Empirical Education

RESEARCH REPORT

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

A Report of a Randomized Experiment in the Maui School District

Jessica Villaruz Cabalo
Boya Ma
Andrew Jaciw
Empirical Education Inc.

October 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 U.S. Department of Education. The purpose of this grant is to improve our ability to conduct small scale experiments to assist local decision-makers. The U.S. Department of Education is not responsible for the content of this report.

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.

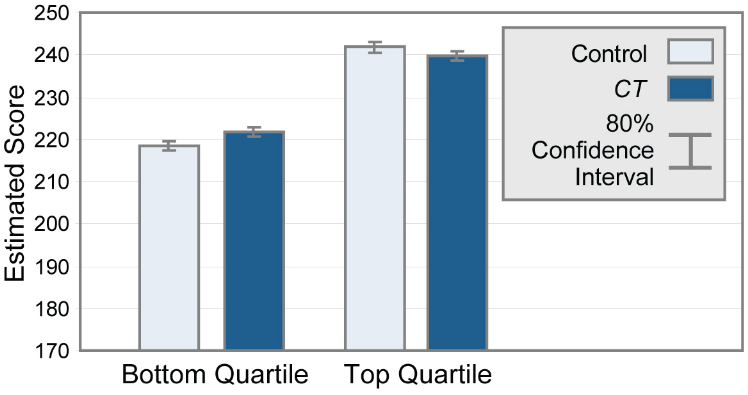
Executive Summary

Introduction. Under the *Math Science Partnership Grant*, the Maui Hawaii Educational Consortium sought scientifically based evidence for the effectiveness of Carnegie Learning's *Cognitive Tutor®* *(CT)* program as part of the adoption process for pre-Algebra programs. During the 2006-2007 school year, we conducted a follow-on study to a previous randomized experiment in the Maui School District on the effectiveness of *CT* in Algebra I. In this second year, the focus was on the newly developed *Bridge to Algebra* program for pre-Algebra. Maui's choice of *CT* was motivated in part by previous research showing substantially positive results in Oklahoma (Morgan & Ritter 2002). Our previous findings in Maui—less positive results for *CT* overall and somewhat negative results for certified teachers—called for additional study with the unique locale and ethnic makeup of Maui.

The research question was whether students in classes using *CT* materials score higher on standardized math assessment, as measured by the Northwest Evaluation Association (NWEA) General Math Test, than those in a control classroom using the pre-Algebra curricula currently in place. The district was also interested in learning whether *CT* is a teacher-friendly tool that could be used feasibly in their setting, whether there would be a differential impact on specific ethnic groups, and whether uncertified teachers would gain more from *CT* than certified teachers.

Findings. We found that most students in both *CT* and control groups improved overall on the NWEA General Math Test. We did not find a difference in student performance in math between groups. Our analysis of the Algebraic Operations sub-strand revealed that many students in both groups did not demonstrate growth in this scale, again with no discernible group differences.

However, for Algebraic Operations outcomes, we found a significant interaction between the pre-test and *CT*: students scoring low before participating in *CT* got more benefit from the program's algebraic operations instructions than students with high initial scores (see bar graph). Moreover, we noted an indication of a differential impact favoring Filipino students over White students on the Algebraic Operations sub-strand. Since the groups of interest (Filipino and Hawaiian/part-Hawaiian students) overall had lower average pretest scores, the results suggest that *CT* may help to reduce the achievement gap between those groups and others.



Differences between *CT* and Control Algebraic Operations Outcomes: Median Pretest Scores in Top and Bottom Quartiles

The district was also interested in *CT's* effectiveness for students taught by certified teachers versus non-certified teachers. In the previous year's study of *Cognitive Tutor for Algebra I*, control students of certified teachers had outperformed control students of non-certified teachers. But the program appeared to have a detrimental effect for certified teachers and no effect for non-certified teachers, both for math overall and for algebraic outcomes. In this experiment on pre-Algebra, we find certified and non-certified teachers performing about the same in their control classes. For math overall (but not the Algebraic Operations sub-strand) we find that *CT* gave the non-certified teachers an advantage.

Our goal was to provide the Maui School District with useful evidence for determining the impact of *CT* within the local setting. Considered as a district pilot, the study adds to the information available on which to base local decisions. Although our study did not provide evidence of a positive impact of *CT* on student math achievement in general, we found some positive effects. Overall, despite the repeated challenges teachers faced in implementation, *CT* was successful in raising student engagement in math and demonstrating, on the algebra-related sub-strand, gains for previously lower-performing students. The program also appeared to be particularly beneficial for non-certified teachers. Because a small number participated in the study, we consider these conclusions for teachers suggestive but not conclusive.

Design and analysis. The design of our Maui experiment was similar to the Oklahoma study, in that pre-Algebra classes were randomly assigned to *CT* or to control. We used a coin toss to assign 32 classes in five Maui schools to use the *CT Bridge to Algebra* program or to continue using the pre-Algebra program currently in place. Each of the 12 teachers involved in the experiment had equal numbers of *CT* and control classes. In their *CT* classes, they used *Bridge to Algebra* for eight to nine months until the NWEA math posttest was administered in May 2007.

The research for this experiment encompasses a multiple methods approach. We collected pre- and posttest math scores from NWEA, and class rosters and demographic information on students and teachers from the district. To measure and document implementation factors and student and teacher interactions with the materials, we also collected qualitative data through classroom observations, phone interviews, and web-based surveys from teachers.

Because our findings differed from those in Oklahoma, this small study illustrates a general caution in interpreting findings from isolated experiments and 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. Here, it is important to interpret the results in relation to what teachers were using in control classes and to the usage patterns, implementations, and applications in *CT* classes. It is also relevant that half the *CT* teachers were using *CT* for the first time; and their initial unfamiliarity may have affected implementation. Finally, the size of this experiment precluded detection of small differences.

Overall teacher impressions. Our qualitative data sources revealed that teachers experienced similar resource challenges in implementing *CT* as in the Algebra study: lack of classroom computers, access to the computer lab, and *CT* materials. Another challenge related to the misalignment between *CT* content and state math standards in middle and high school. Despite these challenges, teachers (and students) reported a generally positive attitude about *CT* overall. Teachers were particularly pleased with how engaged their students were with the *CT* software and the *CT* approach to collaborative learning. (It must be noted that 45% of the teachers reported that this approach affected instructional practices in their control classes. We were not able to determine whether this contamination made a difference in outcomes.)

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

A Report of a Randomized Experiment in the Maui School District

Table of Contents

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 pre-Algebra programs to be considered for adoption. The U.S. 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.

This is a follow-on study to a previous experiment conducted in the Maui School District. In year one, the district wanted to study the effectiveness of Carnegie Learning's *Cognitive Tutor* program in Algebra I. In this second year of research, the district wanted to study the effectiveness of Carnegie Learning's newly developed *Bridge to Algebra* program, a pre-Algebra curriculum. 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) General Math Test, than they would if they had been in a control classroom using the pre-Algebra curricula the Maui schools currently have in place. We also sought to explore the Algebraic Operations sub-strand of this test because the program under study focused on preparation for Algebra.

The experiment started in September of the 2006-2007 school year. We conducted the experiment in five schools in the Maui School District. For the 12 participating teachers, we randomly assigned each of their pre-Algebra 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 eight to nine months during the 2006-2007 school year until the NWEA posttest in math was administered in May 2007.

In addition to their interest in whether *CT* has an impact on student achievement, the district was particularly interested in how much or how little *CT* was implemented as compared to *CT* implementation in the Algebra study, and in how satisfied teachers were with *CT.* They were also interested in whether there would be a differential impact on specific ethnic groups, specifically Native Hawaiians and Filipinos, and whether uncertified teachers would gain more from the *CT* program than certified math teachers.

The choice of the Carnegie Learning *CT* program was motivated in part by previous research that had shown positive results for the Algebra I program. For example, an experiment reported by the publisher showed that the impact of *CT* for Algebra I was substantial (Morgan & Ritter 2002). This research was conducted in a 19,000 student school district in Oklahoma with an ethnic mix of students that included 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 considerable. This metric for effect size gives us a way to standardize across studies that use different outcome measures. While the Oklahoma study showed positive results, our own previous research on the Algebra I program in Maui showed less positive results overall and somewhat negative results for certified teachers. Additional research with the unique locale and ethnic makeup of Maui was necessary.

The design of our experiment in Maui was similar to the Oklahoma study, in that pre-Algebra 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

Our experiment is a comparison of outcomes for classes where *Cognitive Tutor Bridge to Algebra* was in place (the *CT* group) and classes using the regular pre-Algebra curriculum (control group). The outcomes of interest are the student test scores in math, specifically algebra.

This section details the methods used to assess, with some level of confidence, the size of the difference between the *CT* and the control groups (within confidence limits set by the available sample size) and whether the introduction of *CT* was responsible for those differences. We begin with a description and rationale for the experimental design and go on to describe the intervention, the research sites, the sources of data, the composition of the experimental groups and finally the statistical methods used to generate our conclusions about the impact of *CT*.

Experimental Design

With experiments we usually randomize an available sample of cases. Generalization is left to heuristic arguments which include a comparison of the characteristics of the sample with that of the population of interest (e.g., the whole district.) Though we don't have the luxury of randomly sampling cases to be randomized, our results need to express the fact that our sample is just a select group of cases, and the results we get would change if a new sample of teachers or students was selected into the experiment by whatever mechanism. The design of the experiment is based on our best understanding of the amount of variability that we expect due to re-sampling, where our intention is to limit the effect of this 'noise' in order to detect the stable signal (the effect) if it exists. There is always a level of uncertainty and an associated level of imprecision. We think of the uncertainty as related to the likelihood that we would get a different result if we took a new sample of students or of teachers from the same larger population. Our design attempts to efficiently deploy the available resources to reduce uncertainty and improve precision, in other words, to reduce the likelihood that we would get a different result if we tried the experiment again.

An up front effort to fully specify a design or plan for the experiment pays off in two ways:

- First, we identify, before seeing the outcomes, where we expect to see an impact and what factors we expect will moderate the impact. In other words, we specify the research questions up front. In this way, we avoid fishing for results in the data, a process that can lead to mistaking chance differences for differences that are probably important as a basis for decisions. Because some effects will be big simply by chance, "'mining" the data in this way can capitalize on chance—we conclude that there is an effect, when really we're just picking the outcomes that happen to big as a result of chance variation. We can still explore the data after the fact but this is useful mainly for generating ideas about how the new program worked, that is, as hypothesis-generating efforts for motivating future study, rather than as efforts from which we make firm conclusions from our existing study.
- Second, an experimental design will include a determination of how large the study should be in terms of students, teachers and schools in order to get to the desired level of confidence in the results. In the planning stage of the experiment we calculate either how many cases we need to detect a specifically sized difference between the *CT* and control groups, or how big a difference we can detect given the sample size that is available. Technically this is called a power analysis. We will explain how many aspects of design determine the size of the experiment.

How the Sample was Identified

How the participants for the study are chosen largely determines how widely the results can be generalized. In this case, principals who had volunteered to support the use of *CT* had invited a select number of teacher volunteers based on capacity and funding within each school as well as prior experience and use of the *CT* program. All teachers who had volunteered to use *CT* in their classrooms were invited by the district through an email and inter-campus memorandum.

The initial meeting for the experiment was conducted on June 22, 2006 and attended by 17 teachers and several administrators from the district. The meeting included training for the *CT* program, an explanation of the *CT* study, 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 consent and background information form. Nine out of the 17 teachers who attended this initial training participated in the study.

Eight teachers who attended the training did not participate in the study. Three of these teachers did not give a reason for not participating, while the other five gave the following reasons: 1) two teachers taught Special Education, 2) one teacher taught at an alternative learning center, 3) one teacher taught an atypical mix of classes, and 4) one teacher had a substitute for the Fall semester.

As is displayed in Table 1, the background information forms revealed that among the 13 teachers who participated, there was a wide range of experience in teaching pre-Algebra. Eight teachers had 0-3 years of experience, two teachers had 4-6 years, and three teachers had 7-15 years, whereas no teacher had more than 16 years of experience teaching pre-Algebra.

Table 1. Teaching Background

Survey question	0-3 years	4-6 years	7-15 years	16+ years
How many years total have you taught?[a]	31%	8%	39%	23%
How many years have you taught math?[b]	46%	8%	31%	15%
How many years have you taught pre-Algebra?[c]	62%	15%	23%	0%

Note. There were 13 teachers who responded to each of these questions.

[a]Minimum value = 0, maximum = 36, mean = 10.25

[b]Minimum value = 0, maximum = 18, mean = 6.85

[c]Minimum value = 0, maximum = 12, mean = 3.46

Randomization

Since we want to know the impact of *CT*, we have to isolate its impact from all the other factors that might make a difference in how or what schools, teachers, and students do. We want to answer whether *CT* caused a difference. Randomization ensures that, on average, characteristics other than the program that affect the outcome are equally distributed between program and control groups. This distribution prevents us from confusing the program's effects with some other factors, technically called "confounders," that, because they also affect the outcome, would lead to bias if they are unevenly distributed between the groups. For example, randomization helps to ensure that classes of lower scoring students are not selectively assigned to the *CT* or control group.

Organizational Levels Considered in the Experiment

This research works within the organization of schools by not disrupting the existing hierarchy in which students are grouped under teachers who belong to schools. The level in the hierarchy at which we conduct the randomization is generally determined on the basis of the kind of intervention being tested. School-wide reforms call for a school-level randomization while a professional development program can use a teacher-level randomization. Generally, we attempt to identify the lowest level at which the intervention can be implemented without unduly disrupting normal collaboration and without inviting sharing or "contamination" between control and program units. In this experiment, for all of the teachers who volunteered to participate, we randomized classes in approximately equal numbers to the *CT* and control groups. While there is potential for some carry over for the teachers between their teaching in the *CT* and control classes, we believed that the differences in textbooks and especially the restriction of login to the *CT* students provided a sufficient barrier between the two conditions. The outcome measures are student-level test scores in on the NWEA General Math test. Because classes, instead of students, were assigned to *CT* or control, this kind of experiment is often called a “group randomized trial” (Bloom, Bos, & Lee, 1999; Raudenbush, 1997).

What Factors May Moderate the Impact of *CT*?

Our design allows us to consider the extent to which *CT* is more effective for students of certain ethnic backgrounds and for students with certified versus uncertified teachers. These are variables that are measured before the experiment starts, and that we have reason to believe will affect the strength of the effect of *CT*. Technically, these are called potential moderators because they may moderate the impact of *CT*. We measure the strength of the interaction between each moderator and the *CT* effect; that is, we measure whether the effect of *CT* changes as the level of the moderator changes.

How Large a Sample Do We Need?

A process called power analysis was used to plan the number of classes that the experiment will need in order to say with any confidence that the intervention has an impact of a certain size. This is an important part of experimental design and here we walk through the factors considered.

How Small an Impact Do We Need?

The size of the sample needed depends on how small an effect we need to detect. Experiments require a larger sample to detect a smaller impact. It is very important to make an educated guess as to the range of impact a program like the one being tested typically has. From a practical point of view it is also important to know the smallest potential impact that would be considered educationally useful. As a hypothetical example, using percentile ranks as the measure of impact, we may predict that an intervention of this type can often move an average student 15 percentile points. As a sensible matter for educators, however, an improvement as small as 10 percentile points may have practical value. The researcher may then set the smallest effect of interest to be 10 points—the intervention may do better, but if it makes less than a 10 point difference, the practical value will be no different than zero. We can call this the “minimum required effect size”. It is necessary to decide on this value as part of the power analysis since the number of units needed in the sample is related to how small an effect we need to detect. Conversely, with a particular number of units available, we want to know how small an effect we can detect—the so-called “minimum detectable effect size” (MDES). In some cases, there may be positive effects that we can’t detect because they are lower than the MDES.

For the current experiment the design and sample size was adequate for an MDES of 11 percentile points or, in terms of the standard deviation units we introduce below, an effect size of 0.29.

How Much Variation is There Between Classes?

When we randomize at the class level, but the outcome of the interest is a test score of students associated with those classes, we pay special attention to the differences among classes. The greater the differences among those units, the more units we need in the experiment to detect the impact of the intervention. This is because the extra variation among classes adds noise to our measurement, which makes the effect of the intervention, the signal, harder to detect. A larger sample allows us to effectively reduce the level of the noise. If the differences among classes are large and/or the differences within them are small, then the sample size that matters the most for the experiment is the number of classes. If the differences among classes are small so that most of the variation is attributable to differences among students within them, then the sample size that matters most is the number of students. A summary statistic that tells us how the variation is divided up among levels of analysis is the intraclass correlation (ICC). Technically it is the ratio of the variation in the outcome among classes to the total variation. We assume that this is computed before the intervention. For this experiment we assumed an intraclass correlation of .15.

Randomization by Pairs

There are various ways to randomize classes to experimental conditions. For this study, we used a matched-pairs design where we identified pairs of similar classes. First, we consider what the critical characteristics of classes are that we believe affect performance. We use this information to pair classes and then we randomize the members in each pair to the two conditions. Technically, this is a form of blocking and it usually increases how much certainty we have in the difference in the posttest scores that we measure between the *CT* and control groups. In this experiment, classes were matched based on class size and achievement level. Sixteen pairs of classes were assigned using a coin toss to either the *CT* condition or to control.

How Much Value Do We Get From a Pretest?

In order to gain additional precision, we make use of other variables that we know will impact performance. In our experiments, a student's score on a pretest (which may be a test in a subject that is closely related to the outcome measure rather than the same test but given earlier) is almost always the variable most closely associated with the outcome. In this case, the pretest is a "covariate". By including the covariate we can increase precision by "removing" this source of variation in the results. Technically, a covariate-adjusted analysis is called an analysis of covariance (or ANCOVA). In almost all of our analyses we adjust for the effect of the pretest, which is a strong predictor of posttest performance. In this experiment, we assumed a fairly substantial correlation between the pre- and posttests (.80)[1]. In a power analysis determining the number of classes we will need, a good pretest correlation will increase precision and thereby require fewer classes to detect the same level of impact.

Are There Subgroups of Particular Interest?

Often we are interested in whether a program has more impact for a particular student subgroup than others or for a certain group of teachers than others. Where the subgroup is identified within each randomized unit—that is, where each randomized unit has some portion of that subgroup—the impact on the power analysis is minimal. However, if our subgroup of interest is a subtype of the unit of randomization, then, in most cases, we will need to include additional units in the experiment in order to have enough units of each type. In the current experiment, we are interested in ethnic subgroups of students. We were also interested in classes taught by certified teachers versus non-certified teachers. Examining this characteristic depended on a small number of teachers and, while important, can only be considered exploratory.

How Confident Do We Want to be in the Results?

We have described uncertainty in terms of the likelihood that, if we ran the experiment again with a different sample from the same district, we would get the same result. Although results are never exactly identical, we can design the experiment so that the various results we would get would be similar. This scenario is hypothetical because we are not likely to run exactly the same experiment multiple times. An experiment that produces a very high level of confidence that the results of multiple experiments would be very similar requires a larger number of units than an experiment that produces a lower level of confidence or a wider range of likely outcomes for the other hypothetical experiments. Still, we can never be entirely certain of a result. Thus the final step in the power analysis is to determine an acceptable or tolerable level of uncertainty. Conventionally, researchers have called for a high level of certainty, specifically, that getting a result like that observed would happen in only 5% of instances if the program did

[1] That is, we assume that .80*.80=.64 is the proportion of variance in the outcome (i.e., the R-squared) that is accounted for by the covariate, in either condition.

not indeed have an impact. For the purpose of the power analysis for this experiment, we used the 5% criterion although, as we explain later, we report the results using a range of confidence levels.

Sample Size Calculation for This Experiment

Taking all the above factors into consideration, we estimated that 39 classes would constitute a sufficiently large sample to detect a difference as small as 11 percentile points. This means that if the 50th percentile pre-Algebra student in the program group were placed at the end of the experiment in control he or she would be that many points higher (or lower) than the 50th percentile student in the control group. As we explain later in this section, we can also express these as standardized effect sizes or portions of a standard deviation. In that metric the MDES for math is 0.29.

As with the first study, we found 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 representatives in consultation with the researchers decided to move forward with the experiment. We conducted the randomization by class, within teacher, such that each teacher had both *CT* and control classes. By randomizing in this way rather than randomizing among the teachers, we maximized the number of units in the analysis.

Intervention

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 skillometer,[2] 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 graphic representations to solve problems.

Nine out of the twelve *CT* teachers attended three days of professional development led by a Carnegie Learning consultant and received their *CT* materials after the initial meeting. The three *CT* teachers who did not attend the initial training attended a make-up training and received the *CT* materials afterwards. Beyond the initial training, teachers were free to make use of the materials as best suited the needs of their classrooms and students.

Existing Math Program

Survey data revealed that 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 used a variety of pre-Algebra I textbooks from the following publishers: McDougal Littell (a Houghton Mifflin company), Holt, Prentice-Hall, Addison-Wesley, Scott Foresman, and Merrill. During the study, the control group classes continued using these materials as usual.

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 U.S. Census Bureau, the total population in 2006 was 141,300.

[2] Skill bars show students what skills they have mastered and where they need to improve in order to motivate them to take responsibility for their own learning.

Of the adult population, 87.9% have a high school diploma and 26% have a Bachelor's degree. For people reporting one race alone 48 percent were White, 36% Asian, 13% Native Hawaiian and Other Pacific Islander, 2% Other, 1% Black or African American, and fewer than 0.5 percent American Indian and Alaska Native. Twenty-two percent reported two or more races. Nine percent of the people in Maui County were Hispanic. Thirty-four percent of the people in Maui County were White non-Hispanic. People of Hispanic origin may be of any race.

Maui District Schools

The Maui School District covers the Molokai and Lanai school systems. It is the second largest school complex in the state of Hawaii 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 2005-2006 school year. The average ethnic breakdown for the participating schools includes approximately 32% Filipino, 28% Part-Hawaiian, 11% White, 7% percent Japanese, 5% Hawaiian, 3% Hispanic, and 9% Other. An average of 33% of students participated in the National School Lunch Program, while 15% were in Special Education, and 6% were designated as Limited English Proficient.

Data Sources and Collection

The research for this experiment encompasses a multiple methods approach. We collected pre- and post-intervention math scores from NWEA, and class rosters and demographic information on students and teachers from the district. All student and teacher data having any individually identifying characteristics were stripped of such identifiers, and the data were stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA).

We also collected implementation data through the use of classroom observations, phone interviews, and web-based surveys from all participating teachers. We integrated all the information from these multiple sources into a standard data warehouse for the study.

Achievement Measures

The primary pretest and posttest measures, the Northwest Evaluation Association (NWEA) General Math Test (ALT), were administered to the students in the Maui schools in October 2006 and again in May 2007.The test measures achievement in algebra, computations, number sense, geometry, measurement and statistics. It is an adaptive and comprehensive test that reflects the instructional level of each student and measures growth over time. The set of tests consist of multiple levels, with overlapping degrees of difficulty. Prior to administration of the first assessment, a student's appropriate test level is determined by use of a 20-item placement test, referred to as a locator test. For subsequent administrations of the ALT, the student is automatically assigned to a level based on previous results. The tests are scored on a Rasch unIT (RIT) scale, which measures 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. We analyze the overall scale and the sub-strand for Algebraic Operations in the investigation of the impact of *CT*.

Observational and Interview Data

In addition to quantitative data, we also collected qualitative data over the entire period of the experiment, beginning with the randomization meeting and ending with the academic calendar of the district in June 2007. Training observations, classroom observations, informal and formal interviews, multiple teacher surveys, email exchanges, and phone conversations are used to provide both descriptive and quantitative evidence of the implementation.

In general, observational data are used to inform the description of the learning environment, instructional strategies employed by the teachers, and student engagement. These data are minimally coded.

Classroom observations occurred in September and November 2006. Their purpose was to help us understand and document 1) student and teacher interactions with the *CT* materials and 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. We used a standard observation protocol while conducting classroom observations. Our selection of which classes we observed and the length of time for each observation were determined based on the class schedules and time constraints. Some classes occurred simultaneously so we had to go to two classes for 30 minutes each within a single period. Observations ranged in time from 30-50 minutes. In September, a total of four control classes and three *CT* classes in two schools were observed. In November, a total of four *CT* classes in two schools were observed. These classes were visited by a research manager from Empirical Education and, occasionally, by the study's point of contact from the Maui School District.

These observational data, in combination with what we found in the Algebra study, helped us formulate questions for the phone interviews and web-based surveys.

Interview data are used to elaborate survey responses, characterize the teacher's schedule, and to provide descriptions of the overall experience teaching with the *CT* program. Structured phone interviews were conducted with the teachers and with a representative from Carnegie Learning in January 2006. While all teachers were contacted to schedule a phone interview, only six teachers responded. Each interview lasted between 15 and 20 minutes. The purpose of these interviews was to gain an understanding of teachers' attitudes and opinions about *CT* 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).

Survey Data

The quantitative survey data are reported using descriptive statistics; these are summarized by individual teacher and by assignment group (*CT* and control), and are compared by group. Survey data are used to quantify the extent of exposure to the materials (opportunities to learn with the curriculum). The free-response portions of the surveys are minimally coded.

Fifteen web-based surveys were administered to all participating teachers on a bi-weekly schedule from September to April of 2007. We obtained a 100% response rate on all 15 surveys. The content of these surveys covered factors we had identified as possibly influencing *CT*'s implementation, and therefore the results of our study. Table 2 lists the survey dates and the main topic of each survey. In addition to these topics, all surveys asked teachers to report instructional time spent on each program as a way to document program implementation. A final survey addressed questions about teachers' overall experience with the *CT* program as well as the specific units covered throughout the study.

Table 2. Survey Dates and Topics

Survey	Date	Topic
TB	Sept. 1	Teacher Background
CC	Sept. 15	Classroom Context
Survey 1	Sept. 29	Technology Behavior and Attitude
Survey 2	Oct. 13	Existing Math Program
Survey 3	Oct. 27	Program Content
Survey 4	Nov. 10	Planning and Preparation
Survey 5	Nov. 17	Technology Resources
Survey 6	Dec. 15	Assessments
Survey 7	Jan. 19	Interactions with Materials
Survey 8	Jan. 26	Professional Development
Survey 9	Feb. 9	Teacher and Student Collaboration
Survey 10	Feb. 23	Computers and Technology
Survey 11	Mar. 9	Hawaii State Assessment
Survey 12	Mar. 13	Program Content
Survey 13	Apr. 27	Final Survey

Rationale for our Indicators of the Extent of Implementation

Based on what we learned about the local context in the Algebra study and on the needs and interests of the district, we expanded our measures of implementation. The district was particularly interested in how much or how little *CT* was implemented as compared to the *CT* implementation of the Algebra program, and in how satisfied teachers were with *CT.* District personnel were also interested in student engagement because they felt that it is key for improved math achievement. They wanted to know how the *CT* program changed student engagement, if at all.

In the Algebra study, we had collected qualitative data about teacher background, teacher access to and use of materials, professional development and planning, student engagement, collaboration, assessments, and teacher satisfaction with materials. In the pre-Algebra study, we measured the same variables as the in Algebra study but added the Hawaii State Assessment, curricular content and progress, and teacher comfort with technology as strong indicators of program implementation. We also made considerable changes in the content of our survey questions as well as our surveying strategy. In addition, we surveyed teachers about their existing math curricula in order to have balanced information about the *CT* and control programs. In most surveys, we asked questions about the existing math program that were comparable to our questions about the *CT* program. Below we list each category and our rationale for measurement during this pre-Algebra study.

Teacher Access to and Use of Materials

In the Algebra study, we found that lack of access to resources was the main challenge in implementing the *CT* program. Specifically, teachers cited two limitations: 1) receipt of *CT* materials was inconsistent among teachers and was sporadic throughout the year and 2) teachers had varying and inconsistent access to networked computers to use the *CT* software. Therefore, in this pre-Algebra study we continued to measure teacher access to resources.

We measured teacher and student use of the *CT* curricular material. We specifically measured time spent on the *CT* software and time spent with the *CT* textbook (as expressed in percentages). This measurement of time helped us understand how close each class came to Carnegie Learning's suggested implementation of 60% *CT* Text and 40% *CT* software. The need to follow this ratio was reinforced in the *CT* materials as well as in training.

For comparison, we also surveyed teachers on their access to and use of materials for their existing math program.

Planning

It was important to know how much time teachers spent on planning for the *CT* program and their existing math program. We measured average preparation time and lesson planning in a single survey as well as time spent in training and/or accessing additional support from Carnegie Learning and the publisher of their existing math program.

Student Engagement

We measured student engagement by asking teachers to rank the level of student engagement since the introduction of the *CT* program or existing math program for each class. Teachers also provided free response of both positive and negative interactions their students had with the *CT* materials and existing math program.

Teacher Collaboration

The extent of collaboration is an indicator of the degree to which teachers adhered to the *CT* curriculum. Collaborative learning, as indicated in the *CT* materials, is central in the Carnegie Learning philosophy. Therefore we measured the extent to which teachers collaborated with one another and whether teachers enforced collaboration among their students in the *CT* and control classes.

Hawaii State Standards

Hawaii State Assessment

We decided to gather information on teacher preparation for the Hawaii State Assessment (HSA) based on information gathered during the Algebra study. The project POC told researchers that one teacher wanted to discontinue her participation in the study. We found that this particular teacher, along with others, was not continuing implementation of the *CT* program because they felt that the *CT* curricular content was not aligned with the content covered in the HSA. Instead, they used other materials they felt would better prepare their students for the HSA. Therefore, in this pre-Algebra study we measured how well prepared teachers thought their students were for the HSA.

Curricular Content and Progress

Similar to our reason for obtaining information about the HSA, we asked teachers to report the content they covered in each class as a way to assess the alignment of the *CT* program to the Hawaii standards in math. We measured the progress of each class through the content as a possible explanation for any variation in math achievement among classes.

Teacher Comfort with Technology

The district also expected that, since the *CT* program is a technology-integrated curriculum, program implementation is dependent on teachers being comfortable with technology. District personnel also felt that this was important to measure in order to pace the deployment of *CT* in additional classes and schools. Therefore we surveyed teachers on various aspects of technology to help determine teachers' comfort and confidence levels. These include a self-rating of computer skills, use of computers and Internet, confidence in computer use, confidence in use of *CT,* and interest and motivation in learning computer technologies

Teacher Satisfaction with *CT* and Existing Math Program

Finally, we asked teachers about their overall experience and satisfaction with both the *CT* program and their existing math program. We measured this through free response as well as through scales rating levels of satisfaction for each component of the program. The district felt strongly that teacher satisfaction had a considerable impact on program implementation.

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 3 shows the counts of schools, teachers, classes, grades, and students between *CT* and control conditions. Figures in these tables reflect the full number of students in the experiment at the time it began in August 2006.

Table 3. Distribution of *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 7	Grade 8	Grade 9	Grade 10	Grade 12	
CT	5	12	16	78	125	193	4	1	**401**
Control	5	12	16	76	122	207	1	2	**408**
Totals[a]	**5**	**12**	**32**	**154**	**247**	**400**	**5**	**3**	**809**

[a] Some teachers taught more than one class in each group.

Though the experiment began with 809 students, one teacher decided to discontinue participation in the study, leading to a loss of 103 students. This included 52 control students and 51 *CT* students. The following calculations are based on the remaining 706 students.

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

There were two category schemes for ethnicity available for students in this study. Table 4 displays the ethnic makeup of the study participants using the standard categories used in U.S. census data. We can see that a majority of the participants were categorized as Asian and that there is no further breakdown in this scheme. The high *p* value from Fisher's exact test indicates balance on ethnicity between the two conditions.

Table 4. Ethnicity by Typical U.S. Categories for *CT* and Control Groups

Condition	Ethnicity						
	Asian	Hispanic	Native American	Mixed ethnicity	Black	White	Total
Control	197	13	3	95	1	34	343
CT	181	15	3	102	2	33	336
Total	**378**	**28**	**6**	**197**	**3**	**67**	**679**

Statistics	Value	*p* value
Fisher's exact test	<0.01	.93

Note. We are missing this information for 27 students.

A second categorization scheme was used by the Hawaii Department of Education. For all but 80 of the students (18 were missing a designation and 62 were designated as Other), we obtained more detailed ethnic categories that distinguished among the Asian ethnicities. These categories provide a different picture and allow us to identify the ethnicities of particular interest to the project—Filipino students and Hawaiian students (including part-Hawaiians). As we observe in Table 5, these are large groups and, according to both the NWEA Math test overall score and the sub-strand for Algebraic Operations, relatively low scoring on the pretest.

Table 5. Average Pretest Score by Ethnic Group

State ethnicity categories	Number of students	NWEA General Math Test	NWEA Algebraic Operations sub-strand
Filipino	252	216.35	220.46
Part-Hawaiian	163	219.67	223.50
Other	79	216.27	221.05
White	47	223.08	225.03
Japanese	43	224.41	226.08
Spanish, Cuban, Mexican, Puerto Rican	28	215.05	218.09
Hawaiian	29	214.90	221.85
Missing	27	217.85	222.77
Portuguese	15	224.36	225.79
American Indian	6	*	*
Chinese	3	*	*
Samoan	5	*	*
Black	3	*	*
Korean	3	*	*
Indo-Chine	2	*	*

*Mean scores are not provided in order to preserve anonymity.

For ease of analysis, we split the population into four categories that highlight the major categories—Filipino, Hawaiian (including part-Hawaiian), White, and Other—found in the table. Note that mean scores are not provided for categories with fewer than five members. Table 6 shows the counts for each of these four categories broken down into *CT* and control groups. Ethnic categories are reasonably well distributed between the two conditions. These categories will be used in later investigations of differential impact of *CT*.

Table 6. Ethnic Groups of Local Interest for *CT* and Control Groups

Condition	Ethnicity			
	Filipino	Hawaiian/ Part-Hawaiian	White	Other
Control	163	109	26	94
CT	147	98	24	115
Total	**310**	**207**	**50**	**209**
Statistics	**Value**	***p* value**		
Chi-square	3.52	.32		

Teacher Certification

We observe in Table 7 that teacher certification is distributed evenly between the conditions. This is confirmed by Fisher's exact test.

Table 7. Teaching Certification for *CT* and Control Groups

Condition	Non-certified	Certified	Totals
Control	2	4	**6**
CT	3	2	**5**
Totals	**5**	**6**	**11**
Statistics		**Value**	***p* value**
Fisher's exact test		0.32	.57

Note. The teacher who discontinued participation in the *CT* study is excluded.

Achievement Pretests

With randomization, we expect the pretest scores to be equally distributed between *CT* and control groups. As we observe in Table 5, the pretest scores between the two groups are balanced. This result is confirmed when we model the difference, controlling for clustering. In the analyses that follow, we add the pretest covariate in order to increase the precision of the impact estimate. (We note, however, that our impact estimates are unbiased due to randomization, whether or not we factor in the effect of the pretest.)

Table 8. Independent *t* Test of the Difference between Students in *CT* and Control Groups for the NWEA General Math Test

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students	Standard error	Effect size[a]
Control	217.84	12.72	284	0.76	-0.09
CT	219.01	12.22	285	0.72	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
Condition (*CT* – control)	1.17		567	-1.12	.26

[a] The difference we are measuring is not an effect of treatment (the usual sense of effect size) but a result of chance differences in the randomization.

Attrition

Out of a total enrollment of 809 based on fall class rosters, 706 students are considered active, meaning, these students stayed in the experiment from the beginning to the end. Of these, 137 students (or 19%) did not have pretest scores. Of the remaining 569 students, posttest scores are missing for 93 or 16%. Table 9 shows the breakdown by the *CT* and control groups. A Chi-square test indicates no relationship between the rate of 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 137 active students did not have a pretest score.)

Considering all categories of missing data, 333 of 809 or 41% of enrolled students could not be used in the analysis. This may in part be due to students being absent on testing days or the fact that students 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 10 shows that students with a pretest and no posttest scored about the same as students who had both scores. We provide this result as part of the test for differential attrition. A test that controls for clustering yields an even more conservative result with a larger *p* value.

Table 9. Counts of Students Missing Test Score Data

Condition	Having both pre- and posttest scores	Having pretest but missing posttest scores	Totals
CT	237	47	**284**
Control	239	46	**285**
Totals	**476**	**93**	**569**
Chi-square statistics	**DF**	**Value**	***p* value**
	1	0.02	.90

Table 10. Independent *t* Test of the Difference in Pretest Scores between Students with 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[a]
Have pretest scores only	219.16	13.96	93	1.45	0.07
Have both pre- and posttest scores	218.28	12.18	476	0.56	
***t* test for difference between independent means**	**Difference**		**DF**	***t* value**	***p* value**
(Missing posttest) – (Have posttest)	0.88		567	-0.62	.53

[a]The difference we are measuring is not an effect of treatment (the usual sense of effect size) but a result of chance differences in the randomization.

In sum, our tests reveal balance on variables that are expected to affect performance, and lack of differential attrition.

Statistical Equations and Reporting on the Impact of *CT*

Setting Up the Statistical Equation[3]

We put our data for students, teachers, and classes into a system of statistical equations that allow us to obtain estimates of the direction and strength of relationships among factors of interest. The primary relationship of interest is the causal effect of the program on a measure of achievement. We use SAS PROC MIXED (from SAS Institute Inc.) as the primary software tool for these computations. The output of this process are estimates of effects as well as a measure of the level of confidence we can have that the estimate is true of the population to which the experiment is meant to generalize.

Program Impact

A basic question for the experiment was whether, following the intervention, students in *CT* classrooms had higher math scores than those in control classrooms. Answering this is not as simple as comparing the averages of the two groups. The randomization gave us two groups that are equivalent to each other on average in every way, except that one receives *CT* and the other one does not. But as we saw in the section on the formation of the experimental groups, in a single randomization we expect chance imbalances. Adjusting for these random differences gives us a more precise measure of the program's effect. It is also essential that we understand how much confidence we can have that there really is a difference between the two groups,

[3] The term 'statistical equation' refers to a probabilistic model where the outcome of interest is on the left hand side of the equation and terms for systematic and random effects are on the right hand side of the equation. The goal of estimation is to obtain estimates for the effects on the right hand side. Each estimate has a level of uncertainty which is expressed in terms of standard errors or *p* values. The estimate of main interest is for the treatment effect. In this experiment, we model treatment as a fixed effect. With randomized control trials, the modeling equation for which we are estimating effects, takes on a relatively simple form: Each observed outcome is expressed as a linear combination of a treatment indicator, one or more covariates that are used to increase the precision of intervention effect, and usually a series of fixed or random intercepts, which are increments in the outcome that are specific to units. As a result of randomization, the other covariates are distributed in the same way for both the treatment and control groups. For moderator analyses we expand these basic models by including a term that multiplies the treatment indicator with the moderator variable. The coefficient for this term is the moderator effect of interest.

given the size of the effect estimate that we obtain. To appropriately estimate this difference, our equation contains a term for *CT* as well as terms for other important factors such as the student pretest score. The student's prior score, is of course, an important factor in estimating his or her outcome score. By including pretest as a term in the statistical equation, we are able to improve the precision of this estimate because it helps to explain a lot of the variance in the outcomes and makes it easier to isolate the program impact. We also have to account for the fact that students are clustered by classes and teachers. We expect outcomes for students who are in the same class or who have the same teacher to be dependent as a result of shared experiences. We have to add this dependency to our equation or else our confidence levels about the results will be artificially high.

Covariates and Moderators at the Student and Teacher Level

In addition to estimating the average impact, we also include in the equation other variables (called covariates) associated with characteristics of the teachers and students, which we expect to make a difference in the outcomes for the students. For example, as was described above, we add the pretest score into almost all our statistical equations in order to increase precision. In addition, we consider whether there is a difference in the effect of the intervention for different levels of the covariates. For example, we consider whether the program is more effective for higher-performing students than for lower-performing students. We estimate this *difference* (between subgroups) *in the difference* (between the program and control groups) by including an interaction term in the statistical equation. This term multiplies together the variable that indicates whether the student is in the intervention group, and the covariate. We call covariates, that are included in such analyses, potential "moderators" because they may moderate—either increase or decrease—the effect of the program on student outcomes. The value for the interaction term is a measure of the moderating effect of the covariate on the effect of the program.

Fixed and Random Effects

The covariates in our equations measure either 1) fixed characteristics that take on a finite set of values (e.g., there are only two levels of gender); or 2) a set of characteristics that is assumed to have a distribution over a population and where we treat the values that we measure as though they were a random sample from that larger population. The former are called "fixed effects", the latter, "random effects". Random effects add uncertainty to our estimates because they account for sampling variation, or the changes we would observe in the outcomes if we re-sampled units from the same population. Fixed effects produce less uncertainty but also limit the extent to which we can generalize our results.

We usually treat the units that were randomized as "random effects", so that in the statistical equations, our estimates reflect the degree of uncertainty that comes if we were to draw a different sample of units from the same population[4]. This allows us to argue for the generalizability of our findings from a sampling perspective. Treating the units that were randomized as fixed, forces us to use other arguments if our goal is to generalize.

Using random or fixed effects for participating units serves a second function—it allows us to more accurately represent the dependencies among cases that are clustered together (e.g., students in classes.) All the cases that belong to a cluster share an increment in the outcome--either positive or negative—that expresses the dependencies among them. An appropriate measure of uncertainty in our estimate of the program's effectiveness takes into consideration whether there is more variation *within* the larger units or *between* them. All of our statistical equations include a student-level error term. The variation in this term reflects the differences

[4] Although we seldom randomly sample cases from a broader population, and in some situations we use the entire population of cases that is available, we believe that it is still correct to estimate sampling variation (i.e., model random effects). It is entirely conceivable that some part or the whole set of participants at a level end up being replaced by another group (for whatever reason) and it's fair to ask how much change in outcomes we can expect from this substitution.

we see among students that are not accounted for by all the fixed effects and other random effects in our statistical equation.

The choice of terms for each statistical equation is not rigid but depends on the context and the importance of the factors for the question being addressed. The tables reporting the estimates resulting from the computation will provide an explanation of these choices in table notes where necessary for technical review.

Reporting the Results

When we run the computations on the data, we produce several results: among them are effect sizes, the estimates for fixed effects, and *p* values. These are found in all the tables where we report the results.

Effect sizes

We translate the difference between program and control groups into a standardized effect size by dividing the average group difference by the amount of variability in the outcome. The amount of variability is also called the “standard deviation” and can be thought of as the average distance of all the individual scores from the average score (more precisely, it is the square root of the average of squared distances.) Dividing the difference by the standard deviation gives us a value in units of standard deviation rather than units of the scale used by the particular test. This standardized effect size allows us to compare the results we find 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 score and other fixed effects, since that adjustment provides a more precise estimate of the effect by compensating for chance differences in the average pretest of the program and control groups. Theoretically, with many replications of the experiment, these chance differences would wash out so we would expect the adjusted effect size on average to be closer to the true value.

Estimates

We provide estimates to approximate the actual effect size. Any experiment is limited to the small sample of students, teachers, and schools that represent a larger population in a real world (or hypothetical) setting. Essentially we are estimating the population value. When we report an estimate in a table, the value refers to the change in outcome for a one-unit increase in the associated variable. For example, since we code participation in the control group as 0, and participation in the program group as 1, the estimate is essentially the average gain that we expect in going from the control to the program group (while holding other variables constant).

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 a value as large as—or larger than—the absolute value of the one observed when in fact there is no effect. Roughly speaking, it tells us the risk of concluding that the intervention has had an effect when in fact it hasn’t. This mistake is also known as a “false-positive” conclusion. Thus a *p* value of .1 gives us a 10% probability of drawing a false-positive conclusion. This is not to be confused with a common misconception about *p* values: that they tell us the probability of our result being true.

We can also think of the *p* value 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 $p > .20$.

In reporting results with *p* values higher than conventional statistical significance, our goal is to inform the local decision-makers with useful information and provide other researchers with data points that can be synthesized into more general evidence.

Results

Teacher-Level Implementation Results

As we described in the Methods section, we gathered data on a variety of indicators of conditions for implementation and indicators of implementation. Data from three sources provided teacher feedback about the *CT* program and their existing programs and helped us understand the implementation process. Classroom observation, phone interview, and survey data were processed, triangulated, and analyzed as separate data sources. Qualitative data were minimally coded and used as descriptive information only.

Teacher Access to Materials

Specific challenges that teachers reported in this pre-Algebra study about the use of the *CT* program were:

- Lack of access to the computer lab to use the *CT* software (reported by all teachers at one point)
- Limited support to troubleshoot technical problems by their school or the publisher (reported by 18% of the teachers)
- Glitches in the program and other technical difficulties (reported by 64% of the teachers)

When surveyed in mid-September about whether they had adequate resources to properly implement the *CT* program as specified at the *CT* training, 77% reported “No,” as displayed in Table 11. Ninety-two percent of the teachers had adequate resources to implement their existing math program while 8% did not.

Table 11. Resources for Implementation

Survey question	No	Yes
Do you have adequate resources to properly implement the program (as specified at the *CT* training)?	77%	23%
Do you have adequate resources to properly implement your existing math program?	8%	92%
Note. There were 12 teachers who responded to each question. Classroom Context Survey (September 15, 2006)		

By mid-September, 90% of the teachers reported that they did not have any access to computers – either in their classrooms or elsewhere in their schools. Those who had computers did not have the *CT* software set up yet. One-third of the teachers also reported that the *CT* textbooks had not arrived yet. As a consequence, they resorted to using their existing math materials until they received the *CT* materials in late October.

The survey self-reports were confirmed during classroom observations in September. None of the *CT* classes worked on the *CT* software; all observed classes in both groups were working out of their textbooks. Observations of these classes confirmed the reoccurrence of problems experienced in the Algebra study. For some classes, computer time was limited by the need for students to rotate during a single period. Teachers did this so that each student had some access to the computers. Computer lab days were sometimes inconsistent, as teachers had to struggle to secure a time slot.

Table 12 displays the number of networked PC or Mac computers teachers had access to in their classrooms as well as in a library or computer/media lab. We asked teachers twice about the number of networked PC or Mac computers they had access to. By November, all but two respondents had a minimum of 20 available in either their classroom or in a library or computer/media lab so that students could take turns with the *CT* software. Assuming teachers needed at least 30 networked PC computers in one location for all students to work at same time, no teachers had sufficient networked PC or Mac computers for their students in the classroom. Five teachers had sufficient access to computers in the library at the time of the first survey and six had sufficient access at the time of the second survey.

Table 12. How many total networked computers, by type and location, are available for your students' use during class time?

	Survey 5			Survey 10		
Teacher ID	Classroom	Library	Computer/ Media Center	Classroom	Library	Computer/ Media Center
1	17	24	8	17	35	15
2	0	23	10	23	55	6
3	0	0	0	1	28	11
4	16	50	20	0	30	0
5	--	--	--	--	0	0
6	0	60	10	0	60	6
7	0	32	0	--	--	--
8	0	0	0	22	40	36
9	0	60	30	--	--	--
10	0	55	36	21	50	50
11	1	20	11	2	17	10

Survey 5 (November 17, 2006)

Survey 10 (February 23, 2007)

Teacher Use of Materials

The start of *CT* implementation and use of the *CT* software varied greatly across schools and teachers. The use of the *CT* software ranged from partial use in September to no use at all. By the end of January, three teachers had not even started the use of the *CT* software due to their lack of access to computers. Teachers expressed frustration about not being able to use *CT* as it is designed to be used. Teachers commented in phone interviews and/or observations that they believed the program was not fully implemented because of this lack of technical capacity. Two teachers shared these concerns during observations and one teacher expressed in a phone interview that "the CogTutor program is not being implemented as it should be. There's NO TIME, no support, and no teacher collaboration." Teachers recognized that the *CT* textbook and *CT* software were developed to work in tandem such that the math concepts are continually reinforced. However, they found that their resources would not allow them to get their students on the *CT* software 40% of their instructional time. As a result, their students spent most of their time using the *CT* textbook. Table 13 reveals that schools' percentage of time using the *CT* textbook ranged from 66% to 93%, with an average of 78%. Time using the *CT* software ranged from 7% to 34%, with an average of 22%.

Table 13. Indicate what percentage of time you use the *CT* textbook vs. *CT* software for math instruction

	Text	Software
School 1	88%	12%
School 2	78%	22%
School 3	93%	7%
School 4	76%	24%
School 5	66%	34%
Overall	78%	22%

Note: Average use from Survey 1(September 29, 2006) to Survey 13 (April 27, 2007)

Planning

Almost two-thirds of the teachers reported that they don't have enough time during school to plan for math instruction, while the remaining teachers reported that they have enough time to plan for math instruction only sometimes. As displayed in Table 14, every teacher reported that they spend their own time outside of school preparing for math instruction.

Table 14. Planning Time

	No	Sometimes	Yes
Do you have enough time during school to plan for math instruction?	64%	36%	0%
Do you use your own time outside of school to plan for math instruction?	0%	0%	100%

Note. There were 11 teachers who responded to these questions.

Survey 4 (November 10, 2006)

Table 15 provides the average number of minutes per week teachers spent preparing for math instruction. On average they spent approximately 85 minutes per week planning for math instruction during school time and 180 minutes outside of school time. Teachers spent a comparable amount of time during school to prepare for their existing math program as they did for the *CT* program. However, teachers spent 240 minutes preparing for their existing math program outside of school compared to 180 minutes of preparation time for the *CT* program.

Table 15. Preparation Time

How much time do you spend preparing for:		Average number of minutes per week		
		Minimum	Maximum	Mean
***CT* Instruction**	**During School**	5	300	85
	Outside of School	30	480	180
Existing Math Program	**During School**	5	300	90
	Outside of School	55	900	240

Note. There were 11 teachers who responded to these questions.

Survey 4 (November 10, 2006)

Student Engagement

We asked teachers to rate how engaged they thought their students were in both groups. Table 16 details the varying levels of student engagement with the different math materials. Teachers reported that the majority of the students were “very engaged” in the *CT* software. The majority of students were only “somewhat engaged” in the *CT* textbook and the majority of the control students were “somewhat engaged” in the existing math text and activities. While group presentations are essential to Carnegie Learning’s philosophy of student collaboration, only 8% of the students were “very engaged” in the *CT* group presentations.

Table 16. Please rate the average level of student engagement in the following areas

	I don't know	Not at all engaged	Not very engaged	Somewhat engaged	Very engaged
***CT* Textbook**	8 %	0%	0%	50%	42%
***CT* Software**	25%	0%	8%	8%	58%
***CT* Group Presentations**	8%	25%	33%	25%	8%
Existing Math Text	0%	25%	8%	42%	25%
Exiting Math Activities	0%	17%	0%	58%	25%

Note. There were 12 teachers who responded to these questions.

Survey 7 (January 19, 2007)

Nine out of eleven teachers who provided an open response to these survey questions reported that students liked getting on the computer software and its self-paced nature and the real-world problems that they can relate to. Two teachers shared that students enjoyed the group activities as well as the challenges that the problems brought.

Teachers also reported what students disliked about the *CT* program. One-fourth of the teachers said that the students were bored with the *CT* textbook and wanted more time with the *CT* software (versus time with the *CT* textbook). Some students complained about the word problems and the process of working them “backwards.” Six out of eleven teachers commented that their students thought the *CT* content was too difficult, the reading was too dense, and the tests were harder than the lessons.

Teacher Collaboration

While teacher collaboration is enforced in the *CT* materials, we found that teachers had collaboration meetings with other teachers more often for their existing math program than for *CT*. Table 17 displays that more than a quarter of the teachers had collaboration meetings at least monthly for the existing program, while only 9% did so for *CT*.

Table 17. Do you have collaboration meetings with other teachers (e.g. for planning lessons) for your math classes? If so, how often?

	Never	Once/Twice a Semester	At Least Monthly	At Least Weekly	At least Three Times a Week
CT	45%	45 %	9%	0%	0%
Existing Math	18%	55%	27%	0%	0%

Note. There were 11 teachers in each group who responded.

Survey 4 (November 10, 2006)

All teachers said that the *CT* approach to collaborative learning affected the way they teach their *CT* classes. They further specified that collaborative learning allows:

- "more interrogative interplays between students"
- The teacher to be more of a facilitator
- The teacher to "pay more attention as to how the students are grouped"
- More sharing is encouraged.
- Use of peer-teaching
- Students to seek help from peers before asking for further assistance from the teacher

As shown in Table 18, 45% of the teachers reported that the *CT* approach to collaborative learning affected the way they teach their control classes. In the free response portion, teachers specified that:

- "It doesn't really translate."
- " As instructed, I do not carry over any instruction to my existing program."

Table 18. *CT* Approach

Survey question	No	Yes
Has the *CT* approach to collaborative learning carried over into instruction of your existing math program?	55%	45%

Note. There were 13 teachers who responded.

Survey 9 (February 9, 2007)

This result demonstrates that for collaborative aspect of instruction, there was a degree of carry over to the control classes.

Hawaii State Standards

Our surveys revealed that teachers felt that their *CT* students were equally prepared for the Hawaii State Assessment as their students in their control classrooms, as displayed in Table 19. Specifically, 36% of the teachers reported that their students were "somewhat prepared" for the HSA, 36% of the teachers felt that their students were "not very prepared" and 27% of the teachers did not have a real sense of how prepared their students were for the HSA.

Table 19. How prepared do you think your *CT* students were for the HSA?

	I Don't Know	Not At All Prepared	Not Very Prepared	Somewhat Prepared	Very Prepared
CT	27%	0%	36%	36%	0%
Control	27%	0%	36%	36%	0%

Note. There were 11 teachers in each group (*CT* and control) who responded

Survey 12 (March 13, 2007)

Teacher Comfort with Technology

Teachers reported a range of levels in computer skills prior to using *CT*. Forty-six percent of the teachers felt their skill level was "intermediate." Table 20 details how teachers rated their own skill level over a five-month period. After having used *CT*, three teachers reported advancing in skill level, six teachers reported their skill level as the same as what they reported initially, and two teachers reported a declined skill level.

Table 20. Please rate your level of computer skills prior to using *CT* (1 = Basic, 2 = Intermediate, 3 = Advanced)

	Number of Teachers		
	Basic	Intermediate	Advanced
Survey 1	4	6	3
Survey 10	1	9	1

Note. This question was asked on 2 separate surveys. 13 teachers responded to this question on Survey 1 and 11 teachers responded to this question on Survey 10.

Survey 1 (September 29, 2006)

Survey 10 (February 23, 2007)

We asked teachers how much they agreed with statements pertaining to technology and computers as they relate to teaching and learning math. In Table 21, we see that in general, after use of the *CT* program, a few teachers felt more agreeable (than previously reported) that "Technology in general is beneficial in the classroom", "Technology helps students learn math," and that "Computers are beneficial in the classroom," However, one teacher felt less agreeable (than previously reported) that "Computer software can help students learn math" and a few teachers felt less agreeable (than previously reported) that computer software can help enhance their teaching capabilities.

Table 21. Technology and Computers: Using the following scale, please rate how well you agree with the following statements: 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree

Survey question	Survey	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Technology in general is beneficial in the classroom	10	0%	0%	0%	36%	64%
Technology helps students learn math	10	0%	0%	0%	45%	55%
Computers are beneficial in the classroom	10	0%	0%	9%	27%	64%
Computer software can help students learn math	10	0%	0%	9%	27%	64%
Computer software can help enhance my teaching capabilities	10	0%	9%	9%	18%	64%

Note. There were 11 teachers who responded to these questions.
Survey 10 (February 23, 2007)

We asked teachers to rate their level of confidence and comfort in implementing the *CT* program at the start of implementation and in April. Table 22 details the change of confidence and comfort in implementing *CT* over time. From the start of implementing the *CT* program to April, 45% of the teachers moved up in level of confidence,36% of the teachers remained at the same level of confidence and 18% of the teachers went down a level of confidence. From the start of implementing the *CT* program to April, 72% of the teachers moved up in level of comfort, 9% of the teachers remained at the same level of comfort and 18% of the teachers went down a level of comfort.

Table 22. Change in confidence and comfort levels in implementing *CT* over time

	Down	Up	Same
Change in Level of CONFIDENCE while implementing *CT* Program	18%	55%	27%
Change in Level of COMFORT while implementing *CT* Program	18%	73%	9%

Note. Survey question asks for change in levels from start of experiment to April 2007. There were 11 teachers who responded.

We asked teachers to report their interest and motivation in learning about computer technologies. Table 23 displays that seventy-three percent of the teachers reported being very interested and motivated to learn more about computer technologies.

Table 23. How would you rate your interest and motivation in learning about computer technologies?

Not Interested At All	Neutral	Somewhat Interested	Interested	Very Interested
0%	0%	8%	23%	69%

Note. There were 13 teachers who responded.

Survey 1 (September 29, 2006)

Teacher Satisfaction with Math Programs

CT Materials

When asked to select the pieces of the *CT* materials that were most useful, the highest percentage of teachers (58%) selected the *CT* Student Text, as displayed in Table 24. Out of the 12 teachers responding to the survey, 7 teachers selected the *CT* student text, only 4 selected the Student Assignment Book, and only 3 selected the *CT* software, Teacher's Implementation Guide and Teacher's Resources and Assessments. None of the teachers selected the Homework Helper or the Software Implementation Guide.

Table 24. Please Select the Cognitive Tutor Materials that Have Been Most Useful

Materials	Percent of teachers who selected this material
Student Assignment Book	33%
Homework Helper	0%
Cognitive Tutor Software	25%
Teacher's Implementation Guide	25%
Teacher's Resource and Assessments Book	25%
Software Implementation Guide	0%
Student Text	58%

Note. There were 12 teachers who responded. Teachers were able to make more than one choice.

Survey 12 (March 13, 2007)

When asked to comment on the each piece of the *CT* materials, teachers shared that the *CT* textbook was the most useful because it was consumable and it contains visual cues and detailed explanations that students can understand and relate to. The second most useful piece of the *CT* materials was the *CT* student Assignment book because it was consumable and contained activities for extended practice. Teachers also reported the equal usefulness of the *CT* Software, *CT* implementation Guide and *CT* Teacher's Resource and Assessment Book. The *CT* software helped keep the students on task and provided immediate feedback. The *CT* Teacher's Implementation Guide provided additional suggestions for math instruction and the *CT* Teacher's Resource and Assessment Book confirmed learning objectives. Two teachers also shared their preferences in materials, stating:

- "The existing math textbook is preferred, supplemental to the *CT* software."
- "I prefer the regular instruction of my other class, but the software from the cognitive tutoring."

CT Content

When asked to share any opinions about the *CT* content, teachers reported both positive and negative comments. Specifically, individual teachers made the following comments:

- promoted writing and reading comprehension
- was too elementary for high school pre-Algebra students
- was too advanced for their students
- did not align with the high school standards and benchmark mapping for middle school

Existing Math Program

Teachers reported having various textbooks, guides, and supplemental material that compose their existing math program. They shared both positive and negative comments about these materials. Eighty-three percent of the teachers reported liking their existing textbooks because of the hands-on activities, manipulatives and the variety of additional support materials as options for different learners. Forty percent of the teachers also had negative comments. They shared that:

- some of their textbooks were too advanced for their students
- the activities did not promote critical thinking, reading or writing comprehension.
- problem-solving strategies were too difficult for students.

Over 90% of the teachers were either “satisfied” or “very satisfied” with their existing math programs, as displayed in Table 25.

Table 25. Satisfaction Level

Survey question	Not satisfied	Neutral	Satisfied	Very satisfied
How would you rate your level of satisfaction with your existing math program?	0%	8%	75%	17%

Note. There were 12 teachers who responded

Survey 2 (October 13, 2006)

Indicators of Favorable Conditions for Implementation

There is evidence of favorable conditions for implementation, particularly in the data relevant to teacher confidence and comfort with technology and to teacher satisfaction. The data show that teachers were increasingly confident and comfortable using technology. In general, the majority of the teachers were very or somewhat confident and comfortable in using technology at the start of the program. By April, 90% of the teachers felt very or somewhat confident and comfortable in using technology. The majority of teachers agreed or strongly agreed that technology and computers are beneficial to teaching and learning math. Finally, we found that teachers were mostly satisfied or very satisfied with both their existing math program and *CT* materials.

There was also evidence of less than favorable conditions for implementation, specifically in the data on access to resources. Teachers reported challenges such as delayed program implementation and lack of access to computers. The data show that, at the start of the school year, 92% of the teachers had adequate access to resources to implement their existing math program, whereas only 23% had adequate resources to implement the *CT* program. The data reveal that teachers had limited access to networked computers. This partly influenced the use of the *CT* program, specifically the *CT* software. In September, teachers had already started experiencing problems in finding enough computers. By November, while all but two teachers had some access, only five teachers had enough for all students, and none had enough for all students in the classroom. By the end of January, three teachers had not even started the use of the *CT* software due to their lack of access to computers. Only five teachers had sufficient access to computers in the library or computer/media lab in November and only six teachers had sufficient access in February.

All teachers had eventually received the full set of *CT* materials and the full training for the *CT* program. One third of the teachers reported having enough time to plan for math instruction only sometimes. The remainder of the teachers did not feel that there was enough time to plan for math instruction at all.

Indicators of Implementation

The data indicate how much teachers implemented the *CT* program compared to their existing math program. A specific indicator of program implementation is use of materials. At the start of the school year, teachers were able to fully implement their existing math program with their control classes while the *CT* classes had to wait to receive the full set of *CT* materials. Because they didn't have adequate resources, we found that teachers varied in use of *CT* materials. While the *CT* program suggests 40% use of the *CT* software for ideal implementation, teachers reported using the *CT* software 22% of their instructional time on average from September to April.

Another indicator of implementation is use of collaboration. While teachers themselves collaborated more for their existing math program than for the *CT* program, the data show that teachers enforced Carnegie Learning's approach to collaborative learning in their *CT* classes. Teachers supported this because it allowed students to share in their learning, utilize peer-teaching methods, and interact with one another.

We found that nearly half of the teachers reported that the *CT* approach to collaborative learning affected the way they teach their control classes. This carryover of instructional approach from one condition to the other introduces the possible impact of contamination, which may or may not have affected student outcomes.

Summary of Implementation

Overall, our data sources reveal that teachers reported a generally positive attitude about the *CT* program. Teachers reported overall general ease of use of the *CT* program as well as positive interactions with all *CT* materials. Survey responses reveal that some teachers preferred the combination of their existing math textbook and the *CT* software.

We found that surveys, observations, and interviews showed increased student interest and engagement in *CT* classes compared to classes without the program. In general, teachers reported a slight increase in confidence and comfort in implementing *CT* after having used the program for several months. The majority of teachers felt that technology and computers are beneficial and helpful in teaching students how to understand math.

Similar to the Algebra study, challenges teachers continued to face in implementing the *CT* program were the lack of or limited access to resources and delayed implementation due to lack of materials. This caused teachers to not be able to implement the *CT* program as specified. Teachers also reported concerns about the misaligned content with the Hawaii state standards.

Student-Level Impact Results

Our overall outcome measure was the score on the NWEA General Math Test. We also address the outcomes of the Algebraic Operations sub-strand of the NWEA General Math Test 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 both the overall score as well as the score on the Algebraic Operations sub-strand or section of the test, we first estimate the average impact of *CT* on student performance. In the following tables and graphs, these results are presented in terms of effect sizes. We then show the results of additional 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 main 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.

Overall Score on the NWEA General Math Test

Table 26 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 information about all the students in the original sample for whom we have pretest and posttest scores. 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 "Adjusted" row is based on the same sample of students, but uses the effect estimate from a model that adjusts for the effect of the pretest as well as fixed effects used to model group membership above the level of randomization (e.g., pairs). In other words, the adjusted effect size is based on a model that contains the standard effects that are used in most of the models in the analyses that follow.

Table 26. Overview of Sample and Impact of *CT* on the Overall Score: NWEA General Math Test

	Condition	Means	Standard deviations[a]	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]	Percentile standing
Un-adjusted	**Control**	222.37	13.36	237	14	11	0.13	.48	5.17%
	CT	224.71	12.38	239	14	11			
Adjusted	**Control**	222.37	13.36	The same sample is used in both calculations			0.04	.57	1.60%
	CT	223.04[c]	12.38						

[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 unadjusted effect size is Hedges' *g* with the *p* value adjusted for clustering. The *p* value is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The adjusted effect size is the impact estimate from PROC MIXED divided by the estimate of the pooled standard deviation. 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. The *p* value is for the effect estimate from PROC MIXED.

[c] For the adjusted effect size, separate intercepts are modeled for levels of the blocking variable, therefore leading to different estimates for control group performance for each block. 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, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Figure 1 provides a visual representation of specific information in Table 26. The bar graphs represent average performance using the metric of the NWEA General Math 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 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 26). The overall impact on math as an effect size (i.e., in terms of standard deviation units) is 0.04 which is equivalent to a gain of about 1.6 percentile points for the median control group student if the student had received *CT* However, the high *p* value for the treatment effect (.57) indicates that we should have no confidence that the actual difference in average performance 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.

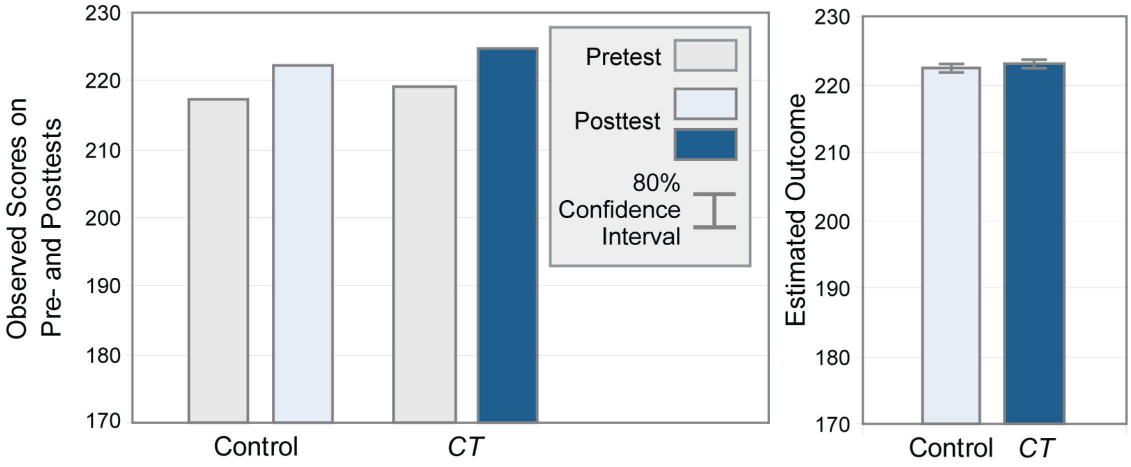


Figure 1. Impact on the Overall Score of the NWEA General Math Test: Unadjusted Pre- and Posttest Means for Control and *CT* (Left); Adjusted Means for Control and *CT* (Right)

Table 27 shows the estimated impact of *CT* on students' performance for the overall score on the NWEA General Math Test[5]. The bottom rows of Table 27 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* works for a student near the middle of the pretest range. The estimate associated with the treatment is .49, which is the estimated difference between the *CT* and control conditions, for a student with an average pretest score. This shows a small positive difference associated with *CT*. The *p* value of .57 gives us no confidence that the true impact is different from zero. In other words, the result could easily reflect a chance difference.

Table 27. Impact of *CT* on Student Performance on the Overall Score of the NWEA General Math Test

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for a control student with an average pretest	227.57	1.57	13	144.84	<.01
Change in outcome for a control student for each unit-increase on the pretest	0.95	0.03	444	30.1	<.01
Effect of *CT* for a student with an average pretest	0.49	0.84	13	0.59	.57
Change in the effect of *CT* for each unit-increase on the pretest	-0.02	0.04	444	-0.38	.71
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Class mean achievement	3.1	1.87		1.65	.05
Within-class variation	29.41	1.97		14.92	<.01

[a]Pairs of classes used for random assignment were modeled as a fixed factor but not included in this table.

[b]Classes were modeled as a random factor.

[5] 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.

As a visual representation of the result described in Table 27, 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.[6]

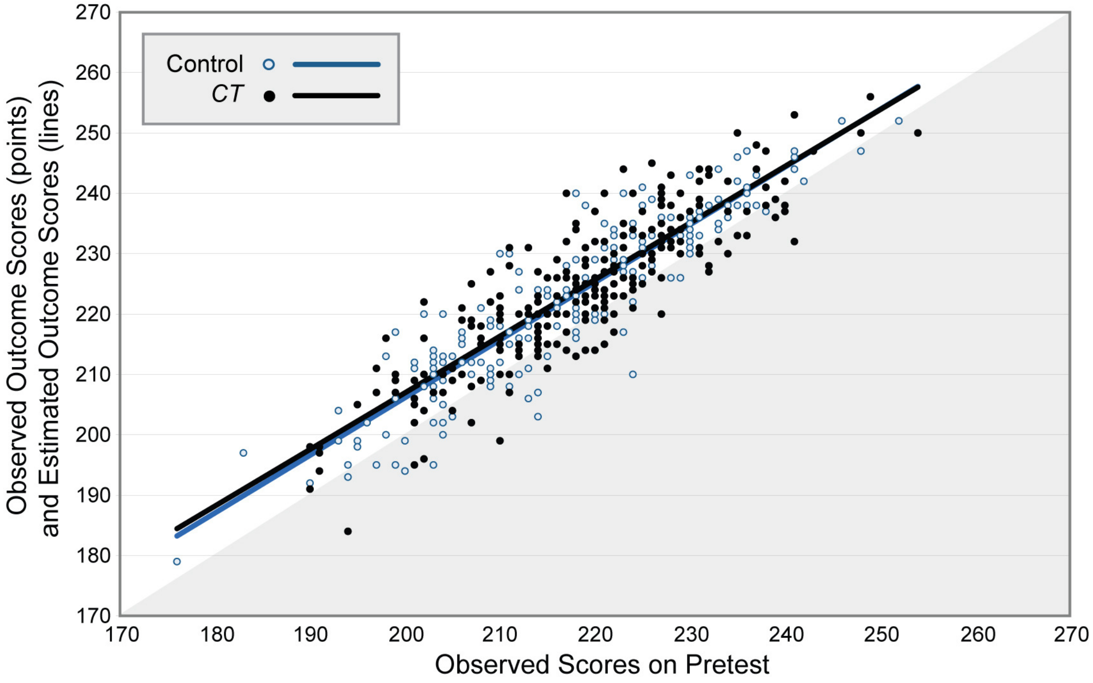


Figure 2. Comparison of Estimated and Actual NWEA General Math Outcomes for Control and *CT* Students

In addition to analyzing the difference between the *CT* and control groups on the overall score on the NWEA General Math Test, we conducted further analyses to determine whether the impact of *CT* on students' performance varied on the specific Algebraic Operations sub-strand.

[6] Pairs were modeled as a fixed factor, resulting in a separate intercept estimate for each pair. To fix the vertical location of the prediction lines, we selected the median estimate for the intercept.

Algebraic Operations

Our next set of calculations addresses Algebra achievement as measured by the Algebraic Operations sub-strand of the test. Table 28 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 26.

Table 28. 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	*p* value[b]	Percentile standing
Un-adjusted	**Control**	225.75	15.28	237	14	11	0.08	.57	3.19%
	CT	227.10	14.03	239	14	11			
Adjusted	**Control**	225.75	15.28	The same sample is used in both calculations			0.04	.65	1.60%
	CT	226.52[c]	14.03						

[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 unadjusted effect size is Hedges' *g* with the *p* value adjusted for clustering. The *p* value is computed using a model that figures in clustering of students in classes but does not adjust for any other covariates. The adjusted effect size is the impact estimate from PROC MIXED divided by the estimate of the pooled standard deviation. 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. The *p* value is for the effect estimate from PROC MIXED.

[c] For the adjusted effect size, separate intercepts are modeled for levels of the blocking variable, therefore leading to different estimates for control group performance for each block. 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, 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 28. The bar graphs represent average performance using the metric of the NWEA test of Algebraic Operations.

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 that includes an adjustment for students' pretest scores and other fixed effects (i.e., this is a visual display of results from the row labeled "Adjusted" in Table 28). The overall impact on the Algebraic Operations sub-strand as an effect size (standard deviation units) is .04, which is equivalent to a gain of about 1.6 percentile points for the median control group student if the student had received *CT* . The high *p* value gives us no confidence that the observed difference occurred for reasons other than chance.

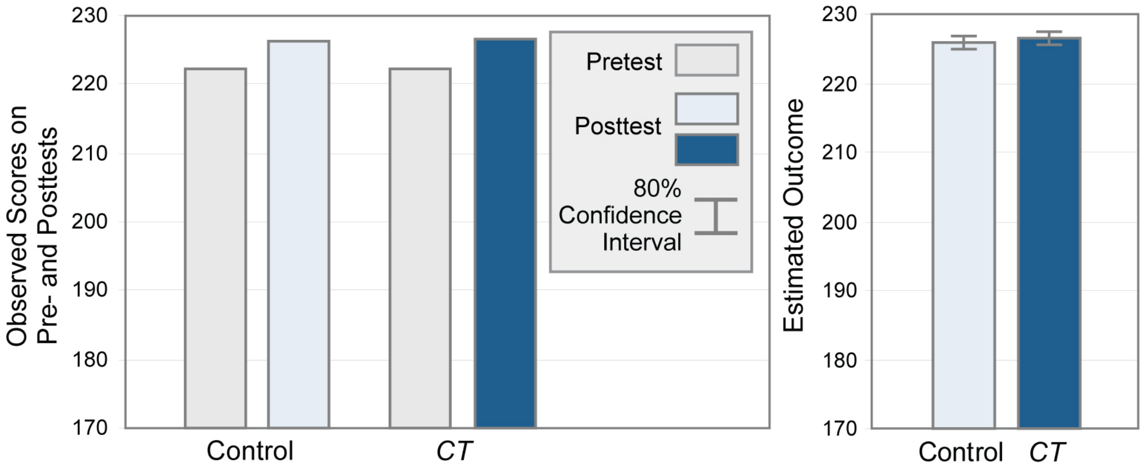


Figure 3. Impact on Algebraic Operations: Unadjusted Pre- and Posttest Means for Control and *CT* (Left); Adjusted Means for Control and *CT* (Right)

Table 29 shows the estimated impact of *CT* on students' performance on the Algebraic Operations sub-strand. For a student with an average score on the pretest, there is roughly a .85-point advantage to being in the *CT* group. The high *p* value of .53 suggests that the observed advantage is easily a chance result; that is, we have no confidence that there is a true advantage. However, we also observe a low *p* value (.02) for the change in the effect of *CT* for each unit-increase on the pretest. This means that the value of *CT* cannot be understood without considering how *CT* and the pretest score work together. Specifically, the higher the pretest score, the lower the impact of *CT*. In other words, there is a diminishing return to the impact of *CT* as the pretest score increases.[7]

[7] In the model used, intercepts are modeled as random at the student and class levels and as fixed at the pair level; however, slopes are not modeled as random; the interaction of pretest with treatment and the corresponding *p* value do not reflect uncertainty due to the re-sampling of classes or teachers.

Table 29. Impact of *CT* on Student Performance on the Algebraic Operations Sub-strand of the NWEA General Math Test

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for a control student with an average pretest	229.36	2.43	13	94.23	<.01
Change in outcome for a control student for each unit-increase on the pretest	0.69	0.05	443	13.66	<.01
Effect of *CT* for a student with an average pretest	0.85	1.32	13	0.65	.53
Change in the effect of *CT* for each unit-increase on the pretest	-0.16	0.07	443	-2.37	.02
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Class mean achievement	5.79	4.9		1.18	.12
Within-teacher variation	102.95	6.92		14.88	<.01

[a] Pairs of classes were modeled as a fixed factor but are not included in this table.

[b] Classes were modeled as a random factor.

This interaction is most readily interpreted through inspection of graphs. As a visual representation of this result, we present a scatterplot in Figure 4, which graphs student growth over the school year in terms of Algebra achievement as measured by the Algebraic Operations sub-strand of the NWEA General Math Test. This graph is interpreted the same way as Figure 2. We are unable to discern an average difference between the two conditions on the posttest. The interaction is evident in the crossing of the prediction lines: it indicates that with these classes and teachers, lower-performing students benefit from treatment; that is, they are helped more by *CT* than are higher-performing students.[8]

[8] The apparent dip of the regression lines below the “no growth line” (i.e., into the gray area) towards the top of the pretest scale, and more generally, the non parallelism between the no growth line and the regression lines is an artifact of the regression process (high-performers on average tend to do less well when retested, and low-performers tend to do better) and has no connection to treatment.

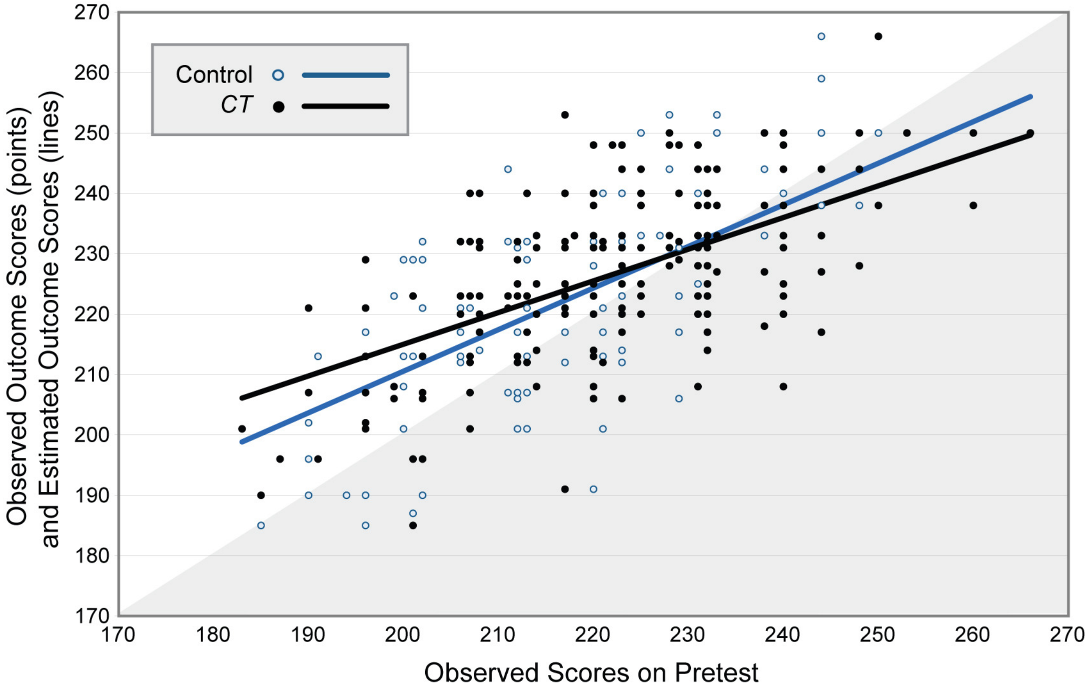


Figure 4. Comparison of Estimated and Actual Algebraic Operations Outcomes for Control and *CT* Students

Figure 5 displays an alternative visual of the results reported in Table 29 by graphically showing the predicted difference between the *CT* and control groups. The 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; here, 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 29, there is evidence of a positive impact for lower performing students, and little or no impact at the higher end of the pretest scale. (The 95% confidence interval does not cross the horizontal axis for the median student in the first quartile, indicating the presence of an effect; in contrast, the 80% confidence interval crosses the horizontal axis for the median student in the top quartile.)

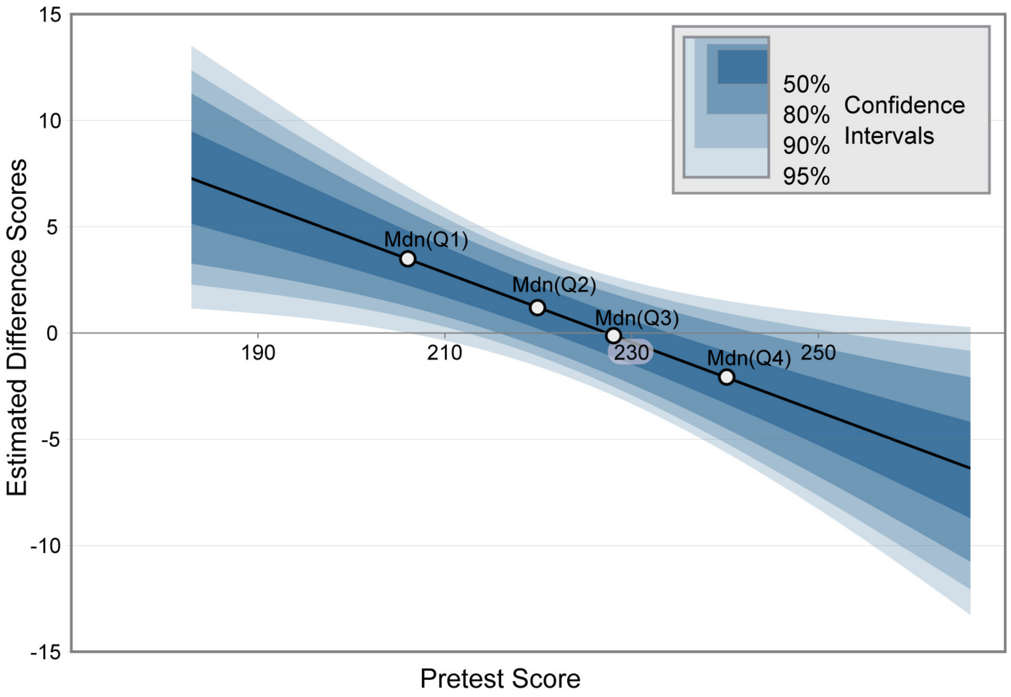


Figure 5. Differences between *CT* and Control Algebraic Operations Outcomes: Median Pretest Scores for Four Quartiles Shown

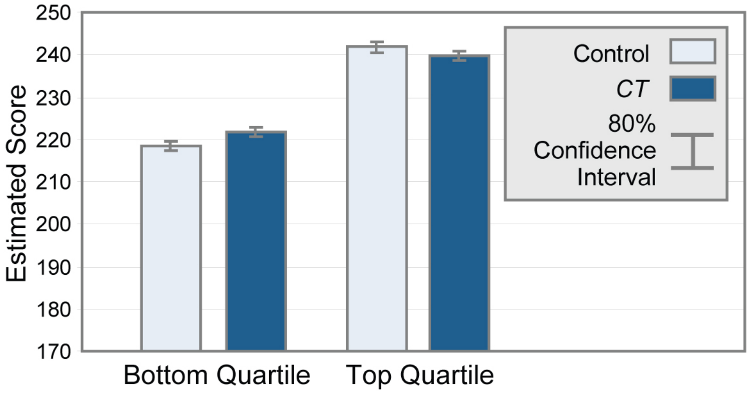


Figure 6. Differences between *CT* and Control Algebraic Operations Outcomes: Median Pretest Scores in Top and Bottom Quartiles

An alternative way of understanding the information in Figure 4 is to represent the result using a bar graph specifically for the students at the median of the top and bottom quartiles of the pretest. Figure 6 presents the estimated difference between *CT* and control for the median student at the two extreme quartiles. Figure 6 indicates that there is an advantage to being in *CT* for the median student in the bottom quartile. The small overlap in confidence intervals for the median student in the top quartile means we have no confidence that such a student would perform differently in the two conditions.

Moderating Effect of Teacher Certification on Student Outcomes

NWEA General Math Test

Following a suggestion by the district's Math Science Partnership (MSP) consultant, we considered whether the impact of *CT* is differentially effective for students who had certified teachers versus those with uncertified teachers. For this experiment, results indicate that there is a slightly negative effect for certified teachers and a strong positive effect for uncertified teachers. Table 30 shows the moderating effect of teacher certification on students' performance on the NWEA General Math Test.

Table 30. Moderating Effect of Teacher Certification on NWEA General Math 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	222.39	1.50	16	147.83	<.01
Change in outcome for each unit-increase on the pretest	0.95	0.02	444	40.96	<.01
Difference (students of certified – students of uncertified teachers) in control outcome	0.37	2.12	16	0.18	.86
Effect of *CT* for non-certified teachers' students	2.34	1.12	16	2.08	.05
Difference (students of certified – students of uncertified teachers) in the effect of *CT*	-3.12	1.46	16	-2.14	.05
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	1.81	1.32		1.37	.09
Within-teacher variation	29.39	1.97		14.91	<.01

[a] Teachers were modeled as a fixed factor but the estimated effects are not included in this table; the predicted value for a control student with an average pretest applies to the reference uncertified teacher.

[b] Classes were modeled as a random factor.

While the control students of certified and non-certified teachers performed similarly, *CT* students who had non-certified teachers outperformed *CT* students with certified teachers. This differential effect is substantiated by the low *p* value (.05) for the difference in the effect of *CT* in the table above. We are confident that the observed benefit of *CT* for students of uncertified teachers is not just a matter of chance. While this finding is intriguing, with only 11 teachers in the sample, we cannot generalize this result to other certified or non-certified teachers.[9]

[9] The teacher effect was modeled as fixed; therefore, we cannot be certain that the interaction would be sustained if teachers were re-sampled. The *p* value for the interaction reflects our level of certainty for a sustained effect on re-sampling students for the same teachers.

As a visual representation of the result described in Table 30,[10] Figure 7 shows the estimated difference between *CT* and control for an average student with a certified teacher versus a non-certified teacher. The low *p* value for the difference in the effect of *CT* is shown graphically by the *CT* bar being higher than the control bar for non-certified teachers, and the reverse being true for the certified teachers (though the overlap of the confidence intervals for the latter difference does not give us confidence the difference did not occur by chance.)

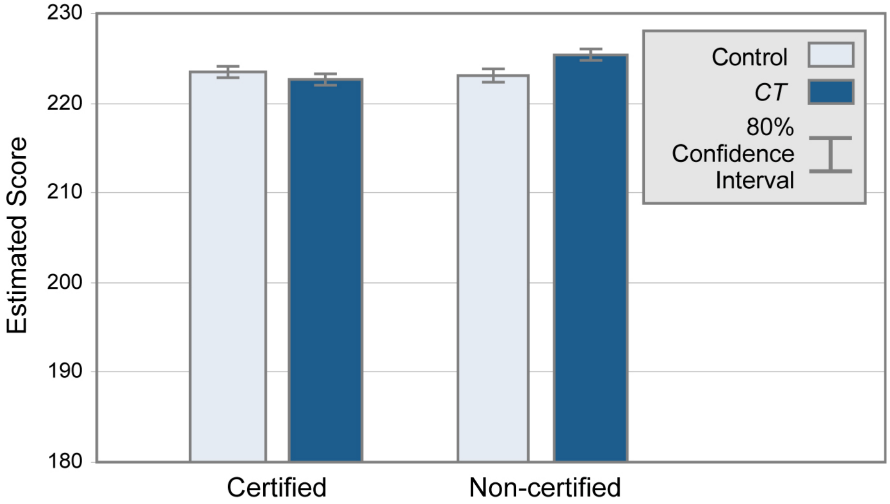


Figure 7. Moderating Effect of Teacher Certification on NWEA General Math Outcomes

[10] The net program effect for students of uncertified teachers is from row 4 (i.e., 2.34), and the net program effect for students of certified teachers is row 4 + row 5 (i.e., 2.34 – 3.12).

NWEA Algebraic Operations

Table 31 shows the same moderator analysis but using the Algebraic Operations sub-strand as the outcome. As noted previously, we chose Algebraic Operations because the program under study focuses on preparation for Algebra.

In contrast to the investigation of the overall score, we did not find an effect of certification, a *CT* effect for students of non-certified teachers, or a differential effect between certified and uncertified teachers.

Table 31. Moderating Effect of Teacher Certification on NWEA Algebraic Operations Outcomes

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for the uncertified teacher's control student with an average pretest	224.23	2.84	16	78.86	<.01
Change in outcome for each unit-increase on the pretest	0.63	0.04	445	17.27	<.01
Difference (students of certified – students of uncertified teachers) in control outcome	-0.60	4.00	16	-0.15	.88
Effect of *CT* for uncertified teachers' students	2.36	2.12	16	1.11	.28
Difference (students of certified – students of uncertified teachers) in the effect of *CT*	-3.01	2.76	16	-1.09	.29
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	6.25	4.79		1.30	.10
Within-teacher variation	107.96	7.24		14.92	<.01

[a] Teachers were modeled as a fixed factor but the estimated effects are not included in this table; the predicted value for a control student with an average pretest applies to an uncertified teacher with an average score on the pretest.

[b] Classes were modeled as a random factor.

Moderating Effect of Ethnic Background on Student Outcomes

Again following a suggestion by the district's MSP consultant, we considered whether the treatment impact varies for students of different ethnicities. The ethnicities of particular interest were Filipino and Hawaiian (including part-Hawaiian). Based on the ethnicity categories used by schools in Hawaii, we divided the remainder into White and Other. Table 5 displayed the pretest levels for the full set of categories before they were consolidated into four. We examined both the overall results for the NWEA General Math Test and for the Algebraic Operations sub-strand.

NWEA General Math Test

We started by considering whether the impact of *CT* was the same for all ethnic groups. The statistical test of this question revealed no variation in the treatment effect across the categories of ethnicity[11]. This does not mean that there is no such interaction; we simply don't have the sample sizes that would be required to detect such differences. If we had a larger sample of teachers and classes, we may have been able to detect small meaningful differences.[12]

As an exploratory exercise, to motivate future study, we provide a table comparing the performance of control students, as well as the treatment effects, by ethnicity. In setting up Table 20, we used Filipino as the reference category because they are the largest ethnic group, and then compared the other categories to that group. The table begins by presenting the results for the Filipino group including the effect of *CT* (row 3). Although this estimate is a positive number, it has a high *p* value, thus giving us no confidence that this difference between *CT* and control Filipino students is not just a chance difference. (The estimated differences provided in this table have taken the pretest score into account. This means that the differences between students of different ethnicities apply to students with the same pretest score.)

[11] We considered the type-3 test of fixed effects for the interaction of treatment with ethnicity. This test determines whether we should have confidence that adding the interaction to the model leads to more variance in the outcome being accounted for than if we don't add this interaction. The high *p* value of .51 that we obtained (table not displayed) gives us no confidence that the interaction accounts for additional variance – the variation we see in the impact of *CT* across ethnic groups is easily due to chance.

[12] This experiment was powered to detect an effect size for an average impact of about .35. Differential effects would need to be larger than the average effect in order to be detected at conventional levels of false positives and false negatives, assuming the same sample size. If differential effects are assumed to be smaller than the average effect then larger samples are needed in order to detect them.

Table 32. Moderating Effect of Ethnic Background on NWEA General Math Test

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Outcome for a Filipino control student with an average pretest	227.89	1.71	13	133.48	<.01
Change in outcome for each unit-increase on the pretest	0.96	0.02	435	40.25	<.01
Effect of *CT* for a Filipino student	1.07	1.09	13	0.98	.35
Difference (White student – Filipino student) in control outcome	-0.49	1.84	66	-0.27	.79
Difference (Other student – Filipino student) in control outcome	-0.44	0.95	66	-0.46	.65
Difference (Hawaiian/Part-Hawaiian student – Filipino student) in control outcome	-0.35	0.94	66	-0.38	.71
Difference (White student – Filipino student) in the effect of *CT*	1.61	2.41	435	0.67	.50
Difference (Other student – Filipino student) in the effect of *CT*	-0.40	1.33	435	-0.30	.77
Difference (Hawaiian/Part-Hawaiian student – Filipino student) in the effect of *CT*	-1.59	1.33	435	-1.19	.23
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Class mean achievement	3.58	2.09		1.71	.04
Within-class variation	30.05	2.04		14.76	<.01

Note: The model controlled for clustering and treated classes as a random factor while modeling pairs as fixed.

NWEA Algebraic Operations

As with the overall score, we tested whether there is a significant amount of variation among ethnic groups in the impact of *CT* on the Algebraic Operations sub-strand.[13] We found that the observed differences could easily be due do chance.

Again, as an exploratory exercise, to motivate future study, we provide a table comparing the performance of control students, as well as the treatment effects, by ethnicity. Table 33 shows the same moderator analysis as was shown in Table 20, this time using the Algebraic Operations sub-strand as the outcome.

The effects in Table 33 suggest a direction for future study. We are especially interested in the possibility that, controlling for pretest, *CT* affects Filipinos and Whites differently. We see a possible divergence in scores between control students in the two ethnic groups (after controlling for

[13] As with overall math, we considered the type-3 test of fixed effects for the interaction of treatment with ethnicity. The high *p* value of .42 that we obtained (table not displayed) gives us no confidence that the interaction accounts for additional variance – the variation we see in the impact of *CT* across ethnic groups is easily due to chance.

differences in the pretest.) An intriguing possibility is that the intervention compensates for the increasing discrepancy in scores between Whites and Filipinos that is observed for control students. Further study would help to establish whether this effect is real, or whether the trends observed in the current study just reflect chance. The caveats about interpreting results that were raised in the previous section apply here as well.

Table 33. Moderating Effect of Ethnic Background on NWEA Algebraic Operations Outcomes

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Outcome for a Filipino control student with an average pretest	229.52	2.92	13	78.50	<.01
Change in outcome for each unit-increase on the pretest	0.64	0.04	435	17.53	<.01
Effect of *CT* for a Filipino student	2.20	1.93	13	1.14	.28
Difference (White student – Filipino student) in control outcome	6.17	3.66	66	1.68	.10
Difference (Other student – Filipino student) in control outcome	2.27	1.81	66	1.26	.21
Difference (Hawaiian/Part-Hawaiian student – Filipino student) in control outcome	-0.84	1.78	66	-0.47	.64
Difference (White student – Filipino student) in the effect of *CT*	-6.24	4.66	435	-1.34	.18
Difference (Other student – Filipino student) in the effect of *CT*	-2.20	2.51	435	-0.87	.38
Difference (Hawaiian/Part-Hawaiian student – Filipino student) in the effect of *CT*	-3.14	2.52	435	-1.25	.21
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Class mean achievement	9.27	6.42		1.44	.07
Within-class variation	108.49	7.36		14.75	<.01

Note. The model controlled for clustering and treated classes as a random factor while modeling pairs as fixed.

To illustrate what we consider an intriguing possibility, in Figure 8 we visually represent the differences in Algebraic Operations outcomes between White students and Filipino students (our reference category) for both the *CT* and control groups. We notice that White students perform better in the control group than in the *CT* group, in contrast to Filipino students, who perform slightly better in the *CT* group than in the control condition. Although the direction of the change is visibly different for the two ethnic groups, the overall test of variation in the impact of *CT* reported in Table 33 gives us no confidence that the differences would be found again if we re-ran the experiment with another sample of students. The trend indicated by this graph suggests that additional study could help establish whether *CT* in fact works differently for the two ethnic groups.

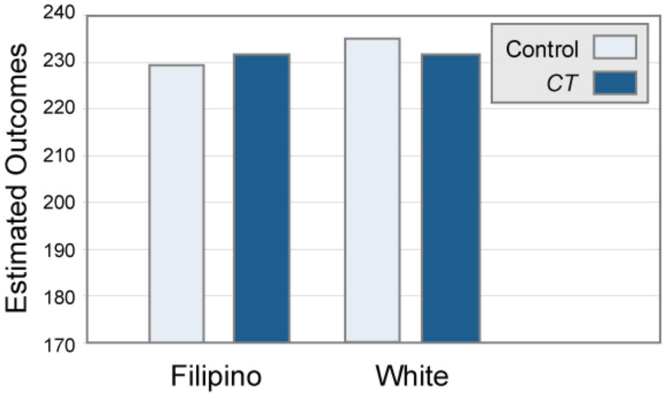


Figure 8. Difference in Algebraic Operations Outcomes between White and Filipino Students in the Control and *CT* Conditions

Discussion

We began our research by asking whether there is a difference in effectiveness between Carnegie Learning's *Cognitive Tutor Bridge to Algebra* program and the pre-Algebra program currently in place in the Maui School District. To assess the impact in this experiment, we used the NWEA General Math Test as a measure of student achievement in pre-Algebra. We looked at the overall math achievement on the NWEA General Math Test as well as the Algebraic Operations sub-strand. As suggested by the district, we also examined whether *CT* was differentially effective 1) for Hawaiian/Part-Hawaiian students or Filipino students versus students in other ethnic groups or 2) for students who had certified teachers versus students who had non-certified teachers. Finally, through qualitative methods, we looked closely at how teachers were implementing *CT* throughout the year.

For this research, we used a randomized control trial. Using a coin toss, we randomly assigned 32 classes to use the *CT Bridge to Algebra Program* or to continue using the pre-Algebra program currently in place. Each of the 12 teachers involved in the experiment had equal numbers of classes in one program and the other.

In our randomized experiment, we found that most students, regardless of condition, improved on the overall math scale used in the NWEA General Math Test. However, the high *p* value for the average treatment effect (.57) indicates that we should have no confidence that the average treatment effect is different from zero. In other words, we did not find a difference in student performance in math, as measured by the NWEA General Math Test, between the *CT* and control groups. When we conducted a further analysis of the Algebraic Operations sub-strand, we found that many students in both conditions did not demonstrate growth in this scale. Again we found that there was no discernible difference between *CT* and control groups in Algebraic Operations. However, for Algebraic Operations outcomes, but not for General Math outcomes, we found a significant interaction between the pre-test and *CT*. That is, students scoring low before participating in *CT* got more benefit from the program's algebraic operations instructions than did students with high initial scores.

A general finding of no difference does not mean that *CT* is ineffective—it appeared to be equally as effective as the existing pre-Algebra program. It is important to interpret these results in relation to what teachers were using in their control classes and to the usage patterns, implementations, and applications in *CT* classes. It is also relevant that this was the first year of use of *CT* for half of the teachers and their initial unfamiliarity may have had an effect on implementation. In addition, this small experiment was not designed to detect differences of less than .35 standard deviation units.[14]

Our communications with the teachers showed a concern with the availability of resources. Our qualitative data sources revealed that they experienced challenges in implementation of *CT* similar to the challenges in the Algebra study. Teachers reported lack of resources, specifically, in the number of classroom computers, access to the computer lab, and *CT* materials. Another challenge that teachers expressed related to the misalignment between the *CT* content and state math standards in middle and high school. Therefore our results must be interpreted in the context of the particular hindrances for the implementation of *CT* in this district, which, as our qualitative data collection methods revealed, were not favorable.

Despite these challenges, teachers (and students) reported a generally positive attitude about the *CT* program overall. Teachers were particularly pleased with how engaged their students were with the *CT* software and the *CT* approach to collaborative learning. It must be noted that 45% of the teachers reported that the *CT* approach to collaborative learning has affected the instructional practices they use in their control classes. This is a form of contamination in a randomized experiment; in this case, we were not able to determine whether it affected our findings.

The district was specifically interested in looking at how the different ethnic groups, particularly the Hawaiian/Part-Hawaiian and Filipino students, performed in math. We examined performance on the NWEA overall test in both groups. We did not find that *CT* had a different effect for different ethnicities. We did notice that, when controlling for pretest score, the estimated effect for the Filipino students is larger than the estimate for the White students on the Algebraic Operations sub-strand. This trend presents interesting implications but requires additional study for confirmation.

In addition, consider the differences in posttest score by initial pretest score. For the Algebraic Operations scale, initially lower scoring students benefited more than the higher scoring students. Since the groups of interest (Filipino and Hawaiian/part-Hawaiian) overall had lower average pretest scores, the results suggest that *CT* may help to reduce the achievement gap between those groups and others.

The district was also interested in learning whether *CT* was effective for students taught by certified teachers versus those with non-certified teachers. In the previous year's study of *Cognitive Tutor for Algebra 1,* control students of certified teachers outperformed control students of non-certified teachers. But the program appeared to have a detrimental effect for certified teachers and no effect for non-certified teachers, both for the overall math scores and for the algebraic outcomes. By contrast, in this experiment on pre-Algebra, we find certified and non-certified teachers performing about the same in their control classes. For the overall score (but not the Algebraic Operations sub-strand) we find that *CT* gave the non-certified teachers an advantage.

Our goal in this research was to provide the Maui School District with evidence that would be useful in determining the impact of *CT* within the local setting. Considered as a district pilot, the study adds to the information available on which to base local decisions. Although our study did not provide evidence of a positive impact of *CT* on student achievement in math in general, we did find some positive effects. Overall, despite the repeated challenges teachers faced in implementation, *CT* was successful in raising student engagement in math and demonstrating, on the Algebra-related sub-strand, gains for previously lower-performing students. The program also appeared to be particularly beneficial for non-certified teachers. These conclusions for teachers can be considered suggestive but not conclusive, since only a small number participated in the study.

This small study illustrates a general caution in interpreting findings from isolated experiments. Our 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

[14] Furthermore, the significant moderating effect assumes fixed classes and teachers in determination of the slopes of the moderator and does not reflect additional uncertainty in the slopes that can result from re-sampling cases at these levels.

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.

Cabalo, J.V., & Vu, M. (2007, May). *Comparative effectiveness of Carnegie Learning's* Cognitive Tutor *Algebra I curriculum: A report of a randomized experiment in Maui School District.* (Empirical Education Rep. No. EEI_EdCT-05-FR-Y1-O.2). Palo Alto, CA: Empirical Education Inc.

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

Carnegie Learning (2005) *Cognitive Tutor Bridge to Algebra Curriculum*. Pittsburgh, PA, Carnegie Learning, Inc.

Hawaii Department of Education (2006). Retrieved on September 1, 2006 from http://doe.k12.hi.us/index.html

Hawaii Department of Education (2007). Henry Perrine Baldwin High School Status and Improvement Report. Retrieved on September 1, 2007 from http://165.248.6.166/data/school.asp?schoolcode=400

Hawaii Department of Education (2007). Lahainaluna High School Status and Improvement Report. Retrieved on September 1, 2007 from http://165.248.6.166/data/school.asp?schoolcode=414

Hawaii Department of Education (2007). Maui High School Status and Improvement Report. Retrieved on September 1, 2007 from http://165.248.6.166/data/school.asp?schoolcode=418

Hawaii Department of Education (2007). Maui Waena Intermediate School Status and Improvement Report. Retrieved on September 1, 2007 from http://165.248.6.166/data/school.asp?schoolcode=428

Hawaii Department of Education (2007). Molokai High School Status and Improvement Report. Retrieved on September 1, 2007 from http://165.248.6.166/data/school.asp?schoolcode=421

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.

U.S. Census Bureau. (2007). 2006 Population Estimate for Maui County. Retrieved on September 1, 2007 from http://www.census.gov/.