For all of the outcomes, we see a positive relationship. We find the smallest positive relationship for the Algebraic Operations sub-strand and the strongest relationship for the Quadratic Equations sub-strand. These regression coefficients mean that the model predicts that the more units students completed on the CT software, the higher their scores were overall and for all of the sub-strands. For every extra unit the student completes, that student can raise his or her “Overall score by .41 – Problem Solving by .56 points, Linear Equations by .29, Algebraic Operations” by .30 and Quadratic Equations by .64.

Table 20. Regressions of Posttest Sub-strands against Units Completed

Outcomes	Regression coefficient for number of units completed	p value
Overall Score	0.33	<.01
Problem Solving	0.38	.01
Linear Equations	0.37	.02
Algebraic Operations	0.16	.23
Quadratic Equations	0.59	<.01

Note. These coefficients are taken from the multilevel model controlling for the pretest and modeling teacher, class as fixed effects.

These high correlations between student achievement and units completed generally confirm the validity of the NWEA test. However, there remains some question as to the construct for Algebraic Operations where achievement showed no correlation to units completed. These results also show that in most areas, the program, when it was used, was working as expected.

Table 21. Results Summary for the Regressions of Posttest Sub-strands against Number of Hours Spent

Outcomes	Regression coefficient for time spent (hours)	p value
Overall Score	0.02	.84
Problem Solving	0.12	.58
Linear Equations	0.05	.80
Algebraic Operations	-0.29	.12
Quadratic Equations	0.18	.40

We can now return to the question of whether there is evidence that lack of time on the computers, because of implementation problems, might explain the poor performance. The results shown in Table 21 indicate that there is a small negative relationship – or no relationship – between time spent and the five outcomes. Achievement as measured by the Algebraic Operations sub-strand, which also showed a negative treatment effect and no relationship to units completed, had a negative correlation with the

time spent variable. We cannot conclude that spending more time is necessarily linked to higher achievement. In addition to the correlations reported above, we conducted an analysis to see if there was a relationship between time spent on Cognitive Tutor and units completed. We found a very low correlation of .08, which also supports the notion that lack of time in the computer lab is not an explanation for low performance.

Discussion

In this first year of implementing the Cognitive Tutor Algebra I Curriculum in the Maui School District and Maui Community College, our experiment found that *Cognitive Tutor* did not have the impact on algebra achievement hoped for by the Math Science Partnership project. The results must be interpreted in the context of the particular resources for the implementation of the program on Maui, which, as our observations, interviews, and surveys revealed, were not favorable. But even taking those difficulties into account, the size of the negative impact, particularly on learning Quadratic

Equations and Algebraic Operations, was considerable. The adjusted effect sizes were -0.33 and -0.25. These are comparable in magnitude to the positive effects found in other studies.

Our communications with the teachers showed a generally positive view of *CT* and a consistent concern with implementation difficulties—particularly access to computers. However, our exploratory analysis of the data provided by the *CT* intervention on student usage gave, at best, very weak support to the lack of computer access as the explanation for the poor results. For the Algebraic Operations sub-strand, which showed a strong negative impact of *CT*, time on the program and number of problems solved were actually negatively related to scores on the outcome test. Many aspects of the implementation other than time in the computer lab may have influenced the results. For example, the late delivery of the product and the fact that the product was new may have had a disruptive effect. If the implementation and start-up problems did result in the poor performance, the results suggest that school districts should avoid undertaking this program without adequate resources and preparation time.

Since we found that the lack of access to computers was not the cause of negative results, we conducted an exploratory analysis of the data provided by the *CT* system on units completed. This indicated that there is an overall positive relationship between the number of units completed and student outcomes on the NWEA test. This also suggested that the NWEA test was a good measure of what was covered in the *Cognitive Tutor* curriculum. We also must caution, however that the nature of the outcome measure (the paper version of the NWEA end-of-course Algebra I test) makes us less confident in our conclusions about lower achieving students. First, it appears that lower achieving students were less likely to take or complete the posttest. Second, the test itself is not sensitive to students scoring at the low end and appears to distort their scores upward due to a floor effect in the paper-version of the NWEA test.

Our best indication for the negative outcome is represented by the analysis of teacher certification. While the certified teachers out performed the uncertified teachers, as one might expect, with their control class, they performed the same as the uncertified teachers with their *CT* classes. Perhaps, because both groups were equally unfamiliar with *CT* their results were the same. With the existing program, however, the more experienced teachers were far more familiar with the methods and content of the existing program than the uncertified teachers (who were equally unfamiliar with both programs). It is possible that over time, with more familiarity with *CT*, both groups may improve. Tracking these teachers into their second year with *CT* may provide some indication as to whether the experienced teachers can overcome the initial deficit.

Our goal in this research was to provide the participating district with evidence that would be useful in determining the impact of the *Cognitive Tutor* program within their local setting. Considered as a local pilot in Maui School District, the study adds to the information available on which to base their district decisions. Our study did not provide evidence of a positive impact of *CT* on student achievement in math. Although the statistical analyses yielded findings of negative or no impact, there are numerous potential explanations for this. Our recommendation is to continue the experiment into an additional year, providing adequate resources for implementation, before undertaking wider deployment.

This small study illustrates a general caution in interpreting findings from isolated experiments. Previous research has indicated that the *Cognitive Tutor* program is effective (Morgan & Ritter, 2002). This experiment demonstrates the importance of conducting multiple replication trials of any application in varying contexts and conditions. Large numbers of trials will begin to build the confidence we can have about the product and, more importantly, they will provide the multiple examples of its functioning with different populations and conditions. Then users of the research will not only have evidence of the product's average impact, but they will also be able to find contexts that are very similar to their own in order to obtain more specific guidance of its likely impact under their conditions.

References

Bloom, H. S., Bos, J. M., & Lee, S., (1999) Using cluster random assignment to measure program impacts. *Evaluation Review*, 23, 445-469.

Carnegie Learning (2006). *Cognitive Tutor*. Retrieved on June 30, 2006 from <http://www.carnegielearning.com/index.cfm>

Morgan, P., & Ritter, S. (2002). *An experimental study of the effects of Cognitive Tutor Algebra I on student knowledge and attitude*. (Available from the Carnegie Learning, Inc., 1200 Penn Avenue, Suite 150, Pittsburgh, PA 15222)

Raudenbush, S. W., (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173-185.