U. S. DEPARTMENT OF EDUCATION

Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort

Report to Congress

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort

Report to Congress
March 2007

Mark Dynarski
Roberto Agodini
Sheila Heaviside
Timothy Novak
Nancy Carey
Larissa Campuzano
Mathematica Policy Research, Inc.

Barbara Means
Robert Murphy
William Penuel
Hal Javitz
Deborah Emery
Willow Sussex
SRI International

NCEE 2007-4005
U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

U. S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Evaluation and Regional Assistance
Phoebe Cottingham
Commissioner

March 2007

Prepared under Contract No.: ED-01-CO-0039/0007 with Mathematica Policy Research, Inc.

To order copies of this report,

- Write to ED Pubs, Education Publications Center, U.S. Department of Education, P.O. Box 1398, Jessup, MD 20794-1398.

- Call in your request toll free to 1-877-4ED-Pubs. If 877 service is not yet available in your area, call 800-872-5327 (800-USA-LEARN). Those who use a telecommunications device for the deaf (TDD) or a teletypewriter (TTY) should call 800-437-0833.

- Fax your request to 301-470-1244.

- Order online at www.edpubs.org.

This report also is available on the Department's Web site at http://www.ed.gov/ies.

Upon request, this report is available in alternate formats such as Braille, large print, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Acknowledgments

This study represents a collaborative effort of many school districts, schools, teachers, researchers, data collection experts, publishers and developers, and staff of the Institute of Education Sciences. We appreciate the willingness of the school districts, schools, and teachers to volunteer for the study, use the products, and respond to many requests for data, perspectives, insights, and access to classrooms. We thank Audrey Pendleton (project officer), Ricky Takai, and Phoebe Cottingham at the Institute of Education Sciences for their support and guidance throughout the study. We also thank John Bailey, Susan Patrick, and Timothy Magner of the U.S. Department of Education's Office of Education Technology for their support and suggestions throughout the study. We benefited from the insights and comments of the evaluation's technical working group: Chris Lonigan, Steven Ross, Tom Cook, Jere Brophy, Michael Kamil, Gary Phillips, Doug Clements, and John Cooney.

We thank Christopher Sanford and Frances Bergland for coordinating the classroom observations and monitoring the collection and processing of the interview and observation data, Amy Hafter for the support to the staff conducting observations, Amy Lewis for design and production of instrumentation for the observations, and Paul Hu for analyzing and documenting the observational data. Melissa Thomas was a key contributor to the development of the survey instruments and helped direct the testing and data collection effort. Kristina Quartana, Tom Barton, Valerie Williams, Melissa Dugger, Timothy Bruursema, and Leonard Brown assisted in managing the data collection, and Marianne Stevenson, Francene Barbour, Susan Golden, Season Bedell-Boyle, and Shelby Pollack also helped collect and manage the data. We thank Nakis Evgeniou, Scott Reid, Mark Beardsley, Ronald Palanca, and Roland Scurato for information systems support; Don Jang and Amang Sukasih for statistical support; and Tim Novak, Carol Razafindrakato, and Alex Bogin for programming the estimation models. We appreciate the expert skills of Jill Miller and Bill Garrett to format and produce the report.

The Authors

Contents

Tables

Figures

Exhibits

Executive Summary

Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort

With computers now commonplace in American classrooms, and districts facing substantial costs of hardware and software, concerns naturally arise about the contribution of this technology to students' learning. The No Child Left Behind Act (P.L. 107-110, section 2421) called for the U.S. Department of Education (ED) to conduct a national study of the effectiveness of educational technology. This legislation also called for the study to use "scientifically based research methods and control groups or control conditions" and to focus on the impact of technology on student academic achievement.

In 2003, ED contracted with Mathematica Policy Research, Inc. (MPR) and SRI International to conduct the study. The team worked with ED to select technology products; recruit school districts, schools, and teachers; test students; observe classrooms; and analyze the data. The study used an experimental design to assess the effects of technology products, with volunteering teachers randomly assigned to use or not use selected products.

The main findings of the study are:

1. **Test Scores Were Not Significantly Higher in Classrooms Using Selected Reading and Mathematics Software Products.** Test scores in treatment classrooms that were randomly assigned to use products did not differ from test scores in control classrooms by statistically significant margins.

2. **Effects Were Correlated With Some Classroom and School Characteristics.** For reading products, effects on overall test scores were correlated with the student-teacher ratio in first grade classrooms and with the amount of time that products were used in fourth grade classrooms. For math products, effects were uncorrelated with classroom and school characteristics.

Study Design

Intervention: Sixteen products were selected by ED based on public submissions and ratings by the study team and expert review panels. Products were grouped into four areas: first grade reading, fourth grade reading, sixth grade math, and algebra.

Participants: Thirty-three districts, 132 schools, and 439 teachers participated in the study. In first grade, 13 districts, 42 schools, and 158 teachers participated. In fourth grade, 11 districts, 43 schools, and 118 teachers participated. In sixth grade, 10 districts, 28 schools, and 81 teachers participated, and for algebra, 10 districts, 23 schools, and 71 teachers participated. Districts and schools could participate in the study at more than one grade level, and some did. Districts were recruited on the basis that they did not already use technology products that were similar to study products in participating schools.

Research Design: Within each school, teachers were randomly assigned to be able to use the study product (the treatment group) or not (the control group). Control group teachers were able to use other technology products that may have been in their classrooms. The study administered tests to students in both types of classrooms near the beginning and end of the school year. The study also observed treatment and control classrooms three times during the school year and collected data from teacher questionnaires and interviews, student records, and product records. Because students were clustered in classrooms, and classrooms were clustered in schools, effects were estimated using hierarchical linear models.

Outcomes Analyzed: Student test scores, classroom activities, and roles of teachers and students.

Educational technology is used for word processing, presentation, spreadsheets, databases, internet search, distance education, virtual schools, interactions with simulations and models, and collaboration over local and global networks. Technology also is used as assistive devices for students with disabilities and to teach concepts or skills that are difficult or impossible to convey without technology. This study is specifically focused on whether students had higher reading or math test scores when teachers had access to selected software products designed to support learning in reading or mathematics. It was not designed to assess the effectiveness of educational technology across its entire spectrum of uses, and the study's findings do not support conclusions about technology's effectiveness beyond the study's context, such as in other subject areas.

This report is the first of two from the study. Whether reading and mathematics software is more effective when teachers have more experience using it is being examined with a second year of data. The second year involves teachers who were in the first data collection (those who are teaching in the same school and at the same grade level or subject area) and a second cohort of students. The second report will present effects for individual products. The current report will present effects for groups of products.

Selecting Technology Products for the Study

The study was based on the voluntary participation of technology product developers, districts and schools, and teachers. Their characteristics provide an important part of the study's structure and context for interpreting its findings.

Before products could be selected, decisions were needed about the study's focus. The legislation mandating the study provided general guidelines but did not describe specifically how the study was to be implemented. A design team consisting of U.S. Department of Education staff, researchers from MPR and its partners, and researchers and educational technology experts recommended that the study

- focus attention on technology products that support reading or math instruction in low-income schools serving the K-12 grades;
- use an experimental design to ensure that measured achievement gains could be attributed to products; and
- base the analysis of student academic achievement on a commonly used standardized test.

The team also identified conditions and practices whose relationships to effectiveness could be studied, and recommended a public process in which developers of technology products would be invited to provide information that a panel would consider in its selection of products for the study. A design report provided discussion and rationales for the recommendations.

A total of 160 submissions were received in response to a public invitation made by ED and MPR in September 2003. A team rated the submissions on evidence of effectiveness (based on previous research conducted by the companies or by other parties), whether products could operate on a scale that was suitable for a national study, and whether companies had the capacity to provide training to schools and teachers on the use of their products. A list of candidate products was then reviewed by two external panels (one each for reading and math). ED selected 16 products for the study from among the recommendations made by the panels and announced the choices in January 2004. ED also identified four grade levels for the study, deciding to study reading products in first and fourth grades and math products in sixth grade and in algebra classes, typically composed of ninth graders. Twelve of the 16 products have either received or been nominated to receive awards (some as recently as 2006) from trade associations, media, parents, and teachers. The study did not determine the number of schools, teachers, and students already using the selected products.

The voluntary aspect of company participation in the study meant that products were not a representative sampling of reading and math technology used in schools. Not all products were submitted for consideration by the study, and most products that were submitted were not selected. Also, products that were selected were able to provide at least some evidence of effectiveness from previous research. ED recognized that selecting ostensibly more effective products could tilt the study toward finding higher levels of effectiveness, but the tilt was viewed as a reasonable tradeoff to avoid investing the study's resources in products that had little or no evidence of effectiveness.

The study was designed to report results for groups of products rather than for individual products. Congress asked whether technology was effective and not how the effectiveness of individual products compared. Further, a study designed to determine the

effectiveness of groups of products required fewer classrooms and schools to achieve a target level of statistical precision and thus had lower costs than a study designed to determine the effectiveness of individual products at the same level of precision. Developers of software products volunteered to participate in the study with the understanding that the results would be reported only for groups of products.

During the course of implementing the study, various parties expressed an interest in knowing results for individual products. To accommodate that interest, the design of the study was modified in its second year of data collection. At the same time, product developers were asked to consent to having individual results about their products reported in the second year of data collection. A report of the results from the second year is forthcoming.

Recruiting Districts and Schools for the Study

After products were selected, the study team began recruiting school districts to participate. The team focused on school districts that had low student achievement and large proportions of students in poverty, but these were general guidelines rather than strict eligibility criteria. The study sought districts and schools that did not already use products like those in the study so that there would be a contrast between the use of technology in treatment and control classrooms. Product vendors suggested many of the districts that ultimately participated in the study. Others had previously participated in studies with MPR or learned of the study from news articles and contacted MPR to express interest.

Interested districts identified schools for the study that fell within the guidelines. Generally, schools were identified by senior district staff based on broad considerations, such as whether schools had adequate technology infrastructure and whether schools were participating in other initiatives. By September 2004, the study had recruited 33 districts and 132 schools to participate. Five districts elected to implement products in two or more grade levels, and one district decided to implement a product in all four grade levels, resulting in 45 combinations of districts and product implementations. Districts and schools in the study had higher-than-average poverty levels and minority student populations (see Table 1).

To implement the experimental design, the study team randomly assigned volunteering teachers in participating schools to use products (the "treatment group") or not (the "control group"). Because of the experimental design, teachers in the treatment and control groups were expected to be equivalent, on average, except that one group is using one of the study's technology products. Aspects of teaching that are difficult or impossible to observe, such as a teacher's ability to motivate students to learn, are "controlled" by the experimental design because teachers were randomly assigned, and therefore should be the same in both groups, on average. The study also used statistical methods to adjust for remaining differences in measured characteristics of schools, teachers, and students, which arise because of sampling variability.

Table 1. Sample Size of the Evaluation of the Effectiveness of Reading and Mathematics Software Products

Subject and Grade Level	Number of Districts	Number of Schools	Number of Teachers[a]	Number of Students[b]
Reading (Grade 1)	14	46	169	2,619
Reading (Grade 4)	11	43	118	2,265
Math (Grade 6)	10	28	81	3,136
Math (Algebra)	10	23	71	1,404
Total	45	140	439	9,424
Unduplicated Total[c]	33	132	439	n.a.

[a]The number of teachers includes the treatment and control teachers.

[b]The number represents students in the analysis sample who were tested in fall 2004 and in spring 2005. The total number of students who were tested at either point in time is larger because some students tested in the fall moved out of their school district by the time of the spring test and some students tested in the spring had moved into study classrooms after the fall test. The total number of students tested was 10,659 in the fall and 9,792 in the spring.

[c]Because nine districts and eight schools are piloting more than one product for the study, the unduplicated total gives the number of unique districts and schools in the study.

n.a. = not applicable.

The experimental design provides a basis for understanding whether software products improve achievement. Teachers in the treatment group were to implement a designated product as part of their reading or math instruction. Teachers in the control group were to teach reading or math as they would have normally, possibly using technology products already available to them. Because the only difference on average between groups is whether teachers were assigned to use study products, test-score differences could be attributed to being assigned to use a product, after allowing for sampling variability.

Because the study implemented products in real schools and with teachers who had not used the products, the findings provide a sense of product effectiveness under real-world conditions of use. While the study worked to ensure that teachers received appropriate training on using products and that technology infrastructures were adequate, vendors rather than the study team were responsible for providing technical assistance and for working with schools and teachers to encourage them to use products more or use them differently. Teachers could decide to stop using products if they believed products were ineffective or difficult to use, or could use products in ways that vendors may not have intended. Because of this feature of the study, the results relate to conditions of use that schools and districts would face if they were purchasing products on their own.

Collecting Achievement and Implementation Data

The study's analyses rely mostly on data from student test scores, classroom observations, and teacher questionnaires and interviews. The study also collected student data items from school district records and incorporated data about districts and schools from the National Center for Education Statistics' *Common Core of Data*.

To measure effects, the team administered a student test in the fall and spring of the 2004-2005 school year. The team used the Stanford Achievement Test (version 9) reading battery for first graders, the Stanford Achievement Test (SAT-10) reading battery for fourth

graders, and the SAT-10 math battery for sixth graders. These tests were administered in fall 2004 and spring 2005. The team also used the Test of Word Reading Efficiency (TOWRE), a short and reliable one-on-one test of reading ability, for first graders to augment measures of reading skills provided by the SAT-9 (Torgesen et al. 1999).

To measure algebra achievement, the study selected Educational Testing Services' (ETS) End-of-Course Algebra Assessment (1997). Because baseline measures of algebra knowledge were not available or were considered unsatisfactory, the study worked with ETS to separate its assessment, which essentially is a final exam, into two components that had equal levels of difficulty. The study randomly selected classrooms either to take part A in the fall and part B in the spring or to take B in the fall and A in the spring. Splitting the test in this way meant that the full test was administered in both the fall and the spring, but each student took only half of the test at each point.

The team also collected scores on district achievement tests if these data were available. The study's administration of its own test provided a consistent measure of achievement across varied districts and schools, but examining findings based on district tests provided a useful check on the robustness of the findings.

Classroom observations were the study's primary basis for assessing product implementation. An observation protocol was developed in spring 2004, and videotapes of classrooms using products were gathered and later used for observer training. The observation protocol was designed to gather similar information in both treatment and control classrooms and across the different grade levels and subject areas in the study. In addition, the protocol was designed to focus on elements of instruction and implementation that could be observed reliably. Each classroom was visited three times during the school year, and observers used the protocol for each observation, which lasted about 1 hour. Observations were complemented by a teacher interview that gathered information about implementation issues. Background information about teachers was also gathered from a questionnaire that teachers completed in November and December 2004.

Summary of Study Findings

The four grade levels essentially comprise substudies within the overall study, and findings are reported separately for each. The study's data collection approach was the same for the four substudies.

The implementation analysis focused on how products were used in classrooms, their extent of usage, issues that resulted from their use, and how their use affected classroom activities. Three implementation findings emerged consistently across the four substudies:

1. **Nearly All Teachers Received Training and Believed the Training Prepared Them to Use the Products.** Vendors trained teachers in summer and early fall of 2004 on using products. Nearly all teachers attended trainings (94 percent to 98 percent, depending on the grade level). At the end of trainings, most teachers reported that they were confident that they were prepared to use the products with their classes. Generally, teachers reported a lower degree of confidence in what they had learned after they began using products in the classroom.

2. **Technical Difficulties Using Products Mostly Were Minor.** Minor technical difficulties in using products, such as issues with students logging in, computers locking up, or hardware problems such as headphones not working, were fairly common. Most of the technical difficulties were easily corrected or worked around. When asked whether they would use the products again, nearly all teachers indicated that they would.

3. **When Products Were Being Used, Students Were More Likely to Engage in Individual Practice and Teachers Were More Likely to Facilitate Student Learning Rather Than Lecture.** Data from classroom observations indicated that, compared to students in control classrooms where the same subject was taught without using the selected products, students using products were more likely to be observed working with academic content on their own and less likely to be listening to a lecture or participating in question-and-answer sessions. Treatment teachers were more likely than control teachers to be observed working with individual students to facilitate their learning (such as by pointing out key ideas or giving hints or suggestions on tackling the task students were working on) rather than leading whole-class activities.

Comparing student test scores for treatment teachers using study products and control teachers not using study products is the study's measure of product effectiveness. Effects on test scores were estimated using a statistical model that accounts for correlations of students within classrooms and classrooms within schools. The robustness of the results was assessed by examining findings using different methods of estimation and using district test scores as outcomes, and the patterns of findings were similar.

Effects of First Grade Technology Products

The first grade study was based on five reading software products that were implemented in 11 districts and 43 schools. The sample included 158 teachers and 2,619 students. The five products were Destination Reading (published by Riverdeep), the Waterford Early Reading Program (published by Pearson Digital Learning), Headsprout (published by Headsprout), Plato Focus (published by Plato), and the Academy of Reading (published by Autoskill).

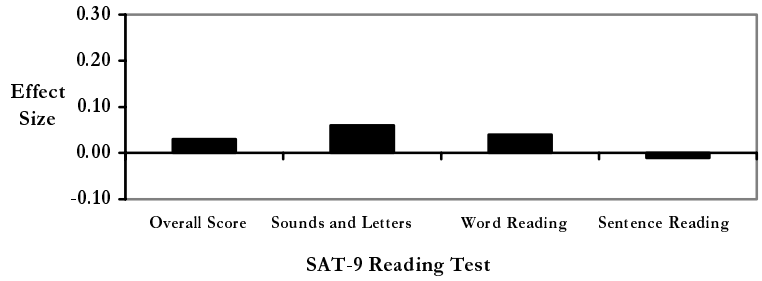
Products provided instruction and demonstration in tutorial modules, allowed students to apply skills in practice modules, and tested students on their ability to apply skills in assessment modules. (The tutorial-practice-assessment modular structure was common for products at other grade levels as well.) Their focus was on improving skills in letter recognition, phonemic awareness, word recognition and word attack, vocabulary building, and text comprehension. The study estimated the average licensing fees for the products to be about $100 a student for the school year, with a range of $53 to $124.

According to records maintained by product software, usage by individual students averaged almost 30 hours a year, which the study estimated to be about 11 percent of reading instructional time. Some control group teachers used technology-based reading products that were not in the study. These products generally allowed students to practice

various skills. Software-based records of student usage of these other products were not collected, but control teachers reported using them about a fifth as much as treatment teachers reported using study products.

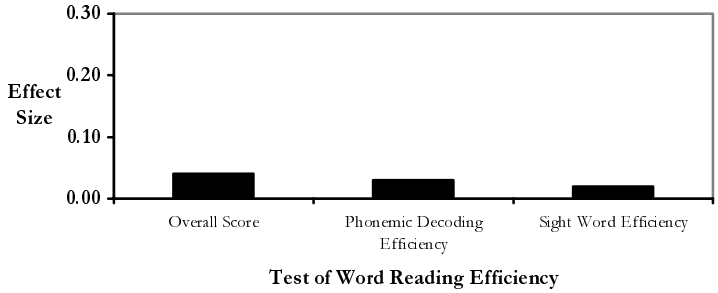
First grade reading products did not affect test scores by amounts that were statistically different from zero. Figure 1 shows observed score differences on the SAT-9 reading test, and Figure 2 shows observed score differences on the Test of Word Reading Efficiency. The differences are shown in "effect size" units, which allow the study to compare results for tests whose scores are reported in different units. (The study's particular measure of effect size is the score difference divided by the standard deviation of the control group test-score.) Effect sizes are consistent for the two tests and their subtests, in the range of -0.01 to 0.06. These effect sizes are equivalent to increases in student percentile ranks of about 0 to 2 points. None is statistically significant.

Figure 1
Effects of First Grade Reading Products on SAT-9 Reading Score



Note: None of the effects is significantly different from zero at the 0.05 level.

Figure 2
Effects of First Grade Reading Products on Test of Word Reading Efficiency Score



Note: None of the effects is significantly different from zero at the 0.05 level

Large differences in effects were observed between schools. Because only a few teachers implemented products in each school, sampling variance (the assignment of teachers to treatment and control groups) can explain much of the observed differences, but the study also investigated whether the differences were correlated with school and classroom characteristics. Relationships between school and classroom characteristics and score differences cannot be interpreted as causal, because districts and schools volunteered to participate in the study and to implement particular products. Their characteristics (many of which the study did not observe) may influence observed effects. For first grade, effects were larger when schools had smaller student-teacher ratios (a measure of class size). Other characteristics, including teacher experience and education, school racial-ethnic composition, and the amount of time that products were used during the school year, were not correlated with effects.

Effects of Fourth Grade Reading Products

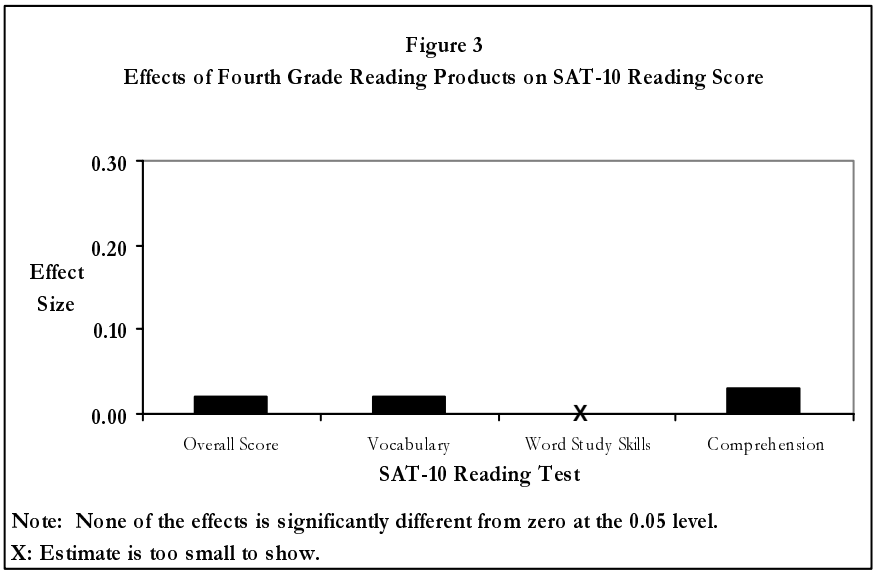
The fourth grade study included four reading products that were implemented in nine districts and 43 schools. The sample included 118 teachers and 2,265 students. The four products were Leapfrog (published by Leaptrack), Read 180 (published by Scholastic), Academy of Reading (published by Autoskill), and KnowledgeBox (published by Pearson Digital Learning).

Three of the four products provided tutorials, practice, and assessment geared to specific reading skills, one as a core reading curriculum and two as supplements to the core curriculum. The fourth product offered teachers access to hundreds of digital resources such as text passages, video clips, images, internet sites, and software modules from which teachers could choose to supplement their reading curriculum. The study estimated the average licensing fees for the products to be about $96 a student for the school year, with a range of $18 to $184.

Annual usage by students for the two fourth grade products that collected this measure in their databases was 7 hours for one product and 20 for the other. Assuming a typical reading instruction period was 90 minutes, students used products for less than 10 percent of reading instructional time (this estimate refers to the computer-based component of products). Treatment teachers also reported scheduling 6 hours of use of other products during the school year, and control teachers reported scheduling 7 hours of use of other products. Treatment teachers also reported spending 1 hour more a week teaching reading than control teachers (the increase was statistically significant).

Fourth grade reading products did not affect test scores by amounts that were statistically different from zero. Figure 3 shows measured effect sizes for the SAT-10 reading test, in effect size units.

Most school and classroom characteristics were not correlated with effects, but effects were larger when teachers reported higher levels of product use. As noted above, these relationships do not have a causal interpretation.



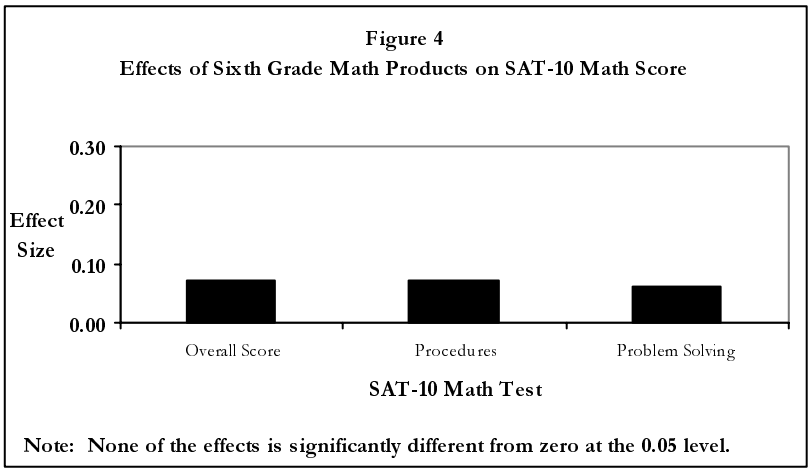
Effects of Sixth Grade Math Products

The sixth grade study included three products that were implemented in 10 districts and 28 schools. The sample included 81 teachers and 3,136 students. The three products were Larson Pre-Algebra (published by Houghton-Mifflin), Achieve Now (published by Plato), and iLearn Math (published by iLearn).

Products provided tutorial and practice opportunities and assessed student skills. Topics covered include operations with fractions, decimals, and percents; plane and coordinate geometry; ratios, rates, and proportions; operations with whole numbers and integers; probability and data analysis; and measurement. Two products were supplements to the math curriculum, and one was intended as a core curriculum. The study estimated the average licensing fees for the products to be about $18 a student for the school year, with a range of $9 to $30.

Student usage was about 17 hours a year, or about 11 percent of math instructional time, according to data from product records (available for two of the three products). In control classrooms, teachers reported about 3 hours of use of other technology products, which was much less than the 51 hours of study product usage reported by treatment teachers.

Sixth grade math products did not affect test scores by amounts that were statistically different from zero (see Figure 4). As with other products, the study observed large effects between schools. However, statistical tests indicated that the school and classroom characteristics measured in the study were not related to the differences in test scores.



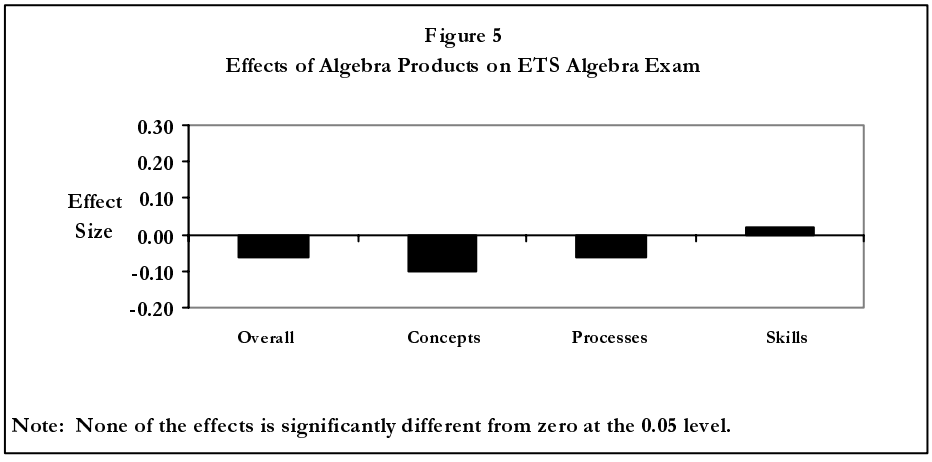
Effects of Algebra Products

The algebra study included three products that were implemented in 10 districts and 23 schools. The sample included 69 classrooms and 1,404 students. The three products were Cognitive Tutor Algebra (published by Carnegie Learning), Plato Algebra (published by Plato), and Larson Algebra (published by Houghton-Mifflin).

Products covered a conventional range of algebra topics. They included functions, linear equations, and inequalities; quadratic equations; linear expressions; polynomials; and so on. One product constituted a full curriculum, and the majority of its activities were carried out in "offline" class periods. The other two were supplements to the regular curriculum. The study estimated the average licensing fees for the products to be about $15 a student for the school year, with a range of $7 to $30.

Product records showed that student usage was 15 hours for the overall sample, equivalent to about 10 percent of math instructional time. Usage averaged 5 to 28 hours, depending on the product.

Algebra products did not affect test scores by amounts that were statistically different from zero (see Figure 5). As with products in the other grade levels, the study observed large differences in effects between schools, but statistical tests indicated that the school and classroom characteristics measured in the study were not related to these differences.



Summary

Congress posed questions about the effectiveness of educational technology and how effectiveness is related to conditions and practices. The study identified reading and mathematics software products based on prior evidence of effectiveness and other criteria and recruited districts, schools, and teachers to implement the products. On average, after one year, products did not increase or decrease test scores by amounts that were statistically different from zero.

For first and fourth grade reading products, the study found several school and classroom characteristics that were correlated with effectiveness, including student-teacher ratios (for first grade) and the amount of time products were used (for fourth grade). The study did not find characteristics related to effectiveness for sixth grade math or algebra. The study also found that products caused teachers to be less likely to lecture and more likely to facilitate, while students using reading or mathematics software products were more likely to be working on their own.

The results reported here are based on schools and teachers who were not using the products in the previous school year. Whether products are more effective when teachers have more experience using them is being examined with a second year of data. The study will involve teachers who were in the first data collection (those who are teaching in the same school and at the same grade level or subject area) and a new group of students. The second-year study will also report results separately for the various products.

Chapter I

Introduction

In the No Child Left Behind Act of 2002, Congress called for a rigorous study of the effectiveness of educational technology for improving student academic achievement. The call for the study was consistent with the growing use of computers in classrooms, demands for accountability in public spending, and the Act's emphasis on using rigorous research methods to study effectiveness. This report presents findings from the study's first year of product implementation and data collection, which corresponded to the 2004-2005 school year.

A. Research on the Effectiveness of Reading and Mathematics Software

Use of software to help teach reading and mathematics skills is common in American schools. In the late nineties, 56 percent of elementary school teachers reported using software products designed to teach English and language arts skills. Similarly, 62 percent of elementary school teachers, 39 percent of middle school teachers, and 22 percent of high school teachers reported using products designed to teach math skills (Anderson and Ronnkvist 1999).

Over the past two decades, numerous studies comparing computer-based and conventional instruction in reading and mathematics have been conducted. Both qualitative research syntheses (Schacter 2001; Sivin-Kachala 1998) and formal meta-analyses of these studies (Blok et al. 2002; Kulik and Kulik 1991; Kulik 1994; Kulik 2003; Murphy et al. 2001; Pearson et al. 2005; Waxman et al. 2003) found that computer-assisted instruction in reading and mathematics generally had a positive effect. Kulik's 1994 meta-analysis, for example, found a positive effect on final examination scores (the effect size—the effect as a proportion of the standard deviation of examination scores—was 0.30).

Murphy et al. (2001) examined a wide range of research studies from the published literature and from software vendors. Of the 195 experimental or quasi-experimental studies conducted between 1993 and 2000 that met the criteria for inclusion, 31 studies met minimum methodological requirements for inclusion in the synthesis. For these studies, researchers estimated an average effect size of .35 for reading and .45 for mathematics.

Despite the fairly sizable number of studies and generally positive findings, meta-analysts have noted that many studies contained weaknesses or design flaws (Murphy et al. 2001; Pearson et al. 2005). Of the technology studies reviewed by Waxman et al. (2003), for example, half had sample sizes of fewer than 50 students. Many studies had no control groups or equivalent comparison groups, leading to questionable validity for claims of effects. Studies with stronger research designs showed smaller effects (Pearson et al. 2005).

Several recent experimental studies examined product effectiveness and reached different conclusions. Rouse and Krueger (2004) evaluated the effectiveness of *Fast ForWord*, a software application based on a set of exercises that neuroscientists had found to produce dramatic reading gains for some children (Merzenich et al. 1996; Tallal et al. 1996). Rouse and Krueger found a small positive effect for *Fast ForWord* on a computer-based measure of language skills but no effect on reading achievement measured using a standardized test. In contrasting their findings with previously reported findings, Rouse and Krueger noted problems in the design of a study by researchers affiliated with the company that distributes *Fast ForWord*. In that study, students were assigned randomly to the treatment group or a control group (Miller et al. 1999), but students who did not complete the treatment (based on a definition of completion provided by the product vendor) were dropped from the sample, thereby invalidating its experimental design.

In contrast, a randomized field trial of *Algebra Cognitive Tutor* reported positive results (Morgan and Ritter 2002). In the study, eight junior high teachers taught some of their classes using the Cognitive Tutor and some using their traditional textbook. Students in the Cognitive Tutor classes scored higher on an end-of-course algebra test developed by Educational Testing Service and also received higher course grades.

Similarly, Nunnery et al. (2006) found positive effects for *Accelerated Reader*, a software product that recommends reading material at a level appropriate for the individual student and provides computer-based quizzes testing student comprehension of the recommended materials. The study randomly assigned 45 teachers from nine elementary schools in a large urban district to use the product or to a control condition.

Thus, while a preponderance of research on the effectiveness of reading and math software suggests positive effects, many studies have been small, methodologically flawed, or sponsored by an organization with an interest in the outcome. More recent studies using rigorous designs have yielded mixed findings. Against this backdrop, a large-scale study of educational technology products can provide useful information about their effectiveness.

B. Design of the National Study

In fall 2002, the U.S. Department of Education (ED) began working with Mathematica Policy Research, Inc. (MPR) and its partners to design the national study called for by Congress. Key recommendations from the design effort were to focus the national study on grades K-12 in schools that served large percentages of students in poverty, to focus on the reading and math subject areas, and to use standardized test scores to measure effectiveness (Agodini et al. 2003). The design effort also recommended that a public submissions process

be used to select the technology products included in the evaluation. The study's main features are summarized in the accompanying box.

The legislation called for the study to have "control groups or control conditions," which, consistent with current practice, was interpreted to mean that the study should use an experimental design. The study used a design in which teachers who were not using one of the study's technology products volunteered to participate in the study. Within each school, these teachers were then randomly assigned to either a treatment group that had access to the assigned product or to a control group that used their conventional teaching approaches. The experimental design was the basis for answering the study's main question: "Do students achieve more when teachers are able to use the selected technology products than when they do not?" Because the only difference, on average, between groups is whether teachers were assigned to use products selected for the study, score differences could be attributed as effects of the products, after accounting for sampling variability. Score differences could be affected by whether control group teachers used other technology products that were not among those selected for the study, an issue that was recognized in the design and will be addressed in the analysis.

The study's main question is equivalent to the question faced by school districts wanting to raise student test scores and considering investing in a technology product to do so: Does purchasing the product lead to higher scores? The study tested whether students in classrooms of treatment group teachers, who were able to use products selected to be in the study, performed better or worse than students in classrooms of control group teachers, who were not able to use those particular products (but may have used other products). In

Study Design

Intervention: Sixteen products were selected by ED based on public submissions and ratings by the study team and expert review panels. Products were grouped into four areas: first grade reading, fourth grade reading, sixth grade math, and algebra.

Participants: Thirty-three districts, 132 schools, and 439 teachers participated in the study. In first grade, 13 districts, 42 schools, and 158 teachers participated. In fourth grade, 11 districts, 43 schools, and 118 teachers participated. In sixth grade, 10 districts, 28 schools, and 81 teachers volunteered, and for algebra, 10 districts, 23 schools, and 71 teachers participated. Districts and schools could participate in the study at more than one grade level, and some did. Districts were recruited on the basis that they did not already use technology products that were similar to study products in participating schools.

Research Design: Within each school, teachers were randomly assigned to be able to use the study product (the treatment group) or not (the control group). Control group teachers were able to use other technology products that may have been in their classrooms. The study administered tests to students in both types of classrooms near the beginning and end of the school year. The study also observed treatment and control classrooms three times during the school year and collected data from teacher questionnaires and interviews, student records, and product records.

Outcomes Analyzed: Student test scores, classroom activities, and roles of teachers and students.

adopting this approach, the national study's design essentially is similar to the designs of many studies of product effectiveness cited above.[1]

The study tests whether selected reading and mathematics products are effective when districts volunteer to participate and schools and teachers volunteer to implement products. The voluntary aspect of the study and the fact that districts and schools were participating in a study may introduce a difference between measured effectiveness reported by the study, effectiveness in actual use, and effectiveness reported by other studies. For example, effectiveness as measured by this study might be higher than effectiveness in actual use because schools and teachers volunteered for the study, and the study also purchased software and hardware for classrooms and schools when upgrades were needed for products to operate. In actual use, products might be placed in classrooms of teachers who do not want to use them and may not use them effectively, which presumably would imply lower levels of effectiveness.

Effectiveness as measured by the study might be lower than effectiveness reported by other studies because products were implemented by teachers who had not used these products before. As later chapters note, some schools and classrooms encountered various difficulties in starting to use or continuing to use products. Teachers could stop using products or reduce their use if teachers believed products were ineffective or difficult to use, or they could use products in ways that vendors may not have intended or predicted. Teachers in control classrooms did not have access to products in the study but could use computers in other ways, such as for web browsing, for office-related functions, and to operate other products. (Later chapters analyze the extent to which they did so.)

Effectiveness as measured by the study might be lower than what has been reported by other studies because this study used an experimental design. Because teachers were randomly assigned as part of the experiment, factors that may have predisposed teachers to use products more effectively and that confound actual effectiveness with teacher characteristics are controlled by the experimental design.

Selecting Products for the Evaluation

In fall 2003, developers and vendors of educational technology products responded to a public invitation and submitted products for possible inclusion in the national study. MPR staff selected 40 of the 160 submissions for further review by two panels of outside experts, one for reading products and one for math products. The three main criteria of the reviews were whether the product had evidence of effectiveness (and, related to this, the validity of the evidence), whether the product was able to operate on a national scale, and whether product developers could train an adequate number of teachers for the national study's

[1]Another possible test of technology's effectiveness would be to introduce computers (the hardware part of educational technology) in classrooms and assess whether test scores or other measures of student learning increase. Answering this question would have required focusing the study on the small number of schools and classrooms that had not yet introduced computers, which would reduce the study's relevance to the much larger number of schools that already used computers.

sample sizes. The panels did not review and compare the instructional features of the products. MPR informed the panel members that they were not limited to MPR's list of 40 products but could review any of the 160 submissions.

In January 2004, ED considered the panel's recommendations and selected 16 products for the study. In selecting products, ED grouped them into four areas: (1) early reading (first grade), (2) reading comprehension (fourth grade), (3) pre-algebra (sixth grade), and (4) algebra (ninth grade).[2] The products ranged widely in their instructional approaches and how long they had been in use. In general, however, the criteria weighted the selection toward products that had evidence of effectiveness from previous research, or, for newer products, evidence that their designs were based on approaches found to be effective by research. Twelve of the 16 products had received awards or been nominated for awards (some as recently as 2006) by trade associations, media, teachers, or parents. The study did not determine the total number of schools, teachers, and students currently using the products.

In the submission process, ED informed developers that the study would focus on the average effectiveness of reading and mathematics software products rather than on the effectiveness of individual products. The intent was to measure whether technology, as represented by the selected reading and mathematics software products, improved academic achievement, rather than how individual products increased achievement. The main findings from the national study were to be based on the combined product results at each of the four grade levels.

The voluntary aspect of company participation in the study meant that products may not be a representative sampling of reading and math technology products that schools could purchase or use. Not all products were submitted for potential inclusion in the study, and most products that were submitted were not selected. Products that were selected may have been more effective than the average product because selected products were able to provide at least some evidence of effectiveness from previous research or a proof of concept. ED recognized that selecting ostensibly more effective products possibly tilted the study toward finding higher levels of effectiveness, but the tilt was viewed as a reasonable tradeoff to avoid investing the study's resources in products that had little or no evidence of effectiveness.

C. Recruiting Districts and Schools for the Study

Sample size targets were set so that the study could detect an effect size on test scores of at least 0.25 at each of its four grade levels (Agodini et al. 2003). This target effect size was a balance between the desire for interventions to substantially close achievement gaps and the reality that few large-scale education evaluations have found effects larger than the

[2]Although many students take algebra in their first year of high school, districts increasingly are moving to have students take algebra in eighth grade. Students in upper years of high school also can take algebra.

target.[3] The team estimated that the study needed to include about 120 schools. These schools, it was estimated, would yield 480 classrooms, 120 in each grade level. Assuming a teacher had an average of 20 students, the sample size target was 2,400 in each grade level.

To be consistent with the No Child Left Behind legislation and to support other study objectives, the study used three criteria to identify potential districts and schools: (1) geographical diversity, (2) high poverty rates, and (3) enough teachers volunteering for the study. As a general guideline, the study wanted geographically diverse districts. The study also wanted districts that had six or more schools receiving Title I funds and, within districts, schools that had high poverty rates. Discretion was exercised for small districts, which may have had fewer than six schools. The poverty rate of a school was not the only factor considered, because some schools with high poverty rates may have been inappropriate for other reasons (such as a lack of technology infrastructure or a lack of interest in participating in the study). Schools also needed to have at least two volunteering teachers at the appropriate grade or subject so that random assignment of teachers within each school could be implemented.

In February 2004, the study staff began contacting school districts to learn whether they were interested in participating in the study and met the study criteria. Developers nominated about 85 percent of the almost 200 districts that the study team contacted. Other nominations came from previous contacts with the study team and from self-nominations by districts and schools that had learned about the study from articles in the media. Study staff visited most of the districts that expressed interest in participating in the study to describe it and to answer questions. In interested districts, administrators worked within the district to identify schools suitable for the study.

As noted above, the study considered schools to be more desirable for the study if their teachers were not already using the reading and mathematics software products in the study (or close substitutes for them). Classrooms in which there were low-intensity computer uses, such as word processing or web browsing, were more desirable because the introduction of a study product would increase the intensity of technology being used for instruction. Later chapters report much lower rates of computer use in the control group than the treatment group, consistent with this approach for identifying appropriate schools. Though schools that participated in the study were not using the products selected for the study, teachers that participated in the study may have used products in the past or at some other school. The study did not gather data about the frequency of teachers with previous experiences using products.

The study assumed that districts would implement only one product, but about a third elected to implement more than one. Because the four grade levels essentially form separate substudies, having the same district implement more than one product did not cause a problem for the study, and it created some cost efficiencies because data collection was more

[3]The target effect size is consistent with the studies noted above. A report from the President's Committee of Advisors on Science and Technology observed that an effect size of 0.25 was at the lower end of the range of effects reported by four meta-analyses of technology studies (PCAST 1997).

clustered. The team encouraged districts that wanted to implement more than one product to implement a second or third product in a different grade level, so that one district would not provide a disproportionately large share of the data for any one grade level. In the end, one district implemented a product in all four grade levels, and another implemented a product in three grade levels. One district implemented two products in the fourth grade level.

By June 2004, nearly all districts had been identified for the study. Agreements to participate were reached by September 2004. A total of 33 districts and 132 schools agreed to participate (Table I.1). The 33 districts are in 19 states, including the populous states of California, Florida, Georgia, Illinois, New Jersey, and Texas. One district (with three schools) later dropped out of the study because technical problems prevented the product the district was attempting to implement from functioning, and one school in another district dropped out. The target number of schools and classrooms was exceeded for grades 1 and 4, but it was not reached for grade 6 and algebra. Because secondary schools have many class sections, the number of students in the study was closer to the target sample size. The target student sample size was exceeded in grade 6. At the actual sample sizes, minimum detectable effect sizes were 0.09, 0.09, 0.13, and 0.12 in the four grade levels, respectively.[4]

Table I.2 shows that the study's emphasis on high-poverty schools resulted in districts having a higher percentage of students eligible for free or reduced-price lunch than the average district. Free and reduced-price lunch rates were 44 percent for schools in the reading substudies and 57 percent for the math substudies, compared to 36 percent nationwide.[5] The study districts also were more likely to be in urban locations (38 percent of districts in the study compared to about 9 percent of districts nationwide) and were larger than the average district in several measures (for example, districts in the reading substudies had about 79 schools and in the math substudies about 126 schools, compared to about 6 schools in the average district). Similarly, the particular schools recruited for the study were more likely to be Title I schools and in urban areas (Table I.3). Consistent with the high percentage of urban schools, study schools had larger enrollments and larger minority student populations than the average school. Schools implementing reading products were similar to schools implementing math products but generally had fewer students (most were elementary schools, whereas nearly all schools implementing math products were middle or high schools).

Schools in the fourth grade study were more highly urbanized and had larger minority student populations than schools in the study's other grade levels.

[4]Minimum detectable effect sizes depend on the distribution of the variance of test scores between students, classrooms, and schools, after accounting for observed student, classroom, and school characteristics. The minimum detectable effect sizes noted in the text are based on estimates of student, classroom, and school residual variances presented in Appendix Table B.1. Detectable effect sizes also depend on the number of teachers assigned to treatment status. The detectable effect sizes are based on the actual treatment assignment rates of 56 percent in the first grade, 53 percent in the fourth grade, 58 percent in the sixth grade, and 55 percent in algebra. The next section describes why the treatment assignment rate differed between grade levels.

[5]The study includes two very large districts. This fact contributes significantly to these characteristics because the two districts contain many high-poverty schools.

Table I.1. Number of Study Districts, Schools, Teachers, and Students by Grade Level, Spring 2005.

Subject and Grade Level	Number of Districts	Number of Schools	Number of Teachers[a]	Number of Students[c]
Reading (Grade 1)	13	42	158	2,619
Reading (Grade 4)	11	43	118	2,265
Math (Grade 6)	10	28	81	3,136
Math (Algebra)	10	23	71	1,404
Total	**44**	**136**	**428**	**9,424**
Unduplicated Total[b]	**33**	**132**	**n.a.**	**n.a.**

[a]The number of teachers includes treatment and control teachers.

[b]The unduplicated total gives the number of unique districts and schools in the study. Nine districts and eight schools piloted more than one product for the study.

[c]The number represents students in the analysis sample tested in both fall 2004 and spring 2005. The total number of students in the study is larger because some students tested in the fall moved out of their school district by the time of the spring test, and some students tested in the spring had moved into study classrooms after the fall test. The total number of students tested was 10,659 in the fall and 9,792 in the spring.

n.a. = not applicable.

Table I.2. Characteristics of Districts in the Study.

Characteristics[a]	Average U.S. District	Districts in the Reading Study	Districts in the Math Study
Number of Title I schools[b]	3.3	34.8	78.5
District location (percentage)			
Urban	8.7	38.1	37.5
Urban fringe	24.9	52.4	43.8
Town	14.7	4.8	6.3
Rural area	51.7	4.8	12.5
Number of schools per district	5.9	78.6	126.4
Number of full-time teachers per district	170	3,642	5,828
Number of students per district	2,988	61,660	103,426
Percentage of students eligible for free or reduced-price lunch[c]	36.1	44.4	56.6
Number of Districts	**15,417**	**21**	**16**

Source: Study tabulations by MPR from the 2001–2002 *Common Core of Data.*

Note: Four districts are in both the reading and math substudies.

[a]Data include districts with one or more regular schools.

[b]Data missing for 6 percent of study districts and 9 percent of districts nationwide.

[c]Data missing for 6 percent of study districts and 10 percent of districts nationwide.

Table I.3. Characteristics of Schools in the Study.

Characteristics[a]	Average U.S. School	Schools in First Grade Study	Schools in Fourth Grade Study	Schools in Sixth Grade Study	Schools in Algebra Study
School location (percentage)					
Urban	24	45	52	36	55
Urban fringe	32	45	48	43	45
Town	12	0	0	4	0
Rural area	32	10	0	18	0
Students per teacher	16	16	16	15	15
Number of students per school	543	626	572	1,073	1,352
Percentage receiving Title I	59	76	88	64	23
Percentage of students eligible for free or reduced-price lunch	42	49	64	71	54
Student race/ethnicity (percentage)					
White	64	44	17	21	29
Black	15	31	57	33	45
Hispanic	15	22	23	42	19
Asian	3	2	3	3	7
Native American	3	<1	<1	<1	<1
Number of Schools[b]	**88,542**	**46**	**43**	**28**	**23**

Source: Study tabulations by MPR from the 2003–2004 *Common Core of Data* (CCD).

[a]Data include regular schools only.

[b]CCD data are missing for 10 study schools.

D. Recruiting and Assigning Teachers

Teachers in participating schools were asked to volunteer by signing a consent form indicating they understood that they would be part of a research study and would implement the product if selected.

In eligible schools (those with two or more volunteering teachers), the study randomly assigned teachers to the product for that school (Figure A.1 in Appendix A shows the flow of teachers into the treatment and control groups). The study randomly assigned 526 teachers and later dropped 98 teachers from the study. The most common reasons for dropping teachers were that teachers who had been randomly assigned were later assigned to teach a different grade level or subject, moved to a different school, retired, or left teaching.

Whether excluded teachers were replaced depended on whether a school could identify a new teacher for the study. If it did, the study either conducted random assignment again (if only one teacher was in the treatment group or control group) or assigned the additional teacher with a 50 percent probability to use the product. Schools were dropped from the study if they had one teacher in either the treatment group or the control group, lost a teacher, and could not replace that teacher. The weighting toward the treatment group ultimately resulted in 56 percent of study teachers being in the treatment group.

E. Collecting Classroom and Student Data

A review of the research literature on the implementation of classroom-based instructional technologies and descriptions of recommended implementation practices provided by software vendors were used to identify the conditions and practices to be measured (Agodini et al. 2005). Relevant literature on the implementation dimensions on which data were collected is summarized briefly below, followed by a brief discussion of data collection methods. Appendix A provides details about the data collection methods.

Research Evidence on Important Implementation Dimensions

Teacher Training and Actions. Using technology products in classrooms places demands on teachers' time and skills. Teachers must prepare the product for student use, monitor and help students as they use the product, maintain the technology, and monitor student progress on it. Since the 1990s, observers have reported that typically only a minority of teachers receives adequate training to manage student use of technology in their classrooms (Kerr 1996; U.S. Congress, Office of Technology Assessment 1995). Recent self-reports from teachers are consistent with these findings. In a recent survey, 86 percent of teachers said that they had a medium or high need for professional development on "how to manage classroom activities that integrate technology" (Adelman et al. 2002).

Amount of Software Use. Tracking usage was important for the study, because research has found that the average classroom does not use products for the amount of time vendors recommend. VanDusen and Worthen (1994) attributed small impacts of integrated learning systems to the fact that the systems were not used for enough time. A recent evaluation of the Waterford Early Reading software in the Los Angeles Unified School District found that first graders spent only 30 percent of the amount of time with the software that Waterford recommended (Hansen et al. 2004).

Locus and Grouping for Product Use. Issues of limited teacher and student access to working computers are often cited as barriers to the integration of technology (Becker et al. 1999; Ertmer 1999; Leggett and Persichitte 1998; Means and Olson 1995). Several studies have reported that barriers to accessing technology are a source of teacher misgivings about using computers for instruction (Adelman et al. 2002; Cuban 2001; Sheingold and Hadley 1990), and teachers cite the inconvenience of scheduling and moving students to a computer lab as reasons for spending less time on technology-dependent curricula (Adelman et al. 2002). Moreover, a study of a statewide implementation of basic skills software found larger achievement gains for math and reading software in schools where the software was used in regular classrooms rather than in computer labs (Mann et al. 1998).

Technical Difficulties and Teacher Support. Computer hardware can be unreliable, computer networks unstable, and technical support inadequate to keep all machines running properly (Cuban 2000; Culp et al. 2003). As a result, on a given day, teachers and students may find themselves without enough working computers and peripherals (such as printers) to effectively use the product. The type and frequency of technical difficulties and other access problems were tracked for each classroom. Access was expected to be less of an issue for the study, because it provided resources to ensure that enough working computers were available. Research also has suggested that teachers need ongoing technical and pedagogical support to use technology well (Adelman et al. 2002; Mills and Ragan 2000). The frequency of technology use has been noted to be associated more with quality of support for integrating technology with curriculum than with the perceived quality of technical troubleshooting and maintenance support. Both of these types of support and teachers' satisfaction with them were measured in the present study.

Product's Role in Instruction. Products can be the core curriculum, provide units that replace some units of the core curriculum, supplement instruction, or provide opportunities to practice beyond those provided by core curriculum materials. Teachers' knowledge of the subject area and the extent to which they view the product to be aligned with the curriculum may influence the extent to which they successfully integrate technology in their classrooms (Ertmer 1999). Researchers have noted that teachers do not use technology when they think it is not connected to the curriculum (Sarama et al. 1998). Questions about whether products were mapped to standards were included in the implementation data collection.

Student and Teacher Roles. One of the arguments made for using technology in classrooms is that it shifts student and teacher roles in ways that lead to more learning. For example, when students are working with software, every student can construct a response to a question, rather than just the student called on by the teacher. Studies of classroom interactions have found that in classes using technology, individual students make many more responses (Schofield and Verban 1988; Worthen et al. 1994), and the teacher shifts out of the role as lecturer and may do more coaching of individuals or small groups (Henriquez and Riconscente 1999; Swan and Mitrani 1993). Students also may be more likely to be on task (Waxman and Huang 1996).

Use of Student Performance Reports. The products provide individual student and classroom reports of performance. Many companies indicated that teachers should review these reports regularly to ensure that students are progressing and to make adjustments as needed. Research on formative assessment reviewed by Black and Wiliam (1998) suggests that teachers' use of information from mid-course assessments can increase learning gains.

Data Collection Methods

The study's data fall into two broad categories: (1) data related to product implementation and classroom characteristics, including product usage and teacher characteristics, and (2) data related to student characteristics, in particular achievement, but also average characteristics of students attending the study's schools. In the first category, the study observed classrooms, interviewed teachers, and administered a questionnaire to teachers. In the second category, the study administered achievement tests, collected records

for individual students, and merged data from public sources about student characteristics of schools.

1. Implementation and Classroom Data

Classroom Observations. A template of a classroom observation protocol was developed for use in observations of both treatment and control classrooms across all grade levels. The structure of the protocol drew on elements of existing observation instruments found to be reliable in the field (Good and Brophy 2003). The content of the protocol was based on research literature on factors associated with using technology (some of which is cited above), as well as on implementation models for each product, which the study worked with product vendors to develop.

The observation protocol had three sections—one each for student activities, technology use, and general observations. Observers collected data at five points during a 50-minute period. They completed a section on student activities every 10 minutes, recording the number of activities taking place, the type of activities, the teacher's role, the percent of students who were off-task, and the number of students using products. They completed a section on product use every 10 minutes, characterizing activities such as teachers' roles in motivating and helping students who were having technical difficulties with the products. They also recorded the total amount of time students used products (if used), the location of the observation, whether other products were used (products that were not the focus of the study), and how products may have been used in ways that differed from the product's implementation model.

Observers were experienced researchers, and a majority had experience conducting observations of classroom instruction. Prior to observations, observers were trained to use the protocol reliably by watching videotapes of classrooms using products in the study. Appendix A provides additional details on the development of the protocol and the training of observers. Observers conducted observations in October and November 2004, in December and January 2005, and in March and April 2005. They completed 98 percent of scheduled classroom observations.

Teacher Interviews. Teachers were interviewed after observations to gather data about product use in treatment classrooms and the use of software in control classrooms. The interview included questions about curriculum and instruction, access and frequency of use, whether students used products at home, and student-grouping strategies. It also included questions about how teachers managed instruction with the product, experiences with technical problems, and the types of information technology and vendor supports available and accessed. Interviewers (classroom observers) selected a response category that best matched the teacher response, or, if a response was difficult to code, they read the response options to teachers, who selected the most appropriate one. Appendix A provides additional details regarding the development of the interview and the training of field staff on its use. Observers interviewed all teachers after the first and third observations (and in some cases, before the observation, to fit with teacher availability). For the second observation, treatment teachers were interviewed on topics related to product use.

Teacher Survey. A questionnaire administered to all teachers complemented the observations and interviews. Based on questionnaires developed by the National Center for Education Statistics and other studies of educational technology, the questionnaire gathered data about teacher demographic and educational background, teaching credentials, experiences using technology, and instructional practices. Teachers also provided information on professional development activities and their first few months of experience using the products, including the availability of technical support and their assessments of the product. The study mailed the 20-minute questionnaire to teachers at their schools in late November and received responses from 96 percent of teachers.

2. Student Data

Student Achievement. For first, fourth, and sixth graders, the study administered the short version of the Stanford Achievement Test (version 10), commonly referred to as the SAT-10. For first and fourth graders, the study used the reading portion, and for sixth graders, the team used the math portion.[6] For first graders, the study supplemented the group-administered SAT with the Test of Word Reading Efficiency (TOWRE), a short one-on-one test considered to be a reliable and valid measure of early reading ability. For algebra, the study used Educational Testing Service's (ETS) End-of-Course Algebra Assessment, which essentially is a final exam. Because baseline assessments of algebra skills were not available, the study worked with ETS to split the test into two equivalent parts that have the same level of difficulty—one administered at the beginning of the course and the other at the end of the course.[7] Classrooms were randomly selected to take either part A in the fall and part B in the spring or to take part B in the fall and part A in the spring. Because most students were taking algebra in a full-year course, the test's timing corresponded roughly to the beginning and end of the school year. A few schools taught algebra in a one-semester course, in which case the test was administered at the beginning and end of the semester. Table I.4 provides information about aspects of the tests including their reliability and norming populations (the ETS test is unnormed but other tests have national norms).

The initial test was administered in fall 2004 at each grade level. Study staff typically administered tests in students' regular classrooms, mostly during October and November 2004 (some testing occurred in December). Tests were administered again beginning in late spring 2005, as close to the end of the school year as possible, to ensure maximum exposure to the products. Tests typically lasted the length of a class period, though in some schools the first grade test was administered in two days because of its length and the age of the students.

[6]For first graders in the fall, the test was the Stanford Early School Achievement Test (SESAT). Because the 10th version was not available at the time the study was purchasing tests in summer 2004, the 9th version was substituted.

[7]The SAT-10 ninth grade test was considered but ultimately not adopted because the ETS test focuses specifically on algebra, whereas the SAT-10 includes other math topic areas such as geometry and statistics. Algebra courses also include students at higher and lower grade levels, which would have required administering different versions of the SAT-10 to students in different grade levels. The study team assumed that skills in these other areas would not align with technology products that focused only on algebra.

Table I.4. Features of Tests Used in the Study.

	Stanford Achievement Test, Ninth Edition (SAT-9), Abbreviated Battery	Stanford Achievement Test, Tenth Edition (SAT-10), Abbreviated Battery	Test of Word Reading Efficiency (TOWRE)	Educational Testing Service End-of-Course Algebra Assessment (Customized)
General Information	Commercially available, used by large number of states and school districts.	Commercially available, used by large number of states and school districts.	Commercially available. Contains two subtests: The Sight Word Efficiency (SWE) subtest assesses the number of real printed words that can be accurately identified within 45 seconds, and the Phonetic Decoding Efficiency (PDE) subtest measures the number of pronounceable printed non-words that can be accurately decoded within 45 seconds.	Full form is commercially available. Test is based on algebra standards of the National Council of Teachers of Mathematics. For the study, ETS separated the test items into two balanced halves with equal levels of difficulty, such that one could be administered in the fall and the other in the spring.
Norm Sample	National norms, based on samples of 250,000 students in spring 1995 and 200,000 in the fall. The average student in the norm sample has a normal-curve-equivalent score of 50, and the standard deviation of normal-curve-equivalent scores is 21.06.	National norms, based on samples of 250,000 students in spring 2002 and samples of 100,000 in fall 2003. The average student in the norm sample has a normal-curve-equivalent score of 50, and the standard deviation of normal-curve-equivalent scores is 21.06.	National norms based on a sample of 1,507 students in 30 states in fall 1997 and spring 1998. The average student in the norm sample has a standard score of 100, and the standard deviation of standard scores is 15.	Not nationally normed.
Reliability and Validity	Internal consistency (KR-20) coefficients ranged from .80s to .90s for most tests and subtests. Evidence of content, criterion-related, and construct validity.	Internal consistency (KR-20) coefficients are .80s to .90s for full multiple-choice battery test and subtests. Evidence of content, criterion-related, and construct validity.	Average alternate forms reliability coefficients exceed .90. Test/retest coefficients range from .83 to .96.	In 2003, information from 20,506 test takers indicated a mean score of 23.3 questions correct out of 50, with reliability of 0.87 and a standard error of measurement of 3.1. The two forms from which the halves of the test used in the study were taken had similar reliability characteristics.

Source: Buros Institute of Mental Measurement: University of Nebraska-Lincoln, Lincoln, NE, 1998; Buros Test Reviews Online; Harcourt Educational Measurement, 2003; Torgesen et al. (1999) Test of Word Reading Efficiency (TOWRE), Examiner's Manual, Pro Ed, Austin, TX. Information about the ETS algebra test was provided in the administrators' manual provided by ETS, and general information about the test can be found at www.ets.org.

Overall, 95 percent of students on fall class rosters completed the fall test, 88 percent of students on spring class rosters completed the spring test, and 85 percent of students who completed the fall test also completed the spring test. Response rates depended on the grade level, with the highest testing rates for first grade (above 90 percent) and the lowest testing rates for algebra (about 80 percent). Response rate differentials for treatment and control classrooms were negligible, usually one or two percentage points. Appendix Table A.6 provides details about student sample sizes and response rates.

Technology's effect on student achievement is reported in "effect size" units. An effect size was calculated as the difference in average test scores between the treatment and control groups, divided by the standard deviation of the control group's test score. This metric allows the study to compare results for tests whose scores are reported in different units. For example, an effect size of 1.0 (or one standard deviation) is equivalent to 21.06 normal curve equivalents (NCEs) on the SAT-10 test and 15 points on the TOWRE. An effect size of 1.0 is equivalent to increasing a student at the 50th percentile to the 84th percentile.

School Records. Student data were gathered from school records starting at the end of the spring semester and continuing through the summer and early fall. Data items included age and race; standardized test scores for the previous and current academic years; and whether students had an individualized education plan, were limited English proficient, or were eligible for the free lunch program. School records data were reasonably complete for student age and gender. However, districts typically did not provide all items, and gaps occurred for test scores and for participation in the free lunch, special education, and English language learning programs in particular. Often, if data items were missing, they were missing for all students in a school. Because the study's estimation approach (discussed in Appendix B) relied on at least some student data being available for each classroom, the gaps in records data led to the decision to use aggregate school-level data for items that were available in the *Common Core of Data* (CCD).[8]

F. Looking Ahead

The report's chapter structure follows the study's grade-level structure. To avoid repeating the same information about the study's methods and approach, details on these are presented in the next chapter on the first grade study, and later chapters focus mostly on the findings. Readers are encouraged to read the second chapter to familiarize themselves with details about the methods.

Appendixes A and B report on the study's data collection approach and its response rates for the various types of data it collected, methods it used to estimate effectiveness, and findings from analyses using alternative estimators and other tests. Appendix B also presents

[8]The *Common Core of Data* (CCD) annually collects fiscal and non-fiscal data about all public schools, public school districts, and state education agencies in the United States. Information about it can be found at http://nces.ed.gov/ccd/. The information used in the study was reported by the CCD for the 2003-2004 school year.

the full set of estimates from the main product-effectiveness models, for readers who want to examine the detailed results.

The report based on a second year of data will focus on whether an additional year of experience using reading and mathematics software products is associated with greater effectiveness. Findings will also be reported for individual products.

Chapter II

Effects of First Grade Reading Software Products

This chapter presents findings on the implementation of first grade reading software products, effects on student test scores, and the relationships between effects and conditions and practices related to product use. The first grade study included five reading software products that were implemented in 11 districts and 43 schools. The sample included 158 teachers and 2,619 students. The five products were Destination Reading (published by Riverdeep), the Waterford Early Reading Program (published by Pearson Digital Learning), Headsprout (published by Headsprout), Plato Focus (published by Plato), and the Academy of Reading (published by Autoskill).

Effects were estimated using a hierarchical linear model (HLM), in which students were nested in classrooms and schools. Because of the study's experimental design, test score differences between classrooms using products and classrooms not using products can be attributed to the products, after allowing for sampling variability. The analysis relied on tests administered to students in fall and spring, as well as on data from student records, teacher questionnaires, and the *Common Core of Data* (for school characteristics).

The hierarchical linear model also was used to examine classroom and school characteristics correlated with effects. The study's design does not support causal statements that these characteristics *determine* effects. Districts and schools were not randomly assigned to use particular products. Instead, they selected products they believed were suitable for their needs. The uncontrolled selection could account for the measured correlations. However, examining the correlations is useful as an indication of whether effectiveness might vary with conditions and practices even if the particular mechanisms producing the variation cannot be established.

A. Implementation Analysis

The main questions addressed in the implementation analysis were (1) What do products do? (2) How much were products used and how did teachers use them? and (3) How did products affect classroom activities or roles? To answer these questions, the study collected data in six areas: features of the products in the study; teacher training and

support on using products; the duration, extent, and location of product use; technical difficulties in using products; role of products in the curriculum; and the effects that product use had on classroom activities. The first three areas relate to products as computer-based tools, and the fourth and fifth relate to products as instructional supports. The sixth area relates to how products may have changed what teachers and students were doing in classrooms and how they were interacting.

Product Features

Understanding the characteristics of the reading software products is useful as a context for the findings. The five reading products in the first grade study all provide tutorials and practice in early reading skills. Skills covered by the software include letter recognition, phonemic awareness, word-recognition and word-attack skills, vocabulary building, and text comprehension. Most of the products provide students with opportunities to read different kinds of text, including informational text, fiction, and poetry. The study estimated the average licensing fees for the products to be about $100 a student for the school year, with a range of $53 to $124.

Study staff assessed product features related to five broad areas of instructional design that are commonly used in the field: (1) tutorial opportunities, (2) practice opportunities, (3) individualization, (4) feedback to teachers, and (5) feedback to students. These categories also are consistent with vendor statements about their products, such as that the product supports individualization of learning or provides students with practice opportunities to supplement instruction. Product features were assessed based on a coding guide developed by the study team. Two coders independently reviewed products, and categories for which inter-coder agreement was 80 percent or greater were retained. For ratings on which initial coders disagreed, a third coder reviewed the product and made a determination. Only computer-based instructional components were assessed. Some products included text components that were not assessed.

Within the categories of the five main features, the study team assessed more detailed aspects of product design. For individualization and for student feedback, products were assessed for their features within three modes: *tutorial*, *practice*, or *assessment*. In tutorial mode, students are being instructed on a skill or concept that is defined, explained, or demonstrated by the product. In practice mode, students answer questions or solve problems in a skill area for which they received instruction (either off or on the computer). In assessment mode, students answer questions or solve problems on instruction they have received, and products tally responses to assess performance or mastery of the content or skill. Further, within individualization, products could sequence modules automatically or allow teachers or students to select modules to create a "learning path" through the product for each of the three modes. The three types of individualization and the three modes interact to create nine possible categories. Similarly, for the "opportunities" feature, products could provide different levels of opportunities in two modes, tutorial and practice, leading to two categories.

For teacher feedback, the team assessed whether the feedback was about student mastery, student learning paths, or classroom performance. For student feedback, the team

assessed whether feedback was "immediate," "mastery," and "diagnostic" and whether it occurred in practice and assessment modes. Feedback often is cited as an important aspect of instruction (Bangert-Drowns et al. 1991; Black and Wiliam 1998; Butler 1987; Crooks 1988, Kluger and deNisi 1996), and learning theorists have stressed the importance of diagnostic feedback that goes beyond whether students got answers right (Bransford et al. 1999).

Table II.1 summarizes assessments of the instructional features of the five products. All provided for individualized instruction (the capability to set individual student learning paths depending on a student's skill level), had numerous tutorial components and practice opportunities, and provided feedback to students during practice. Some products also provided feedback during assessment activities. All provided reports to teachers for the overall class and for each student, separately for the different product modules. One product also provided reports of student responses to individual questions. Three products provided recommendations about modules to which individual students should progress based on their performance.

Exhibit II.1 How Product Features Were Assessed.

Product features were assessed based on a coding guide developed by the study team. Two coders independently reviewed products, and categories were included if inter-coder agreement was 80 percent or greater. For ratings on which initial coders disagreed, a third coder reviewed the product and made a determination.

Inter-coder reliability was reached for five instructional features:

1. Tutorial Opportunities. Activities in which students receive instruction in a skill or concept.
2. Practice Opportunities. Activities in which students practice answering questions or solving problems in a skill area for which they had received some instruction (either on or off the computer).
3. Individualization. The degree to which the product allows for individual learning paths to be specified automatically or by the teacher or student.
4. Feedback to Teachers. Information teachers receive from the product about student mastery of concepts or skills.
5. Feedback to Students. Information students receive from the product about their level of understanding or mastery of a concept or skill during tutorials, practice, or assessments.

Only computer-based instructional components were assessed; some products had non-computer-based components that were not assessed.

Table II.1. Instructional Features of First Grade Reading Products.

Product	1. Tutorial Opportunities	2. Practice Opportunities	3. Individualization[a]									4. Feedback to Teachers			5. Feedback to Students[b]					
			Automatic			Teacher Input			Student Input						Immediate		Mastery		Diagnostic	
			T	P	A	T	P	A	T	P	A	Student Mastery	Learning Paths	Classroom Performance	P	A	P	A	P	A
A	Many	Many	✓	✓	✓	✓	✓	✓				✓	✓	✓	✓			✓		
B	Many	Many	✓	✓		✓	✓	✓				✓	✓	✓	✓				✓	
C	Many	Many	✓	✓		✓	✓	✓	✓	✓		✓		✓	✓				✓	
D	Many	Many	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓		✓	✓	✓	
E	Many	Many	✓	✓		✓	✓	✓				✓		✓	✓	✓		✓	✓	✓

Source: Staff review.

[a]T= Tutorial mode, P = Practice mode, A = Assessment mode.

[b]Immediate feedback: Learner told whether response is correct immediately after completing module; Diagnostic feedback: Learner receives hints or other information concerning the probable source of error; Mastery feedback: Learner informed of the number correct and whether a skill or concept has been acquired ("mastered") after completing a sequence of items. Learner can return to missed items.

Teacher Training and Support

Product vendors trained teachers who would be implementing the products on how to use them. Training generally took place in the host districts (and sometimes the host schools) during summer or early fall of 2004. Training topics included classroom management, curriculum, and standards alignment, and generally teachers had opportunities to practice using the products. Nearly all teachers (94 percent) attended the initial training, according to attendance logs.

On average, vendors provided about one day of training (7.5 hours), varying from 2 to 18 hours across products. Vendors also provided support during the school year. Modes for ongoing support included e-mail or telephone help desks (69 percent of teachers reported receiving this kind of help), product representatives visiting teachers (55 percent), and additional training at schools (39 percent). The need for such additional support is suggested by the finding that by the time of the first classroom observation (generally about mid-fall), when most teachers had begun to use products, the proportion of teachers indicating that the initial training had adequately prepared them had declined from 95 percent at the end of the initial training to 60 percent.

In addition to ensuring that teachers received training, the study team worked with districts to identify hardware and software needs, such as computers, headphones, memory, and operating system upgrades, and the study purchased the upgrades as needed. Common upgrades included desktop and laptop computers, servers, memory, and headphones. The study did not upgrade networking infrastructures, though some purchases of servers enabled products to operate more smoothly on local networks. As noted in the previous chapter, providing hardware and software upgrades may have contributed to higher levels of measured effectiveness, if districts normally would not have been able to purchase the upgrades. The study did not provide software or hardware support for control group teachers.

Study observers noted that first grade treatment teachers averaged about six computers per classroom (classes had average attendance of 18 students). When products were used in school labs, there was typically (but not always) a computer for every student. With the additional technology provided by the study, in principle the number of operable computers appeared adequate for students to reach recommended levels of product use.

Duration and Extent of Product Use

All treatment teachers used the assigned product to some degree, and some products were used heavily during the school year. Researchers asked teachers: "How often (how many sessions per week) and for how many minutes per week does a typical student use the product?" Table II.2 shows the hours that students were exposed to study products (in treatment classrooms) or to the most frequently used other technology-based reading product (in both treatment and control classrooms), based on teacher reports.

Table II.2. Teacher-Reported Use of Study Products and Other Reading Software Products, First Grade

	Study Products			Other Reading Products		
	Treatment Group	Control Group	*p*-value[a]	Treatment Group	Control Group	*p*-value[a]
Percent of teachers using a product	100	1	.00	55	75	.01
Minutes of weekly use	94	(*)	n.a.	18	25	.08
Hours of annual use	48	(*)	n.a.	4	10	.00
Sample size	**89**	**69**		**89**	**69**	

Sources: Classroom observations, teacher interviews.

[a]Tests of difference were conducted using two-level hierarchical models (with a teacher level and a school level), with the treatment effect at level one and between-school variance at level two. The *p*-value is the smallest level of significance at which the null hypothesis that the difference is zero can be rejected.

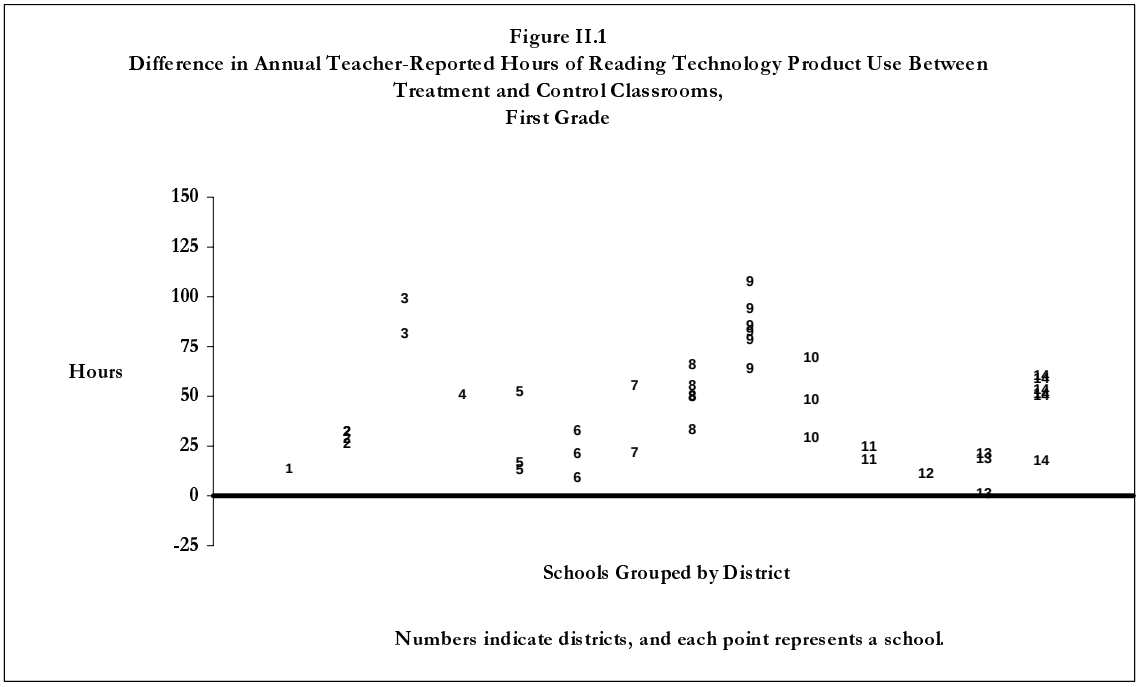
(*) Sample is too small to compute means.

n.a. = not applicable.

Control and treatment teachers could use products other than those in the study, and Table II.2 shows the extent to which they did. About 75 percent of control teachers reported using some other product (possibly more than one), and average use for the control group as a whole (including teachers who did not use products) was 10 hours a year. The most commonly used other products were reading practice or assessment products (one was a product available on the Web at no cost). In contrast, 55 percent of treatment teachers reported using a product other than the one in the study, and average use for all treatment teachers was about 4 hours a year. The sum of use of both types of products (products in the study and products not in the study) was 52 hours a year for treatment teachers versus 10 hours a year for control teachers.

The difference in product use between treatment and control teachers varied by district and school (Figure II.1). Some schools had differences of 20 hours or fewer, whereas others had differences of 100 hours or more. No schools had negative differences, which would arise if control teachers used other products more than treatment teachers used the study product and other products, though one school had a difference that was close to zero. Usage differences also are evident between districts. For example, the difference between treatment and control teacher technology usage was relatively large in district 9 compared to districts 5 and 6. Correlations between usage and product effects will be examined later in the chapter.

Data from product records on student usage also provided information about product use. Four of the five first grade products included databases that tracked the time when students were logged on (some products tracked usage more closely, gathering data on every response and keystroke). Analysis of these data indicated that when students used products, their usage averaged about 23 minutes a day and 29 hours a year (Table II.3). Usage varied



Source: Teacher Inteviews.

Table II.3. Daily and Annual Usage From Product Records.

	Product A	Product B	Product C	Product D	Overall
Total days used during school year	20	122	45	64	76
Minutes of daily use (when used)	34	30	11	14	23
Hours of annual use	11	61	8	15	29

Source: Product data on usage. One product did not collect information on student usage. Overall usage is product usage weighted by student sample sizes.

across products from 11 minutes a day to 34 minutes, and annual usage varied from 8 hours to 61 hours. For a typical 180-day school year, average daily usage is about 10 minutes for all products combined.[9] Assuming that reading is taught in a 90-minute block, products were used for about 11 percent of total reading instruction time.

Unlike product reports of usage, teacher reports of usage are available for all products, and completeness is a useful feature for later analyses. Teacher reports about usage are more inclusive than product usage, which is actual student logged-in time. They can differ because of student absenteeism, computer or network problems, and students working together

[9]Multiplying total days of usage (76) by minutes used (23) yields 1,748 minutes of annual use, which is divided by 180 to arrive at average use of 9.7 minutes per actual school day.

under a single login. However, correlations between the two measures were high (correlation coefficients were 0.80 and 0.81 for weekly and annual usage), and the high correlations provide a basis for using teacher reports of usage in the effectiveness analysis later in the chapter.

The levels of usage found here relate to teachers using products for the first time and within a national study. Both first-time usage and the presence of the study could have affected usage compared to "typical" levels of usage. Whether teachers use products more as they are more experienced is being examined in the second year of the study. How the study itself affected usage is not known. The study may have increased usage because it purchased hardware and software upgrades without needing to go through district procurement processes. Teachers received honoraria for attending training on how to use products (and nearly all teachers attended training). Also, the study team provided feedback to developers about any problems or issues encountered with the product during classroom visits, if teachers asked them do so. On the other hand, districts did not invest their own resources in purchasing products and training teachers, which could have reduced usage because districts did not have a stake in the products.

Technical Difficulties

Study observers noted technical problems during classroom observations in about 20 percent of the time segments they observed (an observation consisted of four to five time segments within a class period). Most problems lasted only a short time and affected only a few students. Common problems involved logging on to products, computers freezing and needing to be rebooted, and trouble with peripheral hardware such as headphones.

Satisfaction with Products

Nearly all teachers (92 percent) said they would use the product in the next school year if given the choice. When asked what they would change, 31 percent said they would focus on classroom management issues related to product use, 15 percent said they would not change anything, 11 percent said they would start using products earlier in the school year, and 11 percent said they would monitor student progress more closely.

Role of Products and Time and Place of Use

Reading software products generally functioned as supplements to the reading curriculum. Four products were expressly designed as supplements, and 96 percent of treatment teachers indicated they used products as supplements. One product could be used as either a supplement to a reading program or as a core reading program, though all treatment teachers used it as a supplement.

Teachers varied in how they scheduled product use. Just over a third (34 percent) scheduled product use exclusively during regular class time; 19 percent scheduled use during "other" times (such as before school, lunchtime, or time usually set aside for science or social science); and the largest group—47 percent—scheduled use both during class time

and at other times. Most teachers (80 percent) used products in regular classrooms, and 20 percent of teachers used products in a computer lab.

Vendors stressed the importance of teachers being present to offer assistance or keep students on task, rather than relying on lab monitors or paraprofessionals. Nearly all first grade treatment teachers (96 percent) reported being present when their students used products. Vendors also stressed the importance of reviewing product reports on student performance; 79 percent of teachers reported reviewing reports two to three times a month or more frequently, and 59 percent reported reviewing performance reports once a week or more.

Impact on Classroom Activities

Previous studies have reported that technology can change the role of teachers (Honey and Henriquez 1996; Linn and Hsi 2000; Sandholtz et al. 1997). This role change is sometimes viewed as the desired outcome, with teachers encouraged not to spend time dispensing information ("the sage on the stage") but instead to act as a facilitator ("the guide on the side"), helping students or small groups who are engaged in their own learning tasks. Because the same observation protocol was used for treatment and control classrooms, the study could contrast classroom activities to assess whether products affected the activities.[10] During observation intervals, teachers were classified as leaders, facilitators, monitors (scanning the class to detect any behavioral issues), or otherwise engaged. Table II.4 shows that during observation intervals (the time periods during which observers noted teacher roles), 56 percent of observations in treatment classrooms coded the teacher's role as "facilitating" compared to 32 percent of control classroom observations ($p < .01$).

Products also changed the typical activity in which students were engaged. Students were more likely to be engaged in individual work during treatment classroom observations than during control classroom observations ($p < .05$). Students in control classes were more likely to be listening to a teacher or participating in a question and answer session directed by the teacher.

[10]The observation protocol called for observers to observe periods during which treatment teachers were using reading products. To the extent possible, control classrooms were observed during the same time period that treatment teachers were using products. For example, if treatment teachers in a particular school used the product during the last half of the reading period, observers attempted to observe control classrooms during the last half of the reading period. If observers had observed classrooms at random times rather than when products were used, the differences shown in the table may have been smaller because products would not be in use during some of the observations.

Previous studies have observed that students may be more academically focused when working on computers (Joyner 2002; Swan et al. 2005). Table II.4 shows that the proportion of treatment and control classroom observations in which more than 90 percent of students were on task was high overall and about the same (85 percent of treatment classrooms and 86 percent of control classrooms).[11]

Table II.4. Activities in Treatment and Control Classrooms, First Grade (Percentage of Observation Intervals).

	Treatment Classrooms	Control Classrooms	*p*-value[a]
Teacher Role (Percent)[b]			
Leader	29	58	.00
Facilitator	56	33	.00
Monitor/observer	17	12	.05
Working on other tasks	6	4	.10
Other	8	5	>.50
Instructional Activity (Percent)[b]			
Individual work	84	40	.00
Lecture	16	26	.01
Question and answer	16	33	.00
Review of student work	2	<1	.05
Other	10	10	>.50
Student On-Task Behavior (Percent)			
Percent of time intervals with more than 90 percent of students on task	85	85	>.50
Sample Size			
Number of classrooms	89	69	
Number of observations	612	795	

Source: Classroom observations at minutes 10, 20, and 30 for each class.

[a]Tests of difference were conducted using two-level hierarchical models (with a teacher level and a school level), with the treatment effect at level one and between-school variance at level two. The *p*-value is the smallest level of significance at which the null hypothesis that the difference is zero can be rejected. If the *p*-value is less than .01, the difference is significant at the 1 percent level. If the *p*-value is less than .05, the difference is significant at the 5 percent level, and so on.

[b]Observers coded "all that apply." Numbers can add to more than 100 percent.

Another aspect of instruction is the amount of time allotted for reading instruction, which is within the control of teachers to some degree and could be influenced by the use of study products. The study asked treatment and control teachers to estimate how much time

[11]Observers estimated the proportion of students who were doing something other than the assigned academic task during a one-minute segment using categories such as walking around the classroom for reasons unrelated to the task, talking with other students on topics unrelated to the task, talking with other students while the teacher was addressing the whole class, sitting at the computer for long periods with no interactions with keyboard or mouse, or sleeping or having their heads on their desks.

they spent on reading instruction. Treatment teachers reported spending an average of 8.7 hours a week compared to 7.9 hours on average for control teachers. The difference of 0.8 hours was not statistically significant ($p = 0.13$).

B. Effects on Reading Test Scores

A central question for the study is whether products resulted in higher test scores. Effects on test scores were estimated using a model that accounted for the nesting of students in classrooms and classrooms in schools. The estimates show that effects on test scores generally were not statistically different from zero and that most teacher and school characteristics were uncorrelated with effects. Different estimation approaches yielded similar estimates. In two districts for which district test scores were available, use of district test scores rather than the study-administered test as the measure of achievement yielded similar results. Appendix B provides details of the robustness analysis.

Characteristics of Treatment and Control Teachers and Students in Treatment and Control Classrooms

The study randomly assigned teachers to treatment and control groups. Whether random assignment achieved its objective can be assessed by examining baseline characteristics of teachers in the two groups. Table II.5 shows teacher characteristics and *p*-values of tests of equivalence. After accounting for multiple comparisons, none of the teacher differences is statistically significant. However, all the characteristics are entered into models to adjust for remaining differences.

The study did not randomly assign students to teachers, but Table II.5 shows that students in treatment and control classrooms were similar on fall test scores, age, and gender. After accounting for multiple comparisons, none of the differences is significant.[12] Standard scores for the TOWRE test, which averaged about 109, place students in about the 67th percentile nationally, higher than their performance on the SESAT, where the average students were at about the 50th percentile. (The simple correlation between scores was about 0.70.) These differences may arise because of differences in what is being tested, differences in how the test is conducted (the TOWRE is one on one and the SESAT is administered in groups), or differences in the norm samples on which the percentiles are based.[13]

[12]The fall test was administered after the school year was underway and, in principle, scores could have been affected by the use of products. The fact that the average treatment group score was not statistically different from the average control group score suggests that the use of products did not have much, if any, effect on fall scores.

[13]A recent study that examined interventions for struggling readers administered a variety of reading tests and also found that student percentile rankings varied depending on the test. See http://www.mathematica-mpr.com/publications/PDFs/CTRGexec.pdf.

Table II.5. Characteristics of Teachers and Students in Treatment and Control Classrooms, First Grade.

	Treatment Classrooms	Control Classrooms	*p*-value of the Difference[a]
Teacher Characteristics			
Teaching experience (years)	11.7	11.6	0.97
Has a master's degree (percent)	38	55	0.03
School has computer specialist (percent)	74	78	0.55
Received professional development on using technology last year (percent)	53	55	0.79
Female (percent)	99	97	0.42
Teacher Sample Size	**89**	**69**	
Student Characteristics			
Female (percent)	49	49	0.93
Age as of October 2004 (years)	6.63	6.67	0.03
Unadjusted score on fall Test of Word Reading Efficiency (standard score)	109.1	109.8	0.43
Unadjusted score on fall SAT-9 Reading Test (NCE)	50.1	50.9	0.62
Student Sample Size	**1,516**	**1,103**	

Sources: Teacher questionnaire, student records, and tests administered by study staff.

Note: Multiple-comparisons testing used the Benjamini-Hochberg procedure, with separate tests for five teacher characteristics and four student characteristics.

[a]Tests of treatment and control differences were conducted using a two-level hierarchical model with classroom treatment status as a fixed effect and school as a random effect (for teachers) and classrooms and schools as random effects (for students). The *p*-value of the difference shown in the table is the *p*-value of the estimated treatment coefficient.

Effects Were Not Statistically Different from Zero

Effects were estimated using a three-level hierarchical linear model. The main outcome at the first level is the spring test score, which is related to student characteristics (age, gender, and the fall score). The second level is a model of a classroom's average test score as a function of classroom characteristics (most importantly, a treatment indicator of whether a classroom was randomly assigned to use a product). The third level is a model of the school's average test score as a function of school characteristics such as proportion of students receiving free or reduced-price lunch, race and ethnic composition, and the proportion of students receiving special education services. These school characteristics also are student characteristics, but nearly complete data from the *Common Core of Data* were available for these items, whereas the study's efforts to collect the same items from school records for individual students resulted in gaps and missing data.

Table II.6 shows score differences for the overall SAT-9 reading test and its three subtests and for the overall Test of Word Reading Efficiency and its two subtests. [14] None of the differences is statistically different from zero. Effect sizes are in the range of -0.01 to 0.06.[15]

Table II.6. Spring Reading Test Score Differences in Treatment and Control Classrooms, First Grade.

	Treatment Classroom Average Score	Control Classroom Average Score	Difference	Effect Size	*p*-value
Stanford Achievement Test Ninth Edition (NCE)					
Overall score	50.20	49.47	0.73	0.03	0.34
Subtest scores					
Sounds and letters	50.67	49.47	1.20	0.06	0.16
Word reading	50.96	50.08	0.88	0.04	0.22
Sentence reading	50.19	50.34	-0.15	-0.01	0.89
Test of Word Reading Efficiency (Standard Score)					
Overall score	112.34	111.83	0.51	0.04	0.57
Subtest scores					
Phonemic decoding efficiency	109.81	109.53	0.28	0.03	0.53
Sight word efficiency	110.42	110.18	0.24	0.02	0.64

Note: See Appendix B for details of the estimation model. Variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity; percent of students in special education; percent eligible for free lunch; and student, classroom, and school random effects.

The treatment classroom average score reported in the table is the control classroom average score plus the treatment effect. It differs from the unadjusted treatment classroom score.

The study examined whether products reduced the proportion of students who were low scorers on the SAT-9 by creating an indicator of whether a student fell in the lower third of scores and estimating a three-level model with that indicator as the outcome. Whether the student was below the 33rd percentile on the pretest was used as a covariate along with the

[14]The main model estimated a single effect for all five products by including an indicator that a student was in a treatment classroom, regardless of the product used in the classroom. This approach essentially averages individual product effects with weights that are proportional to the number of classrooms that products have in the study. Products with more classrooms contribute more to the estimated effect.

[15]Effect sizes are calculated by dividing the score difference shown in the table by the standard deviation of the distribution of spring test scores for the full control group. See Appendix Table A.8(a) for standard deviations.

same set of variables used for the score model above. The results indicated that products did not have a statistically significant effect on whether students were low scorers (Table II.7).[16]

Table II.7. Effect on Percent of Students in Lowest Third of Reading Test Score.

	Treatment Classroom Percentage	Control Classroom Percentage	Difference	Effect Size	*p*-value
Percent of students below 33rd percentile of spring reading test	33.3	34.1	-0.8	-0.02	0.84

Note: Other variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity; percent of students in special education; percent eligible for free lunch; and student, classroom, and school random effects.

The treatment classroom percentage reported in the table is the control classroom percentage plus the treatment effect. It differs from the unadjusted treatment percentage. The effect size is calculated using the Cox Index (the log odds ratio divided by 1.65).

Figures II.2 and II.3 depict the variation of school effect sizes for districts and products.[17] For example, in district 9, effect sizes ranged from almost -0.50 to nearly 0.75 across the six schools in the district. An analysis of variance indicates that 85 percent of the variability of the effect size is within districts and 15 percent is between districts.

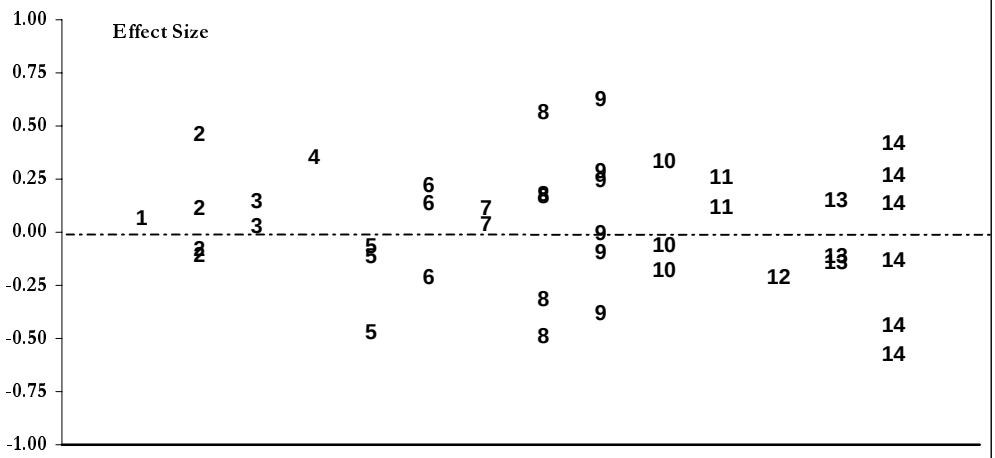
Classroom and School Moderators

Classroom and school characteristics may help explain the variability of effects observed in Figure II.2. The moderating effect of these characteristics was investigated by interacting the treatment indicator with classroom characteristics at the second level and adding an equation to the model's third level that specified the treatment effect as a function of school characteristics. The amount of time teachers reported using study products (shown in Table II.2) also was included in the model, as was the amount of time teachers reported using other products. Other implementation factors included in the model were (1) whether a teacher had adequate time to prepare for product use, (2) whether the teacher indicated

[16]Two-level models also were estimated for students in each quartile (based on the fall score), to assess product effects across the achievement distribution. Estimated effect and *p*-values for the four quartiles were 0.35 (0.82), 0.82 (0.53), 1.77 (0.19), and 0.25 (0.86).

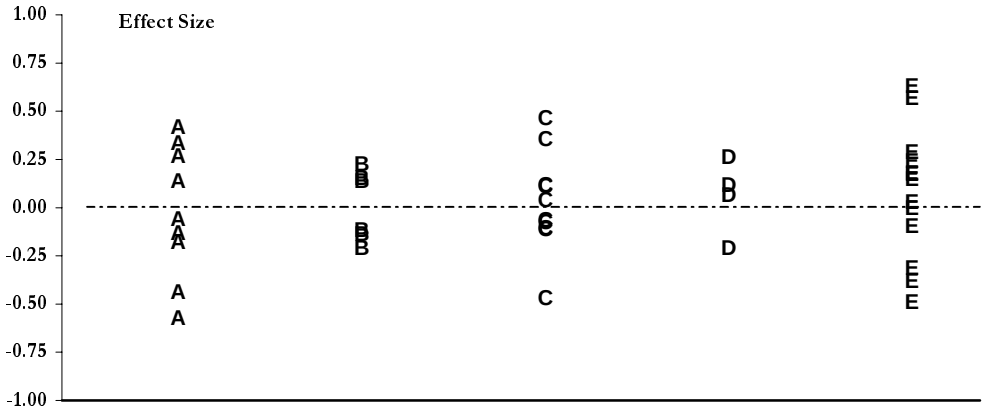
[17]School effects were estimated using a regression model in which test scores are regressed on student and teacher characteristics and the treatment indicator is interacted with an indicator for each school. Standard errors were adjusted for classroom clustering. Effect sizes for schools are calculated by dividing the score difference for each school by the standard deviation of the distribution of control group test scores. An alternate effect size was calculated by dividing the score difference for each school by the standard deviation of the distribution of control group test scores for students in that school. The alternate effect sizes show the same pattern between schools, but are more variable.

Figure II.2
School-Level Effect Sizes by District
First Grade (SAT-9 Reading Score)



Numbers indicate districts, and each point represents a school.

Figure II.3
School-Level Effect Sizes by Product
First Grade (SAT-9 Reading Score)



Letters indicate products, and each point represents a school.

Note: Statistical significance of average effect sizes cannot be inferred from the figure because student and teacher sample sizes differ between schools.

that students had problems accessing the products, and (3) whether the teacher indicated that the school has a computer specialist. The three measures were based on teacher reports.

Table II.8 shows estimates of the relationships of classroom and school characteristics with product effects for the overall SAT-9 score and three subtest scores; Table II.9 shows estimates for the Test of Word Reading Efficiency. The tables present only the moderator estimates and not estimates for all variables in the model. Positive coefficients indicate that a characteristic is associated with larger test score differences between treatment and control classrooms.

Statistical tests indicate that the full set of classroom and school characteristics was not correlated with product effects for the overall SAT-9 score. Results of separate tests of classroom and school characteristics as distinct subsets are also shown and indicate similar results. After adjusting for multiple comparisons, only the student-teacher ratio is statistically significant. Time using study products was not related to effects.

For SAT-9 subtests, statistical tests indicate that school and classroom characteristics were correlated with effects for the sounds and letters and word reading subtests. For the word reading subtest, the student-teacher ratio and the percent of students in special education were individually significant after adjusting for multiple comparisons. For the Test of Word Reading Efficiency, no classroom or school characteristics were correlated with effects.

Because the research design did not randomly assign products to schools with different characteristics, factors that led schools to choose particular products and to have particular experiences with the products could explain the observed correlations. For example, effects were larger when classes had fewer students (smaller student-teacher ratios), but schools with smaller classes may have other characteristics, such as their location, household income, or access to technology in homes—characteristics not measured in the study that may be related to product effectiveness. The inability to rule out these alternative explanations for the observed correlations is a reason to be cautious in interpreting them.

C. Conclusions

This study experimentally tested the effects of selected reading technology products in first grade classrooms that volunteered to participate. Schools participating in the study were more likely than average to be in urban areas and had higher-than-average levels of poverty and students who were Black or Hispanic. Key findings are the following:

1. **Teachers Were Trained on the Products and Used Them**. Nearly all treatment teachers (94 percent) received training from the vendor on how to use the product, and all of them implemented the product to some degree. Products were used for about 11 percent of total reading instruction time.

2. **Effects on Test Scores Were Not Statistically Different from Zero.** Overall reading scores for students in treatment and control classrooms were 50.2 and 49.5, respectively (in normal-curve-equivalent units). The difference was not statistically different from zero.

3. **Most School and Classroom Characteristics Were Uncorrelated With Effects.** Classroom and school characteristics were not correlated with product effects for the overall SAT-9 score. The one exception was the student-teacher ratio. For the sounds and letters and word reading subtests, classroom and school characteristics were correlated with product effects. (The student-teacher ratio and the percent of special education students were individually significant factors.) Time of study product usage did not have a statistically significant correlation with effects for the overall score or subtest scores. No school or classroom characteristics were correlated with effects for the Test of Word Reading Efficiency.

Table II.8. Interactions Between Moderating Variables and Effects: SAT-9 Reading Test, First Grade.

	Total Score (NCE)				Sounds and Letters Subtest Score (NCE)				Word Reading Subtest Score (NCE)				Sentence Reading Subtest Score (NCE)			
	coefficient	standard error	*p*-value		coefficient	standard error	*p*-value		coefficient	standard error	*p*-value		coefficient	standard error	*p*-value	
Classroom Interaction Variables																
Years of teaching experience	0.15	0.08	0.07		0.22	0.09	0.02		0.10	0.08	0.21		0.10	0.09	0.31	
Teacher is female	-1.56	2.66	0.56		-0.57	4.19	0.89		-7.12	2.67	0.01		2.11	3.28	0.52	
Teacher has master's degree	-0.52	1.52	0.73		-1.53	2.29	0.51		1.80	1.39	0.20		-0.63	1.60	0.69	
Annual hours of study product use (in hundreds)	2.25	2.15	0.30		1.68	2.70	0.54		3.72	2.23	0.10		-0.70	1.92	0.72	
Students had problems getting access to product	-1.01	1.29	0.44		-2.51	1.79	0.16		-0.89	1.29	0.49		-0.38	1.22	0.76	
Teacher had adequate time to prepare to use product	2.20	0.92	0.02		3.29	1.29	0.01		1.39	0.96	0.15		0.87	0.86	0.32	
Teacher indicates school has computer specialist	-0.19	1.22	0.88		-2.34	1.72	0.18		2.20	1.16	0.06		0.06	1.15	0.96	
Product is used in classroom	3.08	2.19	0.16		1.19	2.08	0.57		3.47	2.12	0.10		3.28	2.13	0.13	
Teacher participated last year in technology professional development	-1.23	0.92	0.19		-1.25	1.26	0.33		-1.39	1.11	0.22		0.01	0.87	0.99	
School Interaction Variables																
Percent eligible for free lunch	-6.21	4.19	0.15		-6.14	5.68	0.29		-5.67	4.34	0.20		-4.22	4.62	0.37	
Student-teacher ratio	-1.03	0.32	0.00	*	-0.98	0.40	0.02		-1.02	0.31	0.00	*	-0.74	0.42	0.09	
School is in urban area	-0.86	1.43	0.55		-0.69	1.70	0.69		-1.18	1.52	0.44		-1.64	1.70	0.34	
Percent African American students	0.88	3.63	0.81		1.03	5.28	0.85		2.17	3.52	0.54		-0.95	3.64	0.80	
Percent Hispanic students	8.75	4.42	0.06		9.50	5.61	0.10		5.90	4.11	0.16		5.20	5.35	0.34	
Percent special education students	5.21	4.41	0.25		3.45	5.22	0.51		9.11	3.08	0.01	*	0.04	6.19	0.99	
		χ^2	*p*-value			χ^2	*p*-value			χ^2	*p*-value			χ^2	*p*-value	
Chi-Squared Tests																
Classroom and school variables		23.61	0.07			27.03	0.03	‡		28.30	0.02	‡		12.78	*p*>.50	
Classroom variables		14.30	0.11			18.86	0.03	‡		20.32	0.02	‡		6.62	*p*>.50	
School variables		10.50	0.10			6.69	0.35			11.07	0.09			5.72	*p*>.50	

Source: Author calculations. Other variables in the model include student pretest score, age, and gender; teacher experience, education, and gender; and school race-ethnic composition, percent free lunch, urban area, percent special education, and time using other technology products.

*Significantly different from zero based on the Benjamini-Hochberg correction for multiple comparisons, with false discovery rate of 0.05 (see Benjamini and Hochberg 1995). The multiple comparisons correction is computed separately for the overall test score and each subtest score and for the classroom and school domains.

‡Statistically significant at the 0.05 level of significance.

Table II.9. Interactions Between Moderating Variables and Effects: Test of Word Reading Efficiency, First Grade.

	Total Score (Standard Score)			Phonemic Decoding Efficiency (Standard Score)			Sight Word Efficiency (Standard Score)		
	coefficient	standard error	*p*-value	coefficient	standard error	*p*-value	coefficient	standard error	*p*-value
Classroom Interaction Variables									
Years of teaching experience	0.04	0.05	0.47	0.04	0.05	0.44	0.02	0.05	0.62
Teacher is female	-2.56	3.60	0.48	-2.97	3.04	0.33	-1.51	3.30	0.65
Teacher has master's degree	-0.27	0.77	0.72	-0.15	0.86	0.86	-0.29	0.49	0.63
Annual hours of study product use (in hundreds)	-0.39	1.31	0.77	0.11	1.23	0.93	-0.87	1.12	0.44
Students had problems getting access to product	-1.36	0.65	0.04	-1.11	0.66	0.09	-1.11	0.58	0.06
Teacher had adequate time to prepare to use product	0.32	0.49	0.51	0.10	0.48	0.84	0.40	0.43	0.35
Teacher indicates school has computer specialist	1.18	0.75	0.12	0.78	0.64	0.22	1.17	0.70	0.09
Product is used in classroom	0.20	1.11	0.86	0.36	1.20	0.76	0.04	0.87	0.96
Teacher participated last year in technology professional development	0.95	0.61	0.12	0.57	0.59	0.34	1.00	0.47	0.04
School Interaction Variables									
Percent eligible for free lunch	-4.08	2.66	0.13	-5.05	2.59	0.06	-1.74	2.25	0.45
Student-teacher ratio	0.22	0.22	0.33	-0.35	0.22	0.11	0.01	0.19	0.98
School is in urban area	-1.61	0.98	0.11	-1.35	0.95	0.16	-1.33	0.80	0.10
Percent African American students	0.52	2.12	0.81	1.24	2.05	0.55	-0.32	1.69	0.85
Percent Hispanic students	3.50	3.50	0.32	4.67	3.43	0.18	1.11	2.75	0.69
Percent special education students	-0.31	2.06	0.88	-0.90	1.93	0.65	0.20	2.03	0.92
		χ^2	*p*-value		χ^2	*p*-value		χ^2	*p*-value
Chi-squared Tests									
Classroom and school variables		13.83	*p*>.50		12.03	*p*>.50		14.73	*p*>.50
Classroom variables		8.57	*p*>.50		5.20	*p*>.50		10.98	0.28
School variables		7.79	0.25		8.55	0.20		5.63	*p*>.50

Source: Author calculations. Other variables in the model include student pretest score, age, and gender; teacher experience, gender, and education; and school race-ethnicity composition, percent of students eligible for free lunch, percent participating in special education, and percent limited English proficient.

*No estimates are significantly different from zero based on the Benjamini-Hochberg correction for multiple comparisons, with false discovery rate of 0.05 (see Benjamini and Hochberg 1995). The multiple comparisons correction is computed separately for the classroom and school domains.

‡Statistically significant at the 0.05 level of significance.

Chapter III

Effects of Fourth Grade Reading Software Products

This chapter presents findings on the implementation of fourth grade reading software products and their effects on student test scores, and examines the relationships between the effects and school and classroom characteristics. The structure of this chapter follows that of the previous one. The technology products for the fourth grade substudy focus more on reading comprehension than on learning to decode text, but the study's design, data collection approach, and analysis approach were the same. Motivation for various analyses and methodological discussions are not repeated, and the reader is referred to the previous chapter for these discussions.

The fourth grade study included four reading products that were implemented in nine districts and 43 schools. The sample included 118 teachers and 2,265 students. The four products were Leapfrog (published by Leaptrack), Read 180 (published by Scholastic), Academy of Reading (published by Autoskill), and Knowledgebox (published by Pearson Digital Learning).

A. Implementation Analysis

The implementation analysis focused on the same six areas as the first grade study: features of the products in the study; teacher training and support to use products; the duration, extent, and location of product use; technical difficulties in using them; role of products in the curriculum; and the effects that product use had on classroom activities. The analysis was based on data gathered from three observations of classrooms and interviews with teachers, as well as from records collected from the products themselves.

Product Features

Three of the grade 4 study products provided practice and assessment geared to specific reading skills. Skills covered by these three products include aspects of reading comprehension (for example, identifying main ideas and analyzing settings, plots, and characters of stories), vocabulary, and literary analysis. The fourth product was a server-based collection of hundreds of resources (text passages, video clips, images, internet sites,

software modules, and so on) from which teachers could choose resources customized to their local curriculum. The study estimated the average licensing fees for the products to be about $96 a student for the school year, with a range of $18 to $184.

The three products were rated on the five features described in the previous chapter (tutorial opportunities, practice opportunities, individualization, feedback to teachers, and feedback to students). The nature of the fourth product made it incompatible with this product assessment process.

Table III.1 summarizes instructional features for the three fourth grade products that were assessed. The three products received similar ratings for most aspects of their instructional features. They all provide individualized instruction by setting student learning paths according to student skill levels. One product also allows teachers to specify which tutorial units students would receive. All three products provide ample opportunities for practicing reading skills and provide immediate feedback to students when in practice mode. Two products give feedback to students on mastery when used in assessment mode. All three products give feedback to teachers about student performance.

Teacher Training and Support

Product vendors trained teachers to use products during summer and early fall of 2004. Trainings typically were in the host district and sometimes in the host schools. The initial training averaged almost 7 hours, varying from 2.4 hours to 16.5 hours depending on the product. Topics included classroom management and alignment with standards and with the local curriculum; the trainings also gave teachers the opportunity to practice using the products. Nearly all teachers (94 percent) attended the initial training, according to attendance logs. In addition, teachers received other forms of support after initial training. On a questionnaire completed at the end of initial training, 98 percent of teachers said they were confident they could use the product. At the time of the first classroom observation, however, when most teachers had begun to use products, 53 percent told interviewers that the initial training had adequately prepared them to use the product. Ongoing support was delivered through various modes. Product representatives visited teachers (84 percent of teachers reported being visited by a representative), teachers received support through e-mail or telephone help desks (41 percent of teachers), and additional training was provided at schools (59 percent of teachers).

The study team worked with districts to identify hardware and software needs such as computers, headphones, memory, and operating system upgrades. Ultimately, the number of operational computers supported almost a one-to-one ratio of students to computers. Study observers noted that treatment teachers averaged almost 19 computers in their rooms, serving an average of about 20 fourth graders.[18]

[18]One of the products in the fourth grade study, which was the largest in terms of the number of teachers implementing it, provided portable units for each student. These units were not computers in the usual sense but are counted as such for our purposes here because they enabled students to use the software.

Table III.1. Instructional Features of Fourth Grade Reading Products.

Product	1. Tutorial Opportunities	2. Practice Opportunities	3. Individualization[a]									4. Feedback to Teachers			5. Feedback to Students[b]					
			Automatic			Teacher Input			Student Input						Immediate		Mastery		Diagnostic	
			T	P	A	T	P	A	T	P	A	Student Mastery	Learning Paths	Classroom Performance	P	A	P	A	P	A
AA	Some	Many	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓			✓	✓	
BB	Some	Many	✓	✓	✓		✓	✓		✓		✓	✓	✓	✓				✓	
CC	Many	Many	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓		✓	✓		

Source: Staff review.

[a]T = Tutorial mode, P = Practice mode, A = Assessment mode.

[b]Immediate feedback: Learner told whether response is correct immediately after completing module; Diagnostic feedback: Learner receives hints or other information concerning the probable source of error; Mastery feedback: Learner informed of the number correct and whether a skill or concept has been acquired.

Product Use

Treatment teachers used technology products more than control teachers. Teachers reported using study products about 1 to 2 hours a week. All treatment teachers used their assigned study product and reported about 100 minutes of weekly use or about 40 hours of annual use (see Table III.2). About 43 percent of treatment teachers also used other products, generally at lower levels (about 15 minutes of weekly use and 6 hours of annual use) than their use of study products. None of the control group teachers used a study product, but about half reported that they used another product; their usage of products was lower compared to treatment teachers, about 24 minutes of weekly use and 7 hours of annual use. The most commonly used other reading products were reading assessment products and a product that offered supplemental reading instruction along with a tool for creating reading assessments.

Table III.2 Teacher-Reported Use of Study Products and Other Reading Software Products, Fourth Grade.

	Study Products		Other Reading Products		
	Treatment Group	Control Group	Treatment Group	Control Group	*p*-value[a]
Percent of teachers using a product	100	0	43	49	.42
Minutes of weekly use	98	n.a.	15	24	.07
Hours of annual use	40	n.a.	6	7	.79
Sample size	**63**	**55**	**63**	**55**	

Source: Classroom observations and teacher interviews.

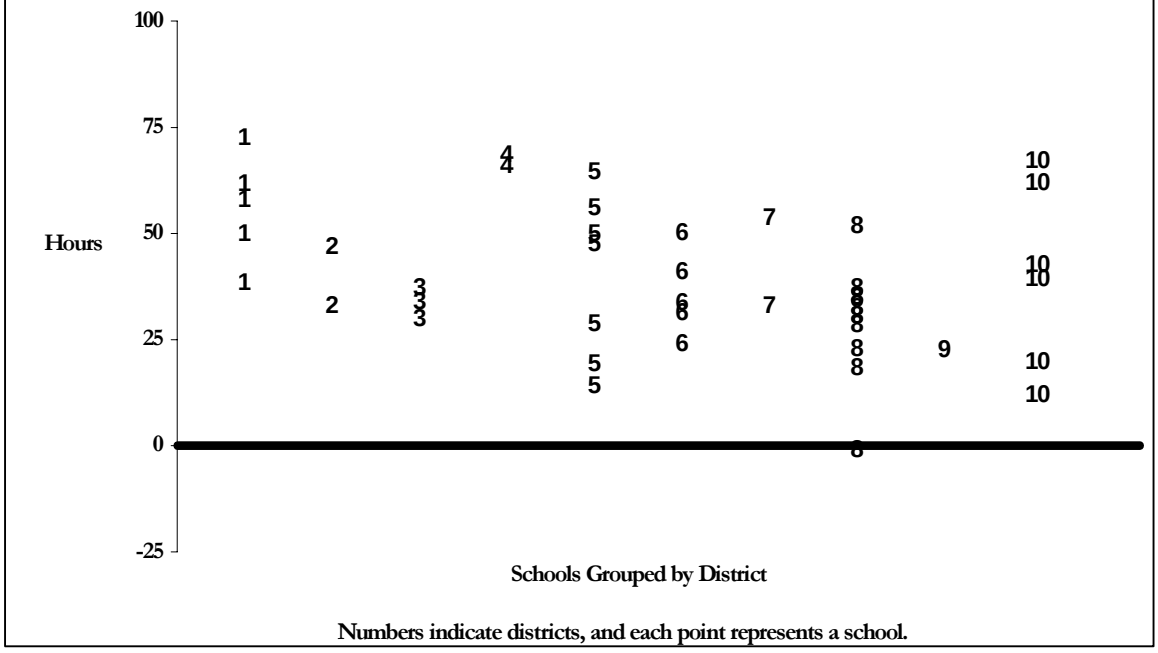
[a]Tests of differences were done using two-level hierarchical models (classroom and schools) with treatment assignment at level one and between-school variance at level two.

n.a. = not applicable.

The difference in product use between treatment and control teachers varied by district and school (Figure III.1). The difference averaged about 40 hours, with some schools having differences of 20 hours or less and others having differences of 75 hours. One school had a difference that was slightly negative, which occurs when control teachers used other products more than treatment teachers used the study product in addition to other products. Usage differences are also evident between districts. For example, the difference between treatment and control teacher product use was relatively large in district 4 compared to district 3.

Three of the four products included databases that provided some data on usage, and two products tracked usage closely. On days students used software, their usage was 12 to 28 minutes for the two products that tracked usage at this level of detail (see Table III.3). Annual use varied from 7 to 20 hours. For a typical 180-day school year, average daily usage was about 3 minutes for one product and 6 minutes for the other product that had adequate

Figure III.1
Difference in Annual Teacher-Reported Hours of Reading Technology Product Use Between Treatment and Control Classrooms, Fourth Grade



Source: Teacher Interviews.

Table III.3. Daily and Annual Usage From Product Records.

	Product A	Product B	Product C	Product D
Total days used during school year	15	90	n.a.	n.a.
Minutes of daily use (when used)	28	12	n.a.	n.a.
Hours of annual use	7	20	7	n.a.

Source: Product data on usage. One product did not yield usable data. One product did not collect total days used or minutes of daily use.

n.a. = not available or usable.

data on use. Assuming that reading is taught in a 90-minute block, the two products were used less than 10 percent of reading instruction time. This estimate refers only to time students spent on computers and not to other time related to products that students may have spent.[19]

[19]As noted in the previous chapter, the study itself may have increased usage because it purchased hardware and software upgrades directly (which allowed it not to go through district procurement processes),

Role of Products in the Curriculum and Time and Place of Use

Three products were used as supplements to the reading curriculum, and a fourth was a complete reading curriculum that included activities not involving computers. Seventy percent of treatment teachers indicated that they used the product as a supplement, and 25 percent said they used it as their core reading curriculum. All 13 teachers using the product designed to be a core curriculum reported using it as their core curriculum.

Teachers varied in how they scheduled product use. Most teachers (70 percent) used products exclusively during regular class time. About a quarter used products during other times (such as before school, lunchtime, or time usually set aside for science or social science), and 6 percent used it during class time and at other times. Products were used in regular classrooms (71 percent of teachers) or in computer labs (29 percent). Consistent with vendor recommendations, nearly all teachers (94 percent) reported being present when their students used products, and nearly all teachers (84 percent) reported reviewing reports two to three times a month or more frequently.

Technical Difficulties

Study observers reported seeing technical problems affecting student use of products in about a third (30 percent) of the time segments they observed (each observation consisted of four to five time segments). Most technical problems were brief and affected few students, the most common being problems logging on to products, having computers freeze and need rebooting, and issues with peripherals such as headphones.

Satisfaction With Products

Nearly all teachers (88 percent) said they would use the product in the next school year if given the choice. When asked what they would change in a second year of implementation, 29 percent said they would not change anything, 16 percent said they would start using products earlier in the school year, and 10 percent said that they would work on classroom management issues related to product use.

Effects on Classroom Activities

Classroom observation data show differences in classroom activities when products were used (see Table III.4).[20] Treatment teachers were much more likely than control

(continued)

paid teachers honoraria for attending training on using products, and relayed information to product developers about issues with using products that the team observed during classroom visits.

[20]The observation protocol called for observers to observe reading periods during which treatment teachers were using products. If observers had observed classrooms at random times, the differences shown in the table may have been smaller because products would not be in use during some of the observations.

Table III.4. Activities in Treatment and Control Classrooms, Fourth Grade (Percentage of Observation Intervals).

	Treatment Classrooms	Control Classrooms	*p*-value[a]
Teacher Role (Percent)[b]			
Leader	19	54	.00
Facilitator	46	32	.02
Monitor/observer	32	24	.03
Working on other tasks	12	5	.00
Other	8	4	.04
Instructional Activity (Percent)[b]			
Individual practice	84	39	.00
Lecture	8	23	.00
Question and answer	15	37	.00
Review of student work	4	5	.33
Other	5	9	.01
Student On-Task Behavior (Percent)			
Percent of time intervals with more than 90 percent of students on task	83	78	.28
Sample Size			
Number of classrooms	63	55	
Number of observations	549	471	

Source: Classroom observations at minutes 10, 20, and 30 for each class.

[a]Tests of differences were conducted using two-level hierarchical models with teacher's treatment status at level one and a school random effect at level two. The *p*-value is the smallest level of significance at which the null hypothesis that the difference between treatment class sessions and control class sessions equals zero can be rejected.

[b]Observers coded "all that apply." Percentages can sum to more than 100 percent.

teachers to act as facilitators and less likely than control teachers to be leaders. Treatment classroom students were more likely than control classroom students to be engaged in "individual practice" ($p < .05$). Participation in individual practice also was the most frequently observed activity in control classrooms, with participating in question-and-answer sessions directed by the teacher and listening to a lecture being the next most common student activities.

Classroom observers noted the proportion of students who were on task (engaged in the assigned academic task) during observation segments. In treatment classrooms, more than 90 percent of students were on task in 83 percent of observed segments. In control classrooms, more than 90 percent of students were on task in 76 percent of observed segments. The difference was not statistically different from zero ($p > .10$).

Whether product use was supplementing or replacing normal reading activities was assessed by asking teachers how much time they spent on reading instruction. Treatment teachers reported spending about 1 hour more on reading instruction (8.4 hours a week compared to 7.4 hours for control teachers), and the difference was statistically different from zero ($p < .05$). This evidence that reading instructional time increased with product use should be kept in mind in interpreting the findings.

B. Effects on Reading Test Scores

Effects on test scores were estimated using a model that accounted for the nesting of students in classrooms and classrooms in schools. The estimates show that effects on test scores generally were not statistically different from zero. Further analysis found that most teacher and school characteristics were uncorrelated with effects, but reported time that products were used was correlated with effects.

Treatment and Control Group Characteristics for Teachers and Students

Whether random assignment achieved its objective of creating equivalent treatment and control groups was assessed by examining characteristics of teachers and students in the two groups. Table III.5 shows teacher and student characteristics and results of statistical tests of equivalence. None of the differences of teacher and student characteristics is statistically different from zero. Nonetheless, the characteristics were entered as covariates to adjust for these differences.

Effects Were Not Statistically Different From Zero

As with first grade products, effects were estimated with a three-level hierarchical linear model. The first level included student characteristics (age, gender, and pretest score). The second level included classroom characteristics (most importantly, the treatment indicator of

Table III.5. Characteristics of Teachers and Students in Treatment and Control Classrooms, Fourth Grade.

	Treatment Classrooms	Control Classrooms	*p*-value[a]
Teacher Characteristics			
Years of experience (percent)	9.1	10.1	0.59
Has a master's degree (percent)	27	31	0.64
School has computer specialist (percent)	78	69	0.29
Received professional development on using technology last year (percent)	52	40	0.18
Female (percent)	81	89	0.22
Teacher Sample Size	**63**	**55**	
Student Characteristics			
Female (percent)	48	52	0.11
Age as of October 2004 (years)	9.7	9.7	0.62
Unadjusted score on fall SAT-10 Reading Test (NCE)	40.3	40.6	0.70
Sample Size	**1,231**	**1,034**	

Source: Teacher questionnaire, student records, and tests administered by study staff.

[a]Tests of treatment and control differences were conducted using a two-level hierarchical model with classroom treatment status as a fixed effect and school as a random effect (for teachers) and classrooms and schools as random effects (for students). The *p*-value of the difference is the *p*-value of the estimated treatment coefficient.

whether a classroom was assigned to use a product); the third level included school characteristics such as proportion of students receiving free or reduced-price lunch, race and ethnic composition, and the proportion that received special education services.

Table III.6 shows the average score differences for the SAT-10 reading test and its three subtests. None of the differences is statistically different from zero, and effect sizes are small. Additional analyses also found that products did not affect whether students were low scorers (Table III.7).[21]

Figures III.2 and III.3 depict the variation in school effect sizes by district and by product.[22] Most of the variability of the score difference is between districts (37 percent of the variance of school effects is within districts and 63 percent is between districts). However, the figure shows that differences vary between schools even within the same district. For example, in district 10, effect sizes ranged between -0.25 and 0.25.

Table III.6. Spring Reading Test Score Differences in Treatment and Control Classrooms, Fourth Grade.

	Treatment Classroom Average Score (NCE)	Control Classroom Average Score (NCE)	Difference	Effect Size	*p*-value
Stanford Achievement Test (Tenth Edition)					
Overall score	42.09	41.68	0.41	0.02	0.48
Subtest scores					
Vocabulary	43.42	43.07	0.35	0.02	0.56
Word study skills	43.05	43.15	-0.10	0.00	0.90
Comprehension	41.34	40.71	0.63	0.03	0.33

Note: See Appendix B for details of the estimation model. Variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity, percent of students in special education, percent eligible for free lunch, and percent limited English proficient; and student, classroom, and school random effects.

The treatment classroom average score reported in the table is the control classroom average score plus the treatment effect. It differs from the unadjusted treatment classroom score.

Effect sizes are based on the standard deviation of the control group for the spring test.

[21]Two-level models also were estimated for students in each quartile (based on the fall score) to assess product effects across the score distribution. Estimated effects (in NCE units) and *p*-values for the four quartiles were -0.32 (0.69), 0.51 (0.64), 1.77 (0.21), and 0.94 (0.71). None are statistically different from zero.

[22]To ensure products cannot be identified, letters used to identify products here do not correspond to letters used earlier in the chapter.

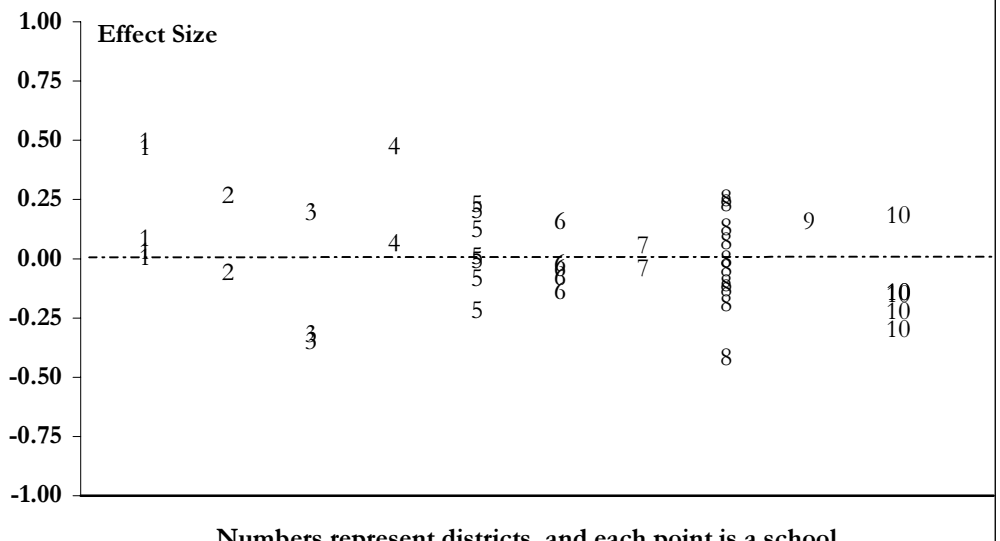
Table III.7. Effect on Percent of Students in Lowest Third of Reading Test Scores.

	Treatment Classroom Percentage	Control Classroom Percentage	Difference	Effect Size	*p*-value
Percent of students below 33rd percentile of spring reading test	52.8	52.9	0.1	0.01	0.97

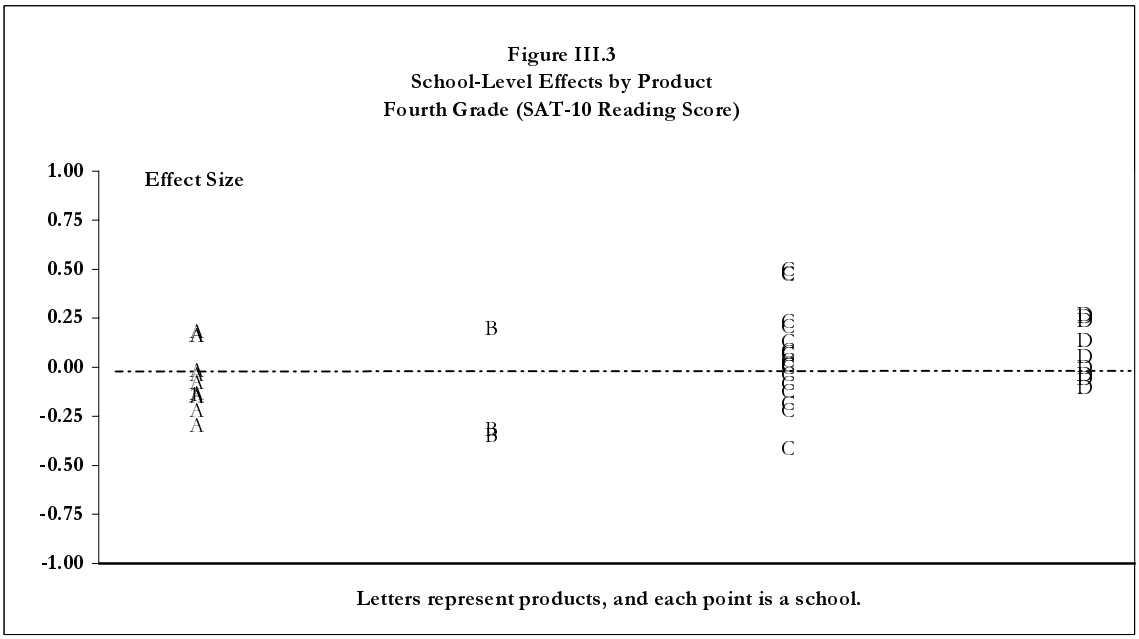
Note: Other variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity; percent of students in special education; percent eligible for free lunch; and student, classroom, and school random effects.

The treatment classroom percentage reported in the table is the control classroom percentage plus the treatment effect. It differs from the unadjusted treatment percentage. The effect size is calculated using the Cox index (the log odds ratio divided by 1.65).

Figure III.2
School-Level Effects by District
Fourth Grade (SAT-10 Reading Score)



Numbers represent districts, and each point is a school.



Note: Statistical significance of average effect sizes cannot be inferred from the figure because student and teacher sample sizes differ between schools.

Classroom and School Moderators

Whether classroom and school characteristics moderate effects was investigated by interacting the treatment effect with classroom characteristics at the second level of the model and by estimating a model of the treatment effect as a function of school characteristics (in the third level). School characteristics included race and ethnicity and the percent of students qualified for free lunches. Product usage (both study products and other products) also was entered.

Table III.8 shows estimates of moderator effects, with positive coefficients indicating that a characteristic is positively correlated with effects and a negative coefficient indicating the opposite. For the overall reading score, statistical tests indicate that classroom and school characteristics are jointly related to effects, but only classroom characteristics are related to effects in separate tests (school characteristics were not statistically significant on their own). The amount of study product usage was a statistically significant characteristic that moderated effects. The estimates indicate that the product effect size would be larger by 0.10 if product usage were larger by a standard deviation (about 17 hours a year). For vocabulary, no characteristics moderated effects. For word reading, the amount of study product use, teacher experience, whether teachers received technology professional development in the previous year, and the percent of students who were Black moderated effects. The amount of use was positively correlated with effects, and other characteristics were negatively correlated with effects.

These moderator relationships need to be interpreted carefully. As noted in the previous chapter, product usage may be related to decisions by teachers and schools after data collection began and possibly after teachers and schools had initial indications of whether the product appeared to be effective.

C. Conclusions

This study experimentally tested the effects of select reading technology products in fourth grade classrooms that volunteered to participate. Although not a nationally representative sample, the schools participating in the study were more likely than average to be in urban areas and had higher-than-average levels of poverty and students who were African American or Hispanic. Key findings are the following:

1. **Teachers Were Trained and Used Products.** Nearly all treatment teachers (94 percent) participated in the initial training from the vendor on using the product, and all implemented the assigned software, at least to some extent. Usage recorded by three of the products was less than 10 percent of reading instruction time.

2. **Differences in Test Scores Were Not Statistically Different From Zero.** Overall reading scores for students in treatment and control classrooms were 42.1 and 41.7, respectively (in normal-curve-equivalent units). The difference was not statistically different from zero.

3. **Some Classroom and School Characteristics Were Correlated With Product Effects.** For the overall score, a significant correlation was found between product effects and product usage. For word reading scores, significant correlations were found between product effects and several characteristics including product usage. The possibility that other factors are related to these characteristics warrants caution in interpreting the correlations.

Table III.8. Interactions Between Moderating Variables and Effects: SAT-10 Reading Test, Fourth Grade.

	Total Score (NCE)				Vocabulary (NCE)				Word Study Skills (NCE)				Comprehension (NCE)			
	coefficient	standard error	*p*-value		coefficient	standard error	*p*-value		coefficient	standard error	*p*-value		coefficient	standard error	*p*-value	
Classroom Interaction Variables																
Years of teaching experience	-0.17	0.07	0.02		-0.08	0.08	0.32		-0.24	0.09	0.01	*	-0.15	0.09	0.13	
Teacher is female	0.27	1.44	0.85		0.58	1.41	0.68		-0.66	1.74	0.70		0.77	2.58	0.77	
Teacher has master's degree	-0.16	1.34	0.91		-0.49	1.51	0.74		-0.66	1.86	0.72		-0.17	1.52	0.91	
Time of study product use (in hundreds)	12.11	2.99	0.00	*	5.97	3.49	0.09		12.92	5.59	0.02		12.78	3.70	0.00	*
Problems getting access to product	0.08	0.86	0.93		-0.02	0.86	0.99		-1.58	1.51	0.30		1.13	1.19	0.35	
Adequate time to prepare to use product	0.26	0.75	0.73		1.84	0.83	0.03		-1.89	1.19	0.11		0.42	1.10	0.70	
School has a computer specialist	-0.23	0.64	0.72		0.47	1.03	0.65		0.85	1.19	0.47		-1.43	0.93	0.13	
Product is used in classroom	2.65	1.07	0.02		2.07	1.08	0.06		4.65	1.46	0.00	*	1.86	1.34	0.17	
Participated last year in technology professional development	-1.13	0.74	0.13		-1.36	0.75	0.07		-2.37	0.96	0.02	*	-0.02	1.03	0.98	
School Interaction Variables																
Percent eligible for free lunch	1.69	2.40	0.48		2.92	2.90	0.32		4.07	3.02	0.18		-0.83	2.96	0.78	
Student-teacher ratio	-0.10	0.15	0.48		-0.02	0.14	0.88		-0.34	0.23	0.14		-0.05	0.16	0.77	
School is in urban area	-1.73	0.83	0.04		-0.95	0.98	0.34		-0.76	1.53	0.62		-2.20	0.95	0.02	
Percent of African American students	-3.72	1.74	0.03		-1.44	1.82	0.43		-6.85	1.93	0.00	*	-2.31	2.46	0.35	
Percent of Hispanic students	-4.67	1.99	0.02		-1.91	2.26	0.40		-8.34	3.39	0.02		-2.76	2.81	0.33	
Percent special education students	-1.57	2.77	0.57		-4.46	3.57	0.21		-4.38	6.67	0.51		0.44	3.28	0.89	
		χ^2	*p*-value			χ^2	*p*-value			χ^2	*p*-value			χ^2	*p*-value	
Chi-Squared Tests																
Classroom and school variables		39.70	0.00	‡		17.91	0.27			37.84	0.00	‡		24.15	0.06	
Classroom variables		26.40	0.00	‡		14.41	0.11			27.41	0.00	‡		15.19	0.09	
School variables		6.48	0.37			2.51	>.50			6.41	0.38			4.16	>.50	

Source: Author calculations. Other variables in the model include student pretest score, age, and gender; teacher experience, education, and gender; and school race-ethnic composition, percent free lunch, urban area, percent special education, and time using other technology products.

*Significantly different from zero based on the Benjamini-Hochberg correction for multiple comparisons, with false discovery rate of 0.05. The multiple comparisons correction is computed separately for classroom and school characteristics.

‡Statistically significant at the 0.05 level of significance.

Chapter IV

Effects of Sixth Grade Math Software Products

This chapter presents implementation findings for sixth grade math technology products and estimates of their effects and the relationships between effects and school and classroom characteristics. The technology products in the sixth grade substudy focused on math and pre-algebra instruction. The study's design, data collection approach, and analysis approach were the same as for reading products. The sixth grade study included three products that were implemented in 10 districts and 28 schools. The sample included 81 teachers and 3,136 students. The three products were Larson Pre-Algebra (published by Houghton-Mifflin), Achieve Now (published by Plato), and iLearn Math (published by iLearn).

A. Implementation Analysis

As noted in previous chapters, the main questions in the implementation analysis related to product features, use in classrooms, and effects on teacher and student activities in classrooms. Classroom use had multiple aspects that were investigated: teacher training and support in using products; the duration, extent, and location of product use; technical difficulties in using products; and role of products in the curriculum.

Product Features

The three products selected for the study are similar to other products that schools use for math instruction in sixth grade, though they do not represent a statistical sample of all available technology in sixth grade math instruction. All three products in the sixth grade study provide tutorial, practice, and assessment opportunities in mathematics. Skills covered include: operations with fractions, decimals, and percents; plane and coordinate geometry; ratios, rates, and proportions; operations with whole numbers and integers; probability and data analysis; and measurement. Each product provides students with opportunities to apply skills through solving practice problems; however, they vary in the level of tutorial assistance. Vendors recommended that products be used between 120 and 225 minutes a week. The study estimated the average licensing fees for the products to be about $18 a student for the school year, with a range of $9 to $30.

Many instructional features of the three products are similar (see Table IV.1). Two products contain many tutorial modules; the other provides only a few tutorial modules. All three provide ample opportunities to practice skills and give students immediate and diagnostic feedback when in practice mode. They all provide for individualization of instruction, automatically setting individual student learning paths depending on a student's skill level. Two products allow teachers to specify manually which tutorial units students should receive and let students indicate skills they would like to practice. All three provide teachers with feedback on average individual student and class performance, and one provides feedback at the level of individual questions. None of the products provides feedback to students in assessment mode.

Teacher Training and Support

Vendor training sessions generally took place in host districts and sometimes host schools during summer or early fall of 2004. The initial training lasted about 6 hours and varied by product from 4 hours to about 8 hours. Training topics included classroom management and alignment with standards and the local curriculum, and training sessions included opportunities to practice with the technology. Nearly all teachers (98 percent) attended the initial training, according to attendance logs. On a questionnaire completed at the end of initial training, 94 percent of teachers said they were confident they could use the product. At the time of the first classroom observation, when most teachers had begun to use products, the proportion of teachers who thought the initial training had adequately prepared them to use the product had fallen from 94 percent to 57 percent. Vendors delivered ongoing support in several modes. Product representatives visited teachers (66 percent of teachers reported receiving this kind of help); vendors also provided support through e-mail or telephone help desks (40 percent) and additional training at schools (30 percent).

The study team worked with districts to identify hardware and software needs such as computers, headphones, memory, and operating system upgrades; it also purchased a set of mobile laptop carts for one district where access to school computer labs was too constrained to support adequate student use. Study observers noted that treatment teachers averaged almost 18 computers in their rooms to serve an average class size of about 22 sixth graders (including the mobile labs purchased by the study).

Duration and Extent of Product Use

Nearly all teachers implemented products to some degree and reported using products for an average of 25 weeks and almost 2 hours a week. Table IV.2 shows the hours that students used study products, according to treatment teachers. Teachers also used other products in addition to study products--27 percent of control teachers and 11 percent of treatment teachers used computer-based math products other than those in the study. Use of other products was less than an hour a year in treatment classes and less than 3 hours a year in control classes (see Table IV.3). None of the control teachers used a study product.

Table IV.1. Instructional Features of Sixth Grade Mathematics Products.

Product	1. Tutorial Opportunities	2. Practice Opportunities	3. Individualization[a]									4. Feedback to Teachers			5. Feedback to Students[b]					
			Automatic			Teacher Input			Student Input						Immediate		Mastery		Diagnostic	
			T	P	A	T	P	A	T	P	A	Student Mastery	Learning Paths	Classroom Performance	P	A	P	A	P	A
A	Few	Many	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓		✓			
B	Many	Many	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	
C	Many	Many	✓	✓	✓							✓		✓	✓				✓	

Source: Staff review.

[a]T= Tutorial mode, P = Practice mode, A = Assessment mode.

[b]Immediate feedback: Learner told whether response is correct immediately after completing module; Diagnostic feedback: Learner receives hints or other information concerning the probable source of error; Mastery feedback: Learner informed of the number correct and whether a skill or concept has been acquired ("mastered").

Table IV.2. Teacher-Reported Use of Study Products and Other Mathematics Products, Sixth Grade.

	Study Products			Other Mathematics Products		
	Treatment Group	Control Group	*p*-value[a]	Treatment Group	Control Group	*p*-value[a]
Percent of teachers using a product	100	0	n.a.	11	27	.03
Minutes of weekly use	116	n.a.	n.a.	2	8	.06
Hours of annual use	51	n.a.	n.a.	<1	3	.07
Sample size	**47**	**34**		**47**	**34**	

Sources: Classroom observations and teacher interviews.

[a]Tests of difference were conducted using a two-level hierarchical model (classrooms and schools), with the treatment effect at level one and between-school variance at level two.

n.a. = not applicable.

Table IV.3. Daily and Annual Usage From Product Records.

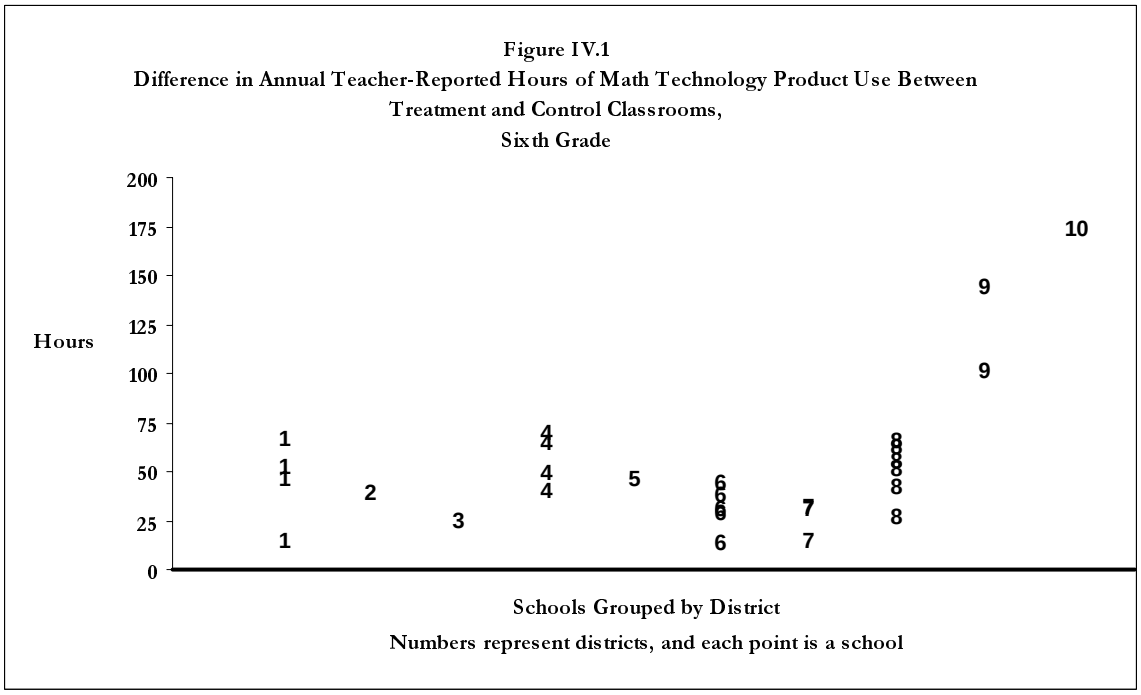
	Product A	Product B	Product C
Total days used during school year	25	95	n.a.
Minutes of daily use (when used)	26	43	n.a.
Hours of annual use	12	73	n.a.

Source: Product records. One product did not maintain data on student usage.

n.a. = not available or usable.

The difference in product use between treatment and control teachers varied by district and school (Figure IV.1) and averaged about 52 hours, with a standard deviation (between schools) of about 35 hours. No schools had a negative difference, and some had high levels of differences in usage, more than 150 hours a year. Usage differences also are evident between districts. For example, the difference between treatment and control teacher technology usage was relatively large in district 7 compared to district 9.

Two products maintained detailed records on student usage on a weekly and annual basis. (The third product did not maintain records on student use.) Data for one product showed 25 days of use, with an average of 26 minutes per use for an average of 12 hours of student use for the school year. The other showed 95 days of use, 43 minutes a day, and 73 hours for the school year. The product with lower usage had a much larger share of the student sample, however—more than 10 times larger than the other product—and therefore



Source: Teacher Interviews.

contributed more to overall usage, estimated to be about 17 hours a year.[23] Assuming math instructional periods were 50 minutes in a 180-day school year, on average, products were used for about 11 percent of math instructional time.[24]

Role of Products in the Curriculum and Time and Place of Use

Two products were designed as supplements to the core mathematics curriculum, and one was designed to function either as the core curriculum or as a source of supplementary instruction. When treatment teachers were asked how they used products in their teaching, 83 percent said they used the product as a supplement and 11 percent said they used it as their core curriculum.

[23]The two types of usage data—scheduled use according to teachers and actual use according to product records—relate to different concepts of usage but were highly correlated ($r = 0.74$ for weekly use and $r = 0.95$ for annual use).

[24]As noted in previous chapters, the study may have increased usage from typical levels of first-time users because it purchased hardware and software upgrades without going through district procurement processes, paid teachers honoraria for attending training on using products, and relayed information to product developers about issues with using products that the team observed during classroom visits.

Most teachers (76 percent) scheduled the product to be used during their regular class time slot. Of the remainder, 11 percent said students used the software during other times, such as before school or during lunch, and 13 percent said their students used the product during class time and at other times.

Products were used in the regular classrooms and in computer labs with about equal frequency (53 percent in classrooms and 47 percent in labs). Consistent with vendor recommendations, nearly all teachers (89 percent) reported being present when their students used products. A total of 51 percent reported reviewing assessment reports two to three times a month or more, and 38 percent reported reviewing assessment reports once a week or more.

Technical Difficulties

Study observers reported technical problems that affected at least one student in about a quarter (28 percent) of the time segments they observed (each observation consisted of four to five time segments). Most technical problems, such as problems logging on to products or having computers freeze and need rebooting, were brief and affected only a few students.

Satisfaction with Products

Nearly all teachers (92 percent) said they would use the product in the next school year if they had the opportunity. When asked what they would change in the second year, 27 percent said they would not change anything and 12 percent said that they would spend more time becoming familiar with the software (no other response category included more than 10 percent of teachers).

Impact on Classroom Activities

Table IV.4 shows that treatment teachers were observed being facilitators during a larger proportion of time segments than control teachers—61 percent compared to 16 percent ($p < .05$).[25] Students in treatment classrooms were also more likely to be engaged in individual work than students in control classrooms—76 percent compared to 26 percent ($p < .05$). Listening to a lecture was the most frequently observed student activity in control classrooms.

[25]The differences were estimated using a two-level hierarchical linear model with teachers in the first level and schools in the second level.

Table IV.4. Activities in Treatment and Control Classrooms, Sixth Grade (Percentage of Observations).

	Treatment Classrooms	Control Classrooms	*p*-value[a]
Teacher Role (Percent)[b]			
Leader	16	66	.00
Facilitator	60	21	.00
Monitor/observer	16	13	.41
Working on other tasks	14	9	.02
Other	4	1	.33
Instructional Activity (Percent)[b]			
Individual work	76	26	.00
Lecture	10	35	.00
Question and answer	5	21	.00
Review of student work	2	14	.00
Other	4	6	.20
Student On-Task Behavior (Percent)			
Percent of intervals with more than 90 percent of students on task	82	71	.06
Sample Size			
Number of classrooms	47	34	
Number of observation intervals	426	291	

Source: Classroom observation intervals at minutes 10, 20, and 30 for each class.

[a]Tests of differences were conducted using two-level hierarchical models with teacher's treatment status at level one and a school random effect at level two. The *p*-value is the smallest level of significance at which the null hypothesis that the difference between treatment class sessions and control class sessions equals zero can be rejected.

[b]Observers coded "all that apply." Percentages can add to more than 100.

B. Effects on Math Test Scores

Effects on test scores were estimated using a model that accounted for the nesting of students in classrooms and classrooms in schools. The estimates show that effects on test scores generally were not statistically different from zero and most teacher and school characteristics were uncorrelated with effects.

Treatment and Control Group Characteristics of Teachers and Students

Teachers and students generally were similar in treatment and control classrooms, with two exceptions (Table IV.5). After adjusting for multiple comparisons, no teacher or student characteristics were statistically different.

Table IV.5. Characteristics of Teachers and Students in Treatment and Control Classrooms, Sixth Grade.

	Treatment Classrooms	Control Classrooms	*p*-value[a]
Teacher Characteristics			
Years of experience (years)	10.3	11.1	0.69
Has a master's degree (percent)	30	35	0.61
School has computer specialist (percent)	91	74	0.03
Received professional development on using technology last year (percent)	38	32	0.47
Female (percent)	65	82	0.07
Teacher Sample Size	**47**	**34**	
Student Characteristics			
Female (percent)	52	53	0.66
Age as of October 2004 (years)	11.6	11.7	0.07
Unadjusted score on fall SAT-10 math test (NCE)	46.4	46.9	0.85
Student Sample Size	**1,878**	**1,258**	

Sources: Teacher questionnaires, student records, and tests administered by study staff.

[a]Tests of treatment and control differences were conducted using a two-level hierarchical model with classroom treatment status as a fixed effect and school as a random effect (for teachers) and classrooms and schools as random effects (for students). The *p*-value of the difference is the *p*-value of the estimated treatment coefficient.

Effects Were Not Statistically Different From Zero

Table IV.6 shows score differences for the overall SAT-10 math test and its two subtests from the hierarchical linear model. Effect sizes are similar, about 5 percent to 7 percent. None were statistically different from zero. An analysis of whether products reduced the proportion of students who scored in the lower third on the SAT-10 (below the 33rd percentile) found that products did not have an effect (Table IV.7).[26]

Figures IV.2 and IV.3 depict the variation in school effect sizes by district and by product.[27] For example, in district 8, effect sizes in the five schools ranged from almost -0.40 to 0.40. An analysis of variance of the school effect size indicates that 62 percent of its variance is between districts and 38 percent is within districts.

[26]Two-level models also were estimated for students in each quartile (based on the fall score). Estimated effects and *p*-values for the four quartiles were 2.21 (0.12), 2.62 (0.07), 2.31 (0.11), and 1.77 (0.32).

[27]School effects were estimated using a regression model in which test scores are regressed on student and teacher characteristics and the treatment indicator is interacted with an indicator for each school. Standard errors were adjusted for classroom clustering.

Table IV.6. Spring Math Test Score Differences in Treatment and Control Classrooms, Sixth Grade.

	Treatment Classroom Average Score (NCE)	Control Classroom Average Score (NCE)	Difference	Effect Size	*p*-value
Stanford Achievement Test (Tenth Edition)					
Overall score	52.20	50.77	1.43	0.07	0.15
Subtest scores					
Procedures	51.81	50.31	1.50	0.07	0.21
Problem solving	52.20	51.08	1.12	0.05	0.18

Source: Author calculations; see Appendix B for details of the estimation model. Variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity, percent of students in special education, and percent eligible for free lunch; and student, classroom, and school random effects.

The treatment classroom average score reported in the table is the control classroom average score plus the treatment effect. It differs from the unadjusted treatment classroom score.

Effect sizes are based on the standard deviation of the control group for the spring test.

Table IV.7. Effect on Percent of Students in Lowest Third of Math Test Scores.

	Treatment Classroom Percentage	Control Classroom Percentage	Difference	Effect Size	*p*-value
Percent of students below 33rd percentile of spring math test	32.4	33.1	-0.7	0.03	0.91

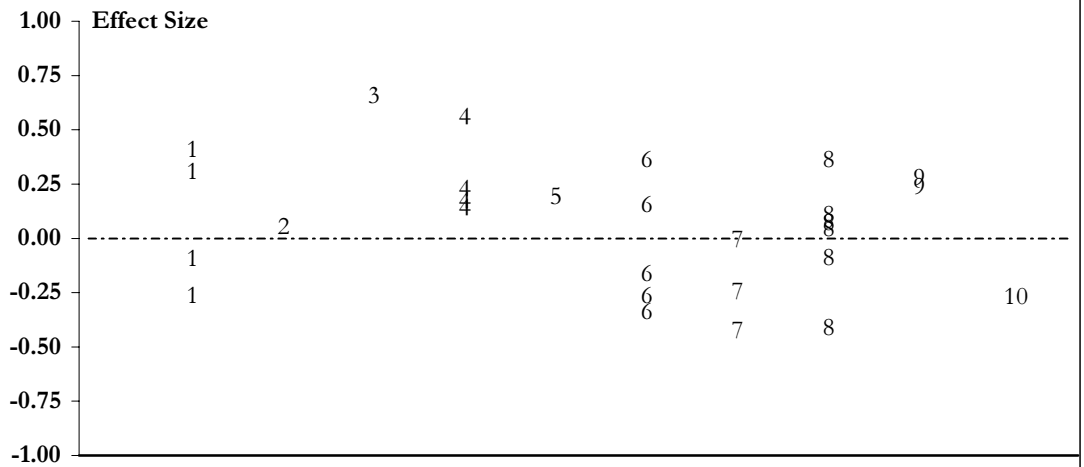
Note: Other variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity; percent of students in special education; percent eligible for free lunch; and student, classroom, and school random effects.

The treatment classroom percentage reported in the table is the control classroom percentage plus the treatment effect. It differs from the unadjusted treatment percentage. The effect size is calculated using the Cox index (the log odds ratio divided by 1.65).

Classroom and School Moderators

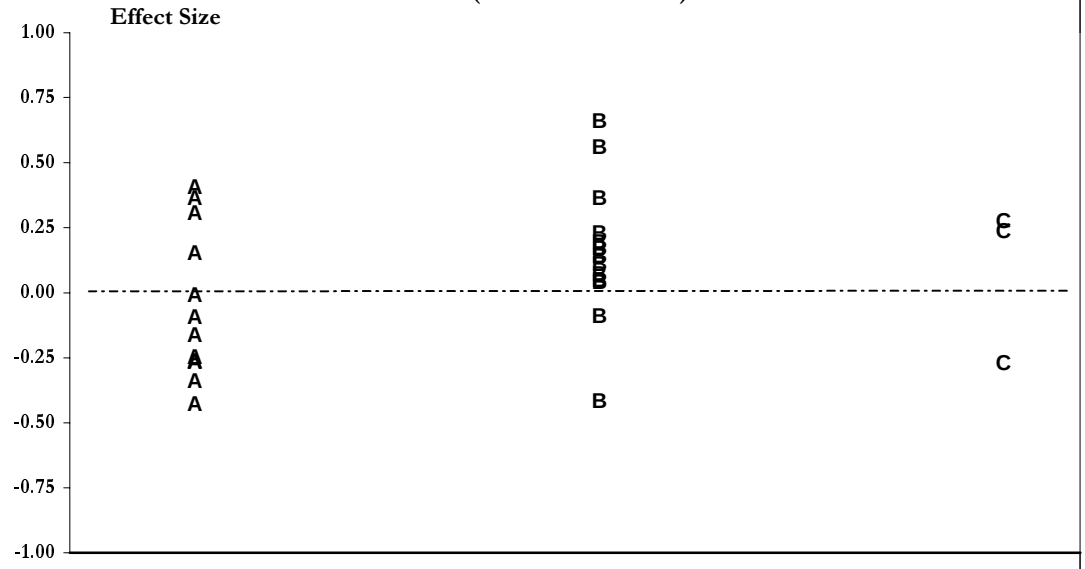
Table IV.8 shows estimates of the relationship between classroom and school characteristics and product effects, with positive coefficients indicating that a characteristic is positively correlated with effects and a negative coefficient indicating the opposite. As a group, classroom and school characteristics were not statistically significantly related to effectiveness, and neither school nor classroom characteristics were related to effects in separate tests. The amount of time study products were used was not correlated with product effects.

Figure IV.2
School-Level Effect Sizes by District
Sixth Grade (SAT-10 Math Score)



Numbers represent districts, and each point is a school.

Figure IV.3
School-Level Effect Sizes by Product
Sixth Grade (SAT-10 Math Score)



Letters represent products, and each point is a school.

Note: Statistical significance of average effect sizes cannot be inferred from the figure because student and teacher sample sizes differ between schools.

Table IV.8. Interactions Between Moderating Variables and Effects: SAT-10 Math Test, Sixth Grade.

	Total Score (NCE)			Procedures (NCE)			Problem Solving (NCE)		
	coefficient	standard error	*p*-value	coefficient	standard error	*p*-value	coefficient	standard error	*p*-value
Classroom Interaction Variables*									
Years of teaching experience	0.14	0.13	0.26	0.04	0.15	0.79	0.21	0.59	0.73
Study product use reported by teacher	-3.60	2.52	0.16	-4.41	2.31	0.06	-2.57	13.23	0.85
Other product use reported by teacher	4.94	11.83	0.68	0.48	9.71	0.96	7.62	64.35	0.91
Students had problems getting access to product	-2.10	1.69	0.22	0.48	2.05	0.82	-2.86	8.38	0.73
Teacher had adequate time to prepare to use product	2.04	1.36	0.14	2.86	1.38	0.04	1.30	6.42	0.84
Teacher indicates school has a computer specialist	0.48	3.27	0.88	-1.24	3.03	0.68	0.98	16.38	0.95
Product is used in classroom	-1.02	1.99	0.61	-2.82	2.29	0.22	-0.22	8.68	0.98
School Interaction Variables*									
Percent eligible for free lunch	-4.18	8.46	0.62	1.45	9.66	0.88	-7.92	39.06	0.84
Student-teacher ratio	-0.41	0.52	0.43	-0.16	0.60	0.79	-0.54	2.28	0.81
School is in urban area	5.12	3.78	0.18	3.29	4.54	0.47	5.83	17.82	0.74
Percent of African American students	-1.17	6.91	0.87	-7.67	8.17	0.35	2.55	28.30	0.93
Percent of Hispanic students	-3.41	6.31	0.59	-6.41	7.36	0.39	-0.84	30.29	0.98
Percent special education students	9.00	11.10	0.42	5.44	11.92	0.65	11.05	56.10	0.85
		χ^2	*p*-value		χ^2	*p*-value		χ^2	*p*-value
Chi-Squared Tests‡									
Classroom and school variables		16.05	0.19		15.90	0.19		17.00	0.15
Classroom variables		7.24	0.30		8.23	0.22		7.95	0.24
School variables		7.20	0.30		6.66	0.35		7.73	0.26

Source: Author calculations; see Appendix B for details of the estimation model. Variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity, percent of students in special education, and percent eligible for free lunch; time of use of other technology products; and student, classroom, and school random effects.

*No characteristics are significantly different from zero based on the Benjamini-Hochberg correction for multiple comparisons, with false discovery rate of 0.05. The multiple-comparisons correction is computed separately for classroom and school characteristics.

‡No chi-squared tests are statistically significant at the 0.05 level of significance.

C. Conclusions

This study experimentally tested the effects of three math software products in sixth grade classrooms that volunteered to participate. The schools participating in the study were more likely to be in urban areas and had higher-than-average levels of poverty and larger percentages of minority students than U.S. public schools as a whole. Key findings include:

1. **Teachers Were Trained and Used Products.** Nearly all treatment teachers (98 percent) received training from the vendor in using the product, and all implemented the assigned product to some extent. Overall usage was estimated at about 11 percent of math instructional time.

2. **Effects on Test Scores Were Not Statistically Different From Zero.** Overall math scores for students in treatment and control classrooms were 52.2 and 50.8, respectively (in normal-curve-equivalent units). The difference was not statistically different from zero.

3. **School and Classroom Characteristics Were Not Related to Product Effects.** Time of product use and other school and classroom characteristics were uncorrelated with product effects.

Chapter V

Effects of Algebra Software Products

This chapter describes the implementation of algebra technology products and presents estimates of their effects and of relationships between effects and conditions and practices in participating classrooms and schools. Many features of the study's design, data collection approach, and analysis approach were the same as for the reading and sixth grade math products substudies. However, the algebra study used the end-of-course algebra exam designed by Educational Testing Service (ETS) as its fall and spring measure of achievement. The end-of-course exam has the benefit of directly focusing on the subject matter and has been used in several other studies of algebra technology products.

The algebra study included three software products that were implemented in 23 schools within 10 districts. The sample included 69 teachers and 1,404 students. The three products were Cognitive Tutor Algebra (published by Carnegie Learning), Plato Algebra (published by Plato), and Larson Algebra (published by Houghton-Mifflin).

A. Implementation Analysis

The implementation analysis focused on six areas: features of the products in the study; teacher training and support on using products; the duration, extent, and location of product use; technical difficulties in using the products; their role in the curriculum; and the effects that product use had on classroom activities.

Product Features

The products in the algebra substudy all provide tutorial, practice, and assessment opportunities in algebra. Topics include common ones in algebra: functions, linear equations and inequalities, quadratic equations, linear expressions, polynomials, systems of equations and inequalities, and data analysis. Products provide students with opportunities, for example, to write, solve, and graph equations and inequalities; solve systems of equations and inequalities; factor polynomials; and analyze and make predictions from data. The study estimated the average licensing fees for the products to be about $15 a student for the school year, with a range of $7 to $30.

The products were rated as similar on many aspects of their instructional design (Table V.1). They all individualize instruction, automatically setting individual student learning paths depending on a student's particular skill level and also letting teachers or students manually select skills to practice. Two products include many tutorial modules; the other has fewer tutorials and provides additional tutorial opportunities in print materials intended to be used with the computer-based component. All three products provide ample opportunities for practicing algebra skills and give immediate diagnostic feedback when used in "practice" mode. Also, they all give teachers feedback on individual student and class performance for groups of items, and one gives teachers recommendations for individualizing students' learning paths based on student scores on assessments.

Teacher Training and Support

Treatment teachers were trained in their host districts or schools during summer or early fall of 2004. The initial training provided by the three algebra product vendors lasted about 12 hours, varying from 4 to 23 hours. Topics included classroom management and alignment with standards and the local curriculum. Teachers were also able to practice using the product. Nearly all teachers (97 percent) attended the initial training, according to attendance logs. On a questionnaire completed at the end of initial training, 81 percent of teachers said they were confident they were prepared to use the product with their students. By the time of the first classroom observation, the proportion of teachers who believed the initial training had adequately prepared them to use the product had dropped from 81 percent to 66 percent. Ongoing support was provided by vendors in various modes. Some had company representatives visit teachers (28 percent of teachers reported receiving this kind of support), supported e-mail or telephone help desks (36 percent of teachers said they were aware of or used this kind of support), and provided additional training at schools (which 14 percent of teachers reported receiving).

The study team purchased some hardware and software upgrades in host schools, but to a lesser extent than in other grade levels. Study observers noted that treatment teachers averaged about 19 computers in the rooms (whether in labs or regular classrooms) where products were being used.

Duration and Extent of Product Use

All treatment teachers implemented the products to some degree (see Table V.2). Treatment teachers reported that they used study products for about 23 weeks during the school year for almost 120 minutes a week, totaling 46 hours for this period. Only 4 percent of treatment teachers and 10 percent of control teachers reported using math products besides those in the study. On average, use of other math software products was about 3 hours a year for control teachers and 4 hours for treatment teachers (Table V.3). The most commonly used products were online study materials to supplement a commercial textbook and a website developed by a state department of education that provided study problems for students to prepare for the state tests. None of the control teachers used a study product.

Table V.1. Instructional Features of Algebra Products.

Product	1. Tutorial Opportunities	2. Practice Opportunities	3. Individualization[a]									4. Feedback to Teachers			5. Feedback to Students[b]					
			Automatic			Teacher Input			Student Input						Immediate		Mastery		Diagnostic	
			T	P	A	T	P	A	T	P	A	Student Mastery	Learning Paths	Classroom Performance	P	A	P	A	P	A
A	Few	Many		✓	✓		✓	✓	✓	✓		✓	✓	✓	✓		✓			
B	Many	Many	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓				✓	
C	Many	Many	✓	✓	✓	✓	✓		✓	✓		✓		✓	✓		✓	✓	✓	

Source: Staff review.

[a]T= Tutorial mode, P = Practice mode, A = Assessment mode.

[b]Immediate feedback: Learner told whether response is correct immediately after completing module; Diagnostic feedback: Learner receives hints or other information concerning the probable source of error; Mastery feedback: Learner informed of the number correct and whether a skill or concept has been acquired.

Table V.2. Teacher-Reported Use of Study Products and Other Math Software.

	Algebra Study Products		Other Mathematics Products		
	Treatment Group	Control Group	Treatment Group	Control Group	*p*-value[a]
Percent of teachers using a product	100	0	4	10	.01*
Minutes of weekly use	118	--	11	10.5	.84
Hours of annual use	46	--	3.6	2.6	.63
Sample size	**39**	**32**	**39**	**32**	

Sources: Classroom observations and teacher interviews.

[a]Tests of mean difference were done within two-level hierarchical models, with the effect of treatment assignment at level one and between-school variance at level two.

*Statistically different at the .05 level.

Table V.3. Daily and Annual Usage.

	Product A	Product B	Product C	Overall
Total days used during school year	40	9	20	22
Minutes of daily use (when used)	41	28	38	34
Hours of annual use	28	5	13	15

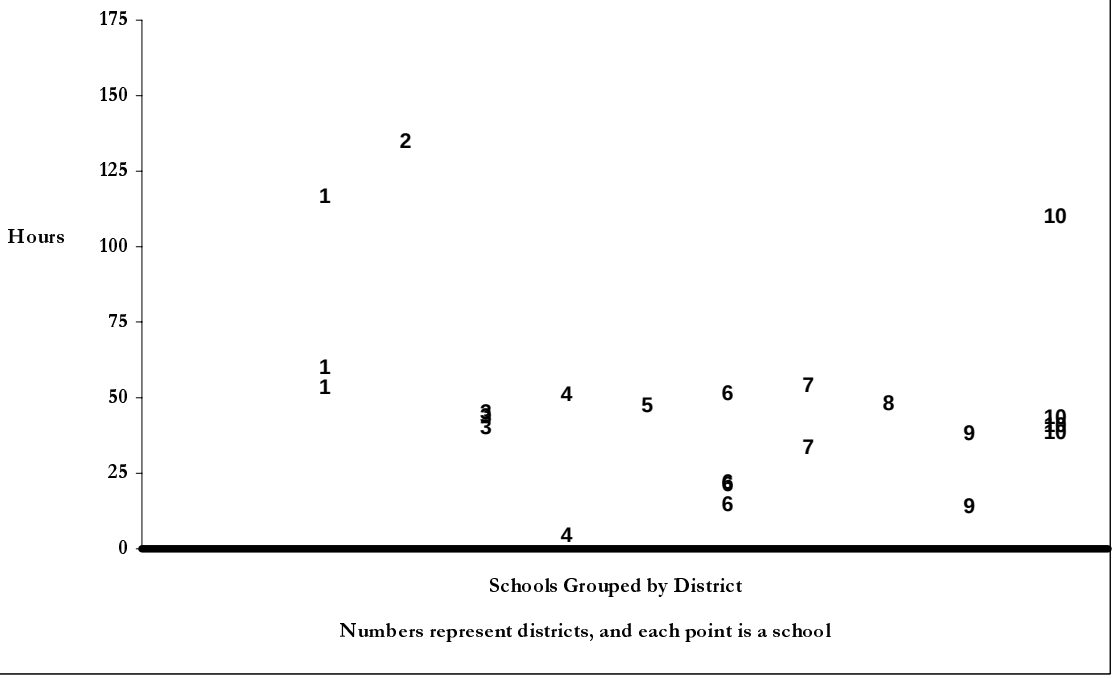
Source: Product records. Overall estimates are based on product estimates weighted by student sample sizes.

The difference in product use between treatment and control teachers varied by district and school (Figure V.1). The difference averaged about 49 hours, with a standard deviation (between schools) of about 31 hours. No schools had a negative difference, and some had high levels of usage, more than 125 hours a year. Usage differences are also evident between districts. For example, the difference between treatment and control teacher technology usage was relatively large in district 1 compared to district 9.

According to product records, actual student usage ranged from 5 to almost 30 hours (see Table V.3).[28] Overall, usage averaged 15 hours for the school year, which is about

[28]Average reported product usage according to teachers (47 hours a year) was larger than average usage from product records (15 hours), and teacher-reported usage and usage according to product records were moderately correlated ($r = 0.56$ for annual use). The degree of correlation suggests that teacher-reported usage is only a rough indicator of actual usage.

Figure V.1
Difference in Annual Teacher-Reported Hours of Math Technology Product Use Between Treatment and Control Classrooms, Algebra



Source: Teacher Interviews.

10 percent of math instructional time (assuming a 50-minute class period and a 180-day school year).[29]

Role of Products in the Curriculum and Time and Place of Use

Two products were intended to supplement the core curriculum, and the third was intended to be the core algebra curriculum. When treatment teachers were asked about the role of products in their algebra curriculum, 72 percent indicated they used the product as a supplement and 28 percent said they used the product as their core curriculum. For teachers using the product that was intended to be a core curriculum, 75 percent said they used it as such.

[29]Overall usage is weighted by student sample size for each product. The usage estimate includes only time using computers and does not include the noncomputer-based class activities that were a major part of the implementation model for one of the products.

Teachers varied in how they scheduled product use. Most teachers (94 percent) used products only during regular class time, and the other 6 percent used it both during class time and at other times. Products were used in computer labs by 84 percent of teachers, and 16 percent of teachers used products in regular classrooms. Consistent with vendor recommendations, nearly all treatment teachers (90 percent) reported being present when their students used products, and 62 percent reported reviewing student reports two to three times a month or more frequently.

Technical Difficulties

Study observers reported technical problems that affected at least one student in 40 percent of the time segments they observed. (Each observation included four to five time segments.) Most technical problems—such as problems logging on to products or having computers freeze and need rebooting—were brief and affected only a few students.

Satisfaction With Products

Nearly all treatment teachers (86 percent) said they would use the product in the next school year if given the choice. When asked what they would change in a second year of implementation, 24 percent said they would attend more closely to alignment of the software with the core curriculum and accountability system; 24 percent said they would start using the product sooner or use it more often.

Impact on Classroom Activities

During observation intervals (the time periods during which observers noted teacher roles), treatment teachers were observed acting as facilitators in a larger proportion of intervals than control teachers, 71 percent compared to 33 percent, $p < .05$ (Table V.4).[30] Students were also more likely to be engaged in individual practice, 88 percent compared to 34 percent ($p < .05$). In control classrooms, students were more likely to be listening to a teacher's lecture, participating in question-and-answer sessions, and reviewing their work with teachers. Students were more likely to be on task when using products, 74 percent compared to 65 percent ($p < .05$).

B. Effects on Algebra Test Scores

Effects on test scores were estimated using a model that accounted for the nesting of students in classrooms and classrooms in schools. The estimates show that effects on test scores generally were not statistically different from zero and most teacher and school characteristics were uncorrelated with effects.

[30]The differences were estimated using a two-level hierarchical linear model with teachers in the first level and schools in the second level.

Table V.4. Activities in Treatment and Control Classrooms.

	Treatment Classrooms	Control Classrooms	*p*-value[a]
Teacher Role (Percent)[b]			
Leader	13	58	.00
Facilitator	71	33	.00
Monitor/observer	18	16	.86
Working on other tasks	7	8	.77
Other	4	1	.08
Instructional Activity (Percent)[b]			
Individual practice	88	34	.00
Lecture	8	38	.00
Question and answer	3	16	.00
Review of student work	1	14	.00
Other	3	5	.04
Student On-Task Behavior (Percent)			
More than 90 percent of students on task	74	65	.02
Sample Size			
Number of classrooms	39	32	
Number of observation intervals	336	282	

Source: Classroom observation intervals at minutes 10, 20, and 30 for each class.

[a]Tests of mean difference conducted within four-level hierarchical models with the effect of treatment assignment modeled at level one and between-school variance modeled at level two.

[b]Observers coded "all that apply."

Treatment and Control Group Characteristics of Teachers and Students

Teacher and student characteristics were similar (Table V.5) for treatment and control conditions. The average score on the ETS algebra exam was 31.5 percent correct for students in treatment classrooms and 32.9 percent correct in control classrooms. Random guessing would yield a score of 25 percent on average (questions had four response categories). However, students were in the early stages of algebra instruction and were taking a final exam in algebra.[31]

[31]The ETS algebra exam consists of two 25-question tests combined in a 50-question test. For purposes of the study, ETS separated the test into its two components, and the study team randomly assigned schools to take one of the components as the fall test and the other as the spring test (half the schools took the test in the pattern A-B and the other half took the test in the pattern B-A). Statistical tests show some differences in average scores on the two components. However, these differences between the two halves of the test do not affect measured effects because the same difference is present for both treatment and control classrooms.

Effects Were Not Statistically Different From Zero

Table V.6 shows average score differences on the ETS algebra exam score and its three subtest scores (concepts, processes, and skills) from the hierarchical linear model. None of the differences is statistically different from zero.

Figures V.2 and V.3 depict the variation of school effect sizes by district and by product.[32] Within districts, effect sizes ranged widely. For example, in district 6, effect sizes in the four schools ranged from about -0.75 to 0.75.

Table V.5. Characteristics of Teachers and Students in Treatment and Control Classrooms.

	Treatment Classrooms	Control Classrooms	*p*-value[a]
Teacher Characteristics			
Years of experience (years)	12.4	10.3	0.32
Has a master's degree (percent)	54	53	0.91
School has computer specialist (percent)	68	72	0.63
Received professional development on using technology last year (percent)	35	34	0.94
Female (percent)	51	69	0.14
Teacher Sample Size	**37**	**32**	
Student Characteristics			
Female (percent)	52	47	0.25
Age as of October 2004 (years)	14.8	14.8	0.57
Unadjusted overall score on fall ETS algebra exam[b]	31.5	33.4	0.23
Sample Size	**774**	**630**	

Sources: Teacher questionnaires, student records, and tests administered by study staff.

[a]Tests of treatment and control differences were conducted using a two-level hierarchical model with classroom treatment status as a fixed effect and school as a random effect (for teachers) and classrooms and schools as random effects (for students). The *p*-value of the difference is the *p*-value of the estimated treatment coefficient.

[b]The ETS test score is the percent of questions answered correctly.

[32]School effects were estimated using a regression model in which test scores are regressed on student and teacher characteristics and the treatment indicator is interacted with an indicator for each school. Standard errors were adjusted for classroom clustering.

Table V.6. ETS Algebra Final Exam Score Differences in Treatment and Control Classrooms.

	Treatment Classroom Average Score[a]	Control Classroom Average Score[a]	Difference	Effect Size	*p*-value
ETS End-of-Course Algebra Exam					
Overall score	37.29	38.15	-0.86	-0.06	0.33
Subtest scores					
Concepts	35.80	37.50	-1.70	-0.10	0.07
Processes	35.07	36.16	-1.09	-0.06	0.27
Skills	41.12	40.73	0.39	0.02	0.79

Source: Author calculations; see Appendix B for details of the estimation model. Variables in the model include student age, gender, and pretest score; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity, percent of students in special education, and percent eligible for free lunch; and student, classroom, and school random effects.

The treatment classroom average score reported in the table is the control classroom average score plus the treatment effect. It differs from the unadjusted treatment classroom score.

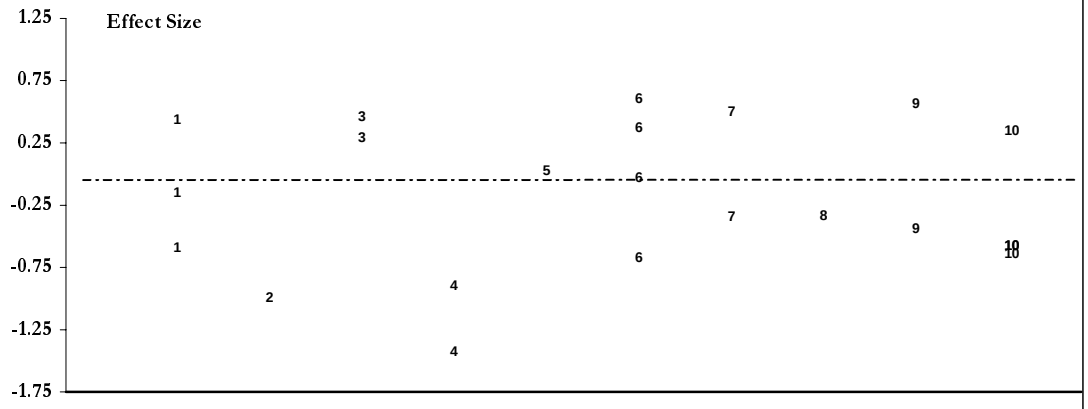
[a]The ETS test score is the percent of questions answered correctly.

Note: Effect sizes are calculated using the control group standard deviations on the spring test.

The spring final exam scores for students in the treatment and control groups (before model-based adjustment) averaged 35 percent for treatment classrooms and 38 percent for control classrooms. Fall scores averaged about 31 percent for both groups, indicating that after a school year of algebra instruction, the average student in the study was able to answer an additional one or two questions correctly (each question was worth four points). In 2003, information from ETS indicates that test takers averaged 46 percent correct, however. Test takers in the study were within a reasonable range of the average after accounting for poverty levels and urbanicity of schools in the study. Evaluations in districts with characteristics similar to districts in this study have found comparable spring ETS exam scores (Shneyderman 2001; Morgan and Ritter 2002). Whether gains were small in these studies is not known because these studies did not administer the test in the fall and spring. However, the similarity of the average spring score in previous studies with scores here suggests that scores in this study are not an aberration.

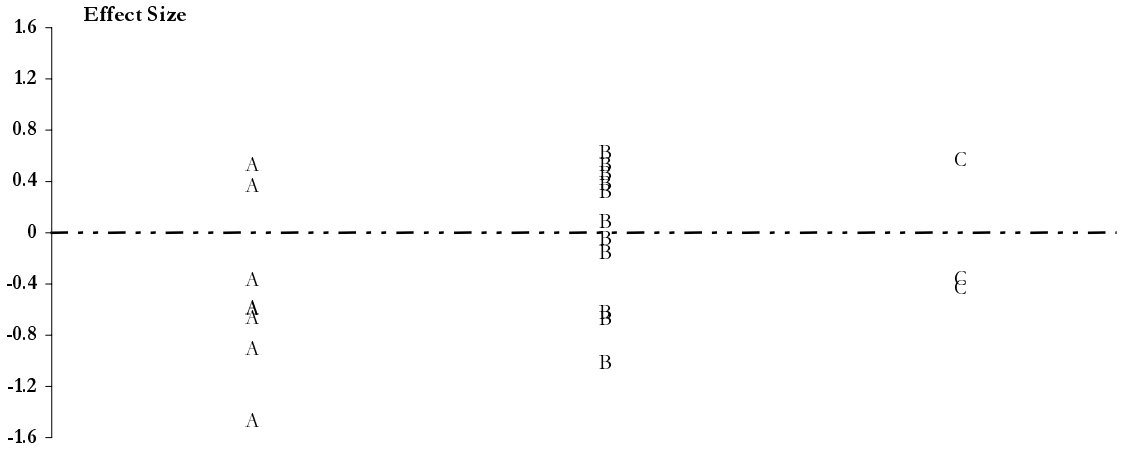
Another way to examine whether the ETS test functioned properly is to look at score differences on district assessments. In two districts where the study was able to gather district test score data from school records, an analysis of the district scores yielded similar directions of score differences and effect sizes were similar (see Appendix Table B.4). The similarities provide support for the ETS algebra test as a reasonable indication of algebra product effectiveness.

Figure V.2
School-Level Effect Sizes by District
Algebra (ETS Final Exam)



Numbers represent districts, and each point is a school.

Figure V.3
School-Level Effect Sizes by Product
Algebra (ETS Final Exam)



Letters represent products, and each point is a school.

Note: Statistical significance of average effect sizes cannot be inferred from the figure because student and teacher sample sizes differ between schools.

Classroom and School Moderators

Whether score differences were related to school and classroom characteristics was investigated using the same approach as in previous chapters. The smaller number of schools participating in the algebra study compared to the other three studies limited the number of variables that could be included in the model. Characteristics included were those that were related to effectiveness in the reading studies: teacher experience, whether teachers had adequate preparation time, time of product usage, and whether study products were used in classrooms. Two school characteristics—the proportion of students eligible for free lunch and the student-teacher ratio—were also included.

The results appear in Table V.7, and the statistical tests at the bottom of the table indicate that classroom and school characteristics were uncorrelated with product effects. Study product usage had a negative correlation with product effects but was not statistically significant (after adjusting for multiple comparisons).

C. Conclusions

This study experimentally tested the effects of select math technology products in algebra classrooms that volunteered to participate. Key findings include the following:

1. **Teachers Were Trained and Used Products.** Nearly all treatment teachers (98 percent) received training from the vendor on using the product, and all implemented the product to some degree.

2. **Effects on Test Scores Were Not Statistically Different From Zero.** Overall math scores for students in treatment and control classrooms were 37.3 percent correct and 38.1 percent correct, respectively. The difference was not statistically different from zero.

3. **Classroom and School Characteristics Were Uncorrelated With Product Effects.** The algebra study included fewer schools, which limited the ability to estimate moderator effects. None of the classroom and school characteristics included in the model was statistically significant.

Table V.7. Interactions Between Moderating Variables and Effects: ETS Algebra Test.

	Coefficient	Standard Error	*p*-value
Classroom Interaction Variables*			
Years of teaching experience	0.05	0.11	0.66
Study product use reported by teachers	-7.04	2.96	0.02
Other product use reported by teachers	-2.83	3.68	0.45
Students had problems getting access to product	-3.78	1.95	0.06
Teacher had adequate time to prepare to use product	-0.44	2.34	0.85
Teacher indicates school has a computer specialist	0.61	1.70	0.72
Product is used in classroom	-0.03	2.99	0.99
School Interaction Variables*			
Percent eligible for free lunch	5.03	4.48	0.27
Student-teacher ratio	0.20	0.46	0.66
		χ^2	*p*-value
Chi-Squared Tests‡			
Classroom and school variables		7.54	>.50
Classroom variables		6.50	0.37
School variables		0.70	>.50

Source: Author calculations; see Appendix B for details of the estimation model. Variables in the model include student age, gender, and the pretest scores; teacher gender, experience, and whether he or she had a master's degree; school race and ethnicity, percent of students in special education, and percent eligible for free lunch; time of use of other technology products; and student, classroom, and school random effects.

*No characteristics are significantly different from zero based on the Benjamini-Hochberg correction for multiple comparisons, with false discovery rate of 0.05. The multiple-comparisons correction is computed separately for classroom and school characteristics.

‡No Chi-squared test is statistically significant at the 0.05 level.

References

Adelman, N., M.B. Donnelly, T. Dove, J. Tiffany-Morales, A. Wayne, and A. Zucker. *Professional Development and Teachers' Uses of Technology.* Subtask 5: Evaluation of Key Factors Impacting the Effective Use of Technology in Schools. Report to the U.S. Department of Education. SRI Project P10474. Arlington, VA: SRI International, 2002.

Agodini, R., M. Dynarski, B. Means, R. Murphy, and L. Rosenberg. "Revised Design for the Evaluation of the Effectiveness of Educational Technology Interventions." Report prepared for Institute of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, Inc., September 2005.

Agodini, R., M. Dynarski, M. Honey, and D. Levin. "The Effectiveness of Educational Technology: Issues and Recommendations for the National Study." Report prepared for Institute of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, Inc., May 2003.

Anderson, R.E., and A. Ronnkvist. *The Presence of Computers in American Schools.* Report 2. Teaching, Learning, and Computing: 1998 National Survey. Irvine, CA: Center for Research on Information Technology and Organizations, 1999.

Bangert-Drowns, R., C. Kulik, J. Kulik, and M. Morgan. "The Instructional Effect of Feedback in Test-Like Events." *Review of Educational Research*, vol. 61, no. 2, summer 1991, pp. 213–238.

Becker, H.J. "Computer-Based Integrated Learning Systems in the Elementary and Middle Grades: A Critical Review and Synthesis of Evaluation Reports." *Journal of Educational Computing Research,* vol. 8, no. 1, 1992, pp. 1-41.

Becker, H.J., J. Ravitz, and Y. Wong. *Teacher and Teacher-Directed Student Use of Computers and Software.* Technical Report #3: Teaching, Learning, and Computation, 1998 National Survey. Irvine, CA: University of California at Irvine, 1999.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Stat. Soc. B., 57, 289-300.

Black, P., and D. Wiliam. "Assessment and Classroom Learning." *Assessment in Education,* vol. 5, no. 1, 1998, pp. 7-74.

Blok, H., R. Oostdam, M.E. Otter, and M. Overmaat. "Computer-Assisted Instruction in Support of Beginning Reading Instruction: A Review." *Review of Educational Research,* vol. 72, 2002, pp. 101-30.

Blumenfeld, P., B. Fishman, J. Krajcik, R.W. Marx, and E. Soloway. "Creating Usable Innovations in Systemic Reform: Scaling Up Technology-Embedded Project-Based Science in Urban Schools." *Educational Psychologist,* vol. 35, no. 3, 2000, pp. 149-164.

Bransford, J.D., A.L. Brown, and R.R. Cocking. "*How People Learn: Brain, Mind, and Experience.* Washington, DC: National Academy Press, 1999.

Buros Institute of Mental Measurements, *Test Reviews Online.* Lincoln, NE: University of Nebraska-Lincoln, 1998.

Butler, R. "Task-Involving and Ego-Involving Properties of Evaluation: Effects of Different Conditions on Motivational Perceptions, Interest, and Performance." *Journal of Educational Psychology*, vol. 79, no. 4, 1987, pp. 474-482.

Cohen, D.K., and D.L. Ball. *Instruction, Capacity, and Improvement.* CPRE Research Report No. RR-043. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education, 1999.

Crooks, T.J. "The Impact of Classroom Evaluation Practices on Students." *Review of Educational Research*, vol. 58, no. 4, winter, 1988, pp. 438-481.

Cuban, L. *Oversold and Underused: Computers in the Classroom*. Cambridge, MA: Harvard University Press, 2001.

Cuban, L. *So Much High-Tech Money Invested, So Little Use and Change in Practice: How Come?* Report to the CCSSO State Educational Technology Leadership Conference – 2000: Preparing Teachers to Meet the Challenge of New Standards with New Technologies. Washington, DC: Council of Chief State School Officers, 2000.

Culp, K.M., M. Honey, and E. Mandinach. *A Retrospective on Twenty Years of Education Technology Policy.* Washington, DC: U.S. Department of Education, Office of Educational Technology, 2003. Accessed August 2, 2005, at [www.nationaledtechplan.org /participate/20years.pdf].

Educational Testing Service. *End-of-Course Algebra Assessment Administrator's Manual.* Princeton, NJ: ETS, 1997.

Ertmer, P.A. "Addressing First- and Second-Order Barriers to Change: Strategies for Technology Integration." *Educational Technology Research and Development,* vol. 47, no. 4, 1999, pp. 47-61.

Good, T.L., and J. Brophy. *Looking in Classrooms.* Boston, MA: Allyn and Bacon, 2003.

Hansen, E.E., L.L. Llosa, and J. Slayton. *Evaluation of the Waterford Early Reading Program as a Supplementary Program in the Los Angeles Unified School District: 2002-03*. Planning, Assessment and Research Division Publication No. 177. Los Angeles: Los Angeles Unified School District, Program Evaluation and Research Branch, 2004.

Henriquez, A., and M. Riconscente. *Rhode Island Teachers and Technology Initiative: Program Evaluation Final Report*. New York: Education Development Corporation, Center for Children and Technology, 1999.

Honey, M., and A. Henriquez. "Union City Interactive Multimedia Education Trial: 1993-1995, Summary Report." New York: Center for Children and Technology, Education Development Center, 1996. Accessed March 1, 2006, at [www2.edc.org/CCT /publications_report_summary.asp?numPubId=88].

Joyner, A. "A Foothold for Handhelds." *American School Board Journal: Special Report*. September 2002. Accessed October 20, 2005, at [www.asbj.com/specialreports /0903SpecialReports/S3.html].

Kerr, S.T. "Technology and the Future of Schooling." In *95th Yearbook of the National Society for the Study of Education,* Part II, edited by S.T. Kerr. Chicago: University of Chicago Press, 1996.

Kluger, A.N., and A. DeNisi. "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory." *Psychological Bulletin*, vol. 119, no. 2, 1996, pp. 254-284.

Kulik, C.C., and J.A. Kulik. "Effectiveness of Computer-Based Instruction: An Updated Analysis. *Computers in Human Behavior,* vol. 7, 1991, pp. 75-94.

Kulik, J.A. *Effects of Using Instructional Technology in Elementary and Secondary Schools: What Controlled Evaluation Studies Say*. Arlington, VA: SRI International, 2003.

Kulik, J.A. "Meta-analytic Studies of Findings on Computer-Based Instruction." In *Technology Assessment in Education and Training*, edited by E.L. Baker and H.F. O'Neil Jr. Hillsdale, NJ: Lawrence Erlbaum, 1994, pp. 9-33.

Leggett, W.P., and K.A. Persichitte. "Blood, Sweat, and Tears: 50 Years of Technology Implementation Obstacles. *TechTrends,* vol. 43, no. 3, 1998, pp. 33-36.

Linn, M.C., and S. Hsi. *Computers, Teachers, Peers: Science Learning Partners*. Mahwah, NJ: Erlbaum, 2000.

Mann, D., C. Shakeshaft, J. Becker, and R. Kottkamp. *West Virginia Story: Achievement Gains from a Statewide Comprehensive Instructional Technology Program*. Santa Monica, CA: Milken Exchange on Educational Technology, 1998.

Means, B., and K. Olson. *Technology and Education Reform: Technical Research Report*. Menlo Park, CA: SRI International, 1995.

Mezernich, M.M., W.M. Jenkins, P. Johnston, C. Schreiner, S.L. Miller, and P. Tallal. "Temporal Processing Deficits of Language-Learning Impaired Ameliorated by Training." *Science,* vol. 271, no. 5, 1996, pp. 77-81.

Miller, S., N. Linn, P. Tallal, M. Merzenich, and W. Jenkins. "Speech and Language Therapy (Reeducation Orthophonique)." *Federation Nationale des Orthophonistes, Special Issues: La Conscience Phonologique*, vol. 197, March 1999, pp. 159-182.

Mills, S.C., and T.J. Ragan. "A Tool for Analyzing Implementation Fidelity of an Integrated Learning System (ILS)." *Educational Technology Research & Development,* vol. 48, 2000, pp. 21-41.

Morgan, P. and S. Ritter. "An Experimental Study of the Effects of Cognitive Tutor Algebra I on Student Knowledge and Attitude." Pittsburgh, PA: Carnegie Learning, Inc., May 16, 2002.

Murphy, R., W. Penuel, B. Means, C. Korbak, and A. Whaley. *E-DESK: A Review of Recent Evidence on the Effectiveness of Discrete Educational Software.* Menlo Park, CA: SRI International, 2001.

Nunnery, J.A., S.M. Ross, and A. McDonald. "A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6." *Journal of Education for Students Placed at Risk,* vol. 11, no. 1, 2006, pp. 1-18.

Pearson, P.D., R.E. Ferdig, R.L. Blomeyer Jr., and J. Moran. *The Effects of Technology on Reading Performance in the Middle-School Grades: A Meta-analysis with Recommendations for Policy.* Naperville, IL: Learning Point Associates, 2005.

President's Committee of Advisors on Science and Technology [PCAST], Panel on Educational Technology Report to the President on the Use of Technology to Strengthen K-12 Education in the United States, March 1997. Accessed March 14, 2006, at [www.ostp.gov/PCAST/k-12ed.html].

Rouse, C.E., and A.B. Krueger. "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically-Based' Reading Program." *Economics of Education Review,* vol. 23, no. 4, 2004, pp. 323-338.

Sandholtz, J.H., C. Ringstaff, and D. Dwyer. T*eaching With Technology: Creating Student-Centered Classrooms.* New York: Teachers College Press, 1997.

Sarama, J., D.H. Clements, and J.J. Henry. "Network of Influences in an Implementation of a Mathematics Curriculum Innovation." *International Journal of Computers for Mathematical Learning,* vol. 3, no. 2, 1998, pp. 113-48.

Schacter, J. *The Impact of Educational Technology on Student Achievement: What the Most Current Research Has to Say.* Santa Monica, CA: Milken Exchange on Education Technology, 2001.

Schofield, J.W., and D. Verban. "Computer Usage in Teaching Mathematics: Issues Which Need Answers." In *Effective Mathematics Teaching,* vol. I, edited by D.A. Grouws and T.J. Cooney. Hillsdale, NJ: Erlbaum, 1988, pp. 169-193.

Sheingold, K., and M. Hadley. *Accomplished Teachers: Integrating Computers into Classroom Practice.* New York: Center for Technology in Education, Bank Street College of Education, 1990.

Shneyderman, A. *Evaluation of the Cognitive Tutor Algebra I Program.* Miami-Dade County Public Schools, Office of Evaluation and Research, September 2001.

Sivin-Kachala, J. *Report on the Effectiveness of Technology in Schools, 1990-1997.* Washington, DC: Software Publishers Association, 1998.

Swan, K., and M. Mitrani. "The Changing Nature of Teaching and Learning in Computer-Based Classrooms. *Journal of Research on Computing in Education,* vol. 26, 1993, pp. 40-54.

Swan, K., M. van 't Hooft, A. Kratcoski, and D. Unger. "Uses and Effects of Mobile Computing Devices in K-8 Classrooms." *Journal of Research on Technology in Education,* vol. 38, no. 1, 2005, pp. 99-112.

Tallal, P., S.L. Miller, G. Bedi, G. Byma, X. Wang, S.S. Nagarajan, C. Schreiner, W.M. Jenkins, and M.M. Mezenich. "Language Comprehension in Language-Learning Impaired Children Improved with Acoustically Modified Speech." *Science,* vol. 271, no. 5, 1996, pp. 81-84.

Torgesen, J., R. Wagner, and C. Rashotte. *Test of Word Reading Efficiency: Examiner's Manual.* Austin, TX: Pro-Ed, Inc., 1999.

U.S. Congress, Office of Technology Assessment. *Teachers and Technology: Making the Connection.* OTA-HER-616. Washington, DC: U.S. Government Printing Office, 1995.

VanDusen, L.M., and B.R. Worthen. "The Impact of Integrated Learning Systems Implementation on Student Outcomes: Implications for Research and Evaluation." *International Journal of Educational Research,* vol. 51, 1994, pp. 13-24.

Waxman, H.C., and S.L. Huang. "Differences by Level of Technology Use on Students' Motivation, Anxiety, and Classroom Learning Environment in Mathematics." *Journal of Educational Technology Systems,* vol. 25, no. 1, 1996, pp. 67-77.

Waxman, H.C., M-F Lin, and G.M. Michko. *A Meta-analysis of the Effectiveness of Teaching and Learning with Technology on Student Outcomes.* Naperville, IL: Learning Point Associates, 2003.

Worthen, B.R., L.M. Van Dusen, and P.J. Sailor. "A Comparative Study of the Impact of Integrated Learning Systems on Students' Time-on-Task." *International Journal of Educational Research,* vol. 21, 1994, pp. 25-37.

Appendix A

Data Collection Approach and Response Rates

This appendix describes the study's approach for data collection and classroom observations. It also provides more detail about response rates and issues with some of the data.

The study's data collection began when teachers were randomly assigned, which generated the initial listing of teachers and ultimately the listing of students to test and classrooms to observe. Not all students in teacher classrooms were tested. The two criteria for testing students in the fall were that parental consent was received and that students did not have barriers to testing (disability or language issues). The spring test included students who had been tested in the fall as well as students who had entered study classrooms after the fall test was administered. The study attempted to collect records for all students in the spring testing sample.

A. Teacher Sample

The study recruited 134 schools in 34 school districts. One district dropped out before the start of the school year (after teacher random assignment had been conducted), and another district dropped out after the fall test had been conducted because the technology product did not work correctly.

Figure A.1 shows the flow of teachers into the study. Of the 531 teacher consent forms received, five teachers were excluded before random assignment because their schools had not recruited an adequate number of teachers (a minimum of two teachers in each grade was needed for random assignment). The remaining 526 teachers were randomly assigned. Odds of assignment to use a product varied depending on how many teachers in a school consented. In schools with an even number of teachers, the odds were 50-50. In schools with an odd number of teachers consenting, the odds favored the treatment group by one (two to one if three teachers consented, three to two if five teachers consented, and four to three if seven teachers consented; few schools had more than seven teachers consenting).

After initial random assignment had been conducted, 98 teachers were excluded from the study for various reasons, which is shown in Figure A.1. The most common reason (applying to 47 teachers) was that teachers received new teaching assignments or left their schools for other positions. Another 30 teachers were excluded because their districts decided not to participate. Attrition also sometimes resulted in fewer than two sampled teachers in a school, and these schools were dropped (12 teachers). Four teachers declined to participate after random assignment was conducted (three in the treatment group and one in the control group). As the figure shows, treatment and control status had relatively little effect on teachers being excluded from the study. Table A.1 shows final counts of teachers in the sample by grade level and assignment status.

Figure A.1. Flow of Teachers Through Study.

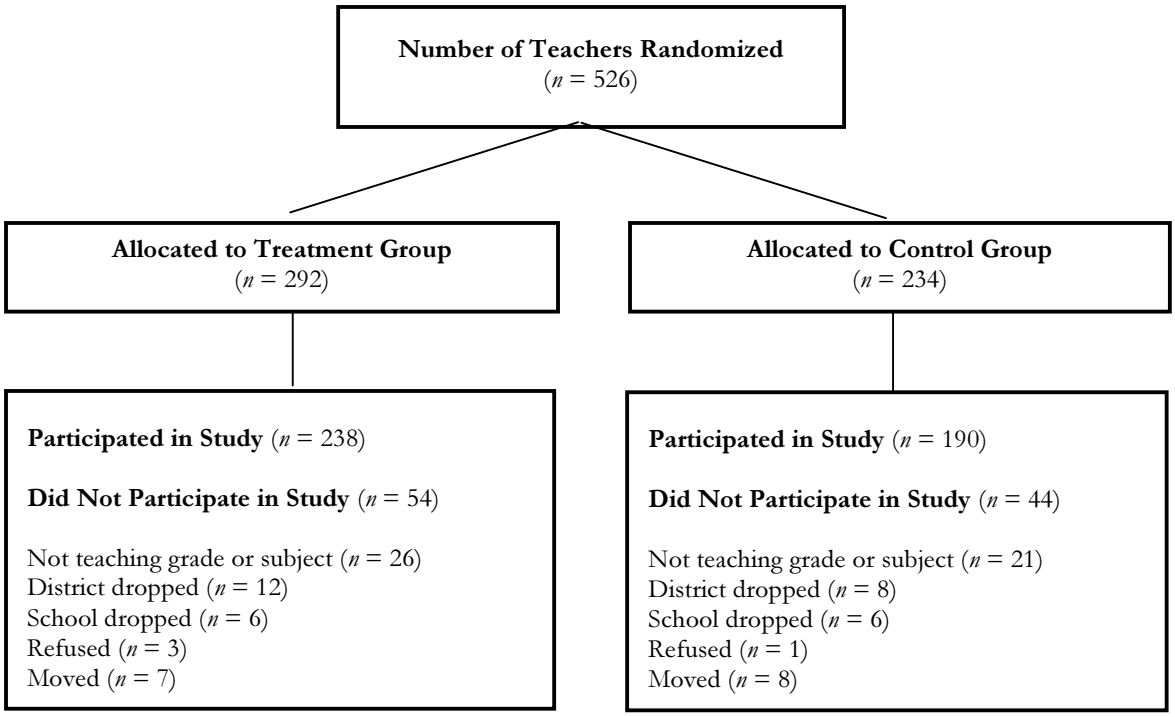


Table A.1. Number of Teachers Participating by Grade Level and Random Assignment Status, 2004-2005 School Year.

	Number of Teachers		
	Total	Treatment	Control
Total	428	238	190
Grade 1 reading	158	89	69
Grade 4 reading	118	63	55
Grade 6 mathematics	81	47	34
Algebra 1	71	39	32

B. Teacher Survey

In November 2004, teacher questionnaires were mailed to schools, and a second mailing was sent in December to teachers at their home addresses. Additional prompts to complete the questionnaire included e-mails and telephone messages, additional mailings as needed, and telephone interviews through March 2005. Ultimately, 94 percent of teachers completed a questionnaire. Completion rates ranged from 91 percent of fourth grade teachers to 96 percent of first grade teachers (Table A.2).

Table A.2. Total Number of Teachers in the Study and the Number and Percent Completing the Teacher Survey, 2004-2005 School Year.

	Teachers		
	Total	Number Completing Survey	Percent
Total	428	401	94
Grade 1 reading	158	152	96
Grade 4 reading	118	107	91
Grade 6 mathematics	81	76	94
Algebra 1	71	66	93

C. Student Data Collection

Students were tested as early or late as was feasible in the school year. Reading achievement tests were administered in grades 1 and 4, and math tests in grade 6 and in algebra. Data from students' school records were also collected.

Obtaining Class Lists and Parent Consent

At the beginning of the school year, class lists for each teacher in the study were obtained from schools and were the basis for letters to parents requesting consent. The letter explained the purpose of the study and described all data collection activities involving students. Letters (which were translated and available in eight languages) were sent to the parents via mail or were sent home with students, depending upon the preference of the school. A brochure with answers to frequently asked questions was also included in the parent permission packets along with a toll-free number for parents who had additional questions. Of the 12,700 students on teachers' fall or spring semester classroom lists, 3,789 students attended schools requiring active consent and 8,911 students attended schools requiring passive consent (Table A.3).

Parent consent was obtained for 93 percent of students, 80 percent in active consent districts and 99 percent in passive consent districts. Treatment status may have had some effect on consent rates: Consent was obtained for 95 percent of students in treatment classrooms and 91 percent of students in control classrooms (Table A.4). The difference arose from control classrooms for which active consent was needed. In treatment

Table A.3. Number and Percent of Eligible Students by Type of Consent, 2004-2005 School Year.

All Students			Students in Passive Consent Sites			Students in Active Consent Sites		
	With Consent			With Consent			With Consent	
Total	Number	Percent	Total	Number	Percent	Total	Number	Percent
12,700	11,869	93	8,911	8,839	99	3,789	3,030	80

Table A.4. Eligible Students in Study Classrooms with Parental Consent to Participate in the Study, 2004-2005 School Year.

	All Eligible Students			Treatment Students			Control Students		
		With Consent			With Consent			With Consent	
Technology Cluster	All	Number	Percent	All	Number	Percent	All	Number	Percent
Total	12,700	11,869	93	7,133	6,791	95	5,567	5,078	91
Grade 1 early reading	3,170	3,111	98	1,791	1,783	99	1,379	1,328	96
Grade 4 reading	2,926	2,766	95	1,587	1,524	96	1,339	1,242	93
Grade 6 mathematics	3,919	3,725	95	2,341	2,271	97	1,578	1,454	92
Algebra 1	2,685	2,267	84	1,414	1,213	86	1,271	1,054	83

classrooms, 86 percent of parents gave consent, and in control classrooms, 71 percent of parents gave consent.

Student Sample

The study received parent consent for 11,869 students. Table A.5 shows students by classroom assignment status, with 57 percent of students in treatment classrooms and 43 percent in control classrooms. These proportions match closely the allocations of teachers shown in Table A.1.

Table A.5. Number and Percent of Eligible Student Sample by Assignment Group and Grade, 2004-2005 School Year.

	Eligible Student Sample	In Treatment Classrooms		In Control Classrooms	
	All	Number	Percent	Number	Percent
Total	11,869	6,791	57	5,078	43
Grade 1 reading	3,111	1,783	57	1,328	43
Grade 4 reading	2,766	1,524	55	1,242	45
Grade 6 mathematics	3,725	2,271	61	1,454	39
Algebra 1	2,267	1,213	54	1,054	46

Student Tests

Student tests were administered during regular class periods in the fall and spring. In addition, the Test of Word Reading Efficiency (TOWRE), an individual assessment of reading fluency, was administered to first graders in the fall and spring at about the same time as the SAT-9 was administered. Figure A.2 lists the test and subtests that the study administered.

The study tested 10,683 students in fall 2004 (Table A.6). The total represents about 95 percent of students who were on class rosters at that time, who did not have disabilities or language problems precluding testing, and whose parents had given consent. Response rates ranged from 85 percent in algebra classrooms to 99 percent in sixth grade classrooms.

Figure A.2. Achievement Tests Administered by the Study

	Fall 2004 Test	**Spring 2005 Test**
Grade 1	Stanford Early School Achievement Test (SESAT 2, Form S), Fourth Edition Test of Word Reading Efficiency	Stanford Achievement Test, Abbreviated Primary 1, Ninth Edition (Form S) Test of Word Reading Efficiency
Grade 4	Stanford Achievement Test Abbreviated Battery Primary 3, Tenth Edition	Stanford Achievement Test Abbreviated Battery Intermediate 1, Tenth Edition
Grade 6	Stanford Achievement Test Abbreviated Battery Intermediate 2, Tenth Edition	Stanford Achievement Test Abbreviated Battery Intermediate 3, Tenth Edition
Algebra 1	Educational Testing Services' End-of-Course Algebra Test	Educational Testing Services' End-of-Course Algebra Test

Table A.6. Number of Students and Percent Tested in Fall and Spring, 2004-2005 School Year.

	Eligible Students In Treatment Classrooms	Eligible Students In Control Classrooms	Tested Eligible Students In Treatment Classrooms	Tested Eligible Students In Control Classrooms	Response Rate, Treatment Classrooms	Response Rate, Control Classrooms
First Grade SESAT						
Fall	1,712	1,266	1,660	1,218	97%	96%
Spring	1,707	1,266	1,623	1,197	95%	95%
First Grade TOWRE						
Fall	1,712	1,266	1,651	1,218	96%	96%
Spring	1,707	1,266	1,617	1,204	95%	95%
Fourth Grade SAT-10						
Fall	2,149	1,402	2,109	1,365	98%	97%
Spring	1,450	1,206	1,316	1,110	91%	92%
Sixth Grade SAT-10						
Fall	2,149	1,402	2,109	1,365	98%	97%
Spring	2,217	1,428	2,022	1,330	91%	93%
ETS Algebra Test						
Fall	1,139	1,000	963	828	85%	83%
Spring	1,155	1,000	958	811	83%	81%
Total						
Fall	6,449	4,844	6,134	4,549	95%	94%
Spring	6,529	4,900	5,919	4,448	91%	91%

The spring test was administered about 4 to 6 weeks before the last day of the school year. At follow-up, 10,367 students were tested. The response rate of 87 percent ranged from 78 percent in algebra to 90 percent in first and sixth grades for standardized tests. Ninety-five percent of eligible first graders took the TOWRE in the spring.

Students who were tested in both fall 2004 and in spring 2005 are the basis of the estimates presented in the main report (see Table A.7). Eighty-five percent of eligible students (who were enrolled and had parent consent) took both the fall and spring test, ranging from 90 percent for first graders to 67 percent for algebra students.

Table A.7 Number of Students Tested In Both Fall and Spring.

	Number	Percent
All	9,458	85
First Grade SAT	2,653	90
First Grade TOWRE	2,644	89
Fourth Grade SAT	2,265	87
Sixth Grade SAT	3,136	89
Algebra ETS	1,404	67

Imputing Missing Data

Some data were missing because students did not take all tests or subtests, students did not take the TOWRE after taking the SESAT (or vice versa), or school districts did not provide records. The largest number of missing tests and subtests occurred in first grade. In some districts, the first grade test was administered on two adjoining days because of its length, and students being absent on either day affected whether some parts of the test were missing.

Components of the test scores and student age and gender were imputed using the Markov Chain Monte Carlo (MCMC) method in SAS 9. The imputation was done five times separately for students in treatment and control classrooms. The method was tested by making subtest scores for random samples of students "missing" and examining correlations between imputed scores and actual scores. The correlations were high, in the range of 90 percent to 95 percent for different samples, indicating that the MCMC method was imputing scores that were close to the actual scores. The HLM estimation and other statistical procedures used the five imputed data sets and calculated variances of the estimates that incorporated the added variance from the imputation. The increase in variances associated with the imputation ranged from less than 1 percent for many items to 9 percent for age of first grade students. There were fewer missing data items in fourth and sixth grades, and the corresponding increases in variances from imputation were smaller. No missing data were imputed for algebra.

School Records

The study collected student records data from files prepared by districts or by abstracting data from paper files. Records data collection began in late spring 2005. The data items on the records form included attendance, limited English proficiency, sex, age, eligibility for free- or reduced-price lunch program, disabilities, special education plans or services, grade promotion and retention, grades, and standardized test scores administered during the 2003-2004 and 2004-2005 school years. The study obtained data from student records for 93 percent of students who were eligible for the fall or spring testing (not shown in tables). Response rates ranged from 90 percent for first graders to 97 percent for sixth graders.

Classroom Observations

The study team used implementation models, previous research, and other instruments to design a classroom observation protocol. The draft was tested with selected products in spring 2004, and videotaped examples of product use were gathered and later used during training of observers.

To ensure the validity and reliability of the observation protocol, a coding guide was developed along with supplementary materials to help orient researchers to the products during training. The guide included definitions and examples of each code and also specified when observations were to be made. The training focused on preparing observers to code responses accurately and reliably. Direct instruction in the use of the protocols was provided and anchored with video examples. Observers watched an anchor video of a real classroom using software and discussed how they would code the interaction depicted in the video. Additional videotape segments functioned as reliability tests for observers. Observers with agreement scores of less than 80 percent with expert coders were excluded from the observer pool. (Three trainees were excluded on this basis.)

Observation of classrooms occurred at three points during the school year and focused on a class period during which products were being used (treatment classrooms) or the equivalent class period (control classrooms). The observation protocol called for observational data to be recorded at five points during a 50-minute period. The protocol had three sections: student activities, product use, and general observations. Observers recorded the number of student activities taking place, the type of activity, the teacher's role, the percent of students who were off task, and the number of students using the product. Observers coded data about product use for a 10-minute period, including teacher roles in motivating and helping students who were using products and whether teachers or students had technical difficulties with hardware or software. The general observation section covered the entire class period, and observers recorded information about the amount of time students used the product (if it was used), where the product was used, and whether other products were used.

To assess reliability of the observation protocol in the field, in a sample of districts where there were two or more observers, researchers conducted at least one reliability observation with two observers coding the same time segment in the same classroom. A total of 68 reliability observations were completed across the three site visits. The observation codes from the observer pairs were then used to assess agreement rates. The overall agreement rate (perfect agreement) across observation variables was 78 percent.

The teacher interview protocols were designed to gather in-depth data on product implementation and contexts of use. The core elements included time spent on curriculum and instruction, access and frequency of use, use as the core curriculum or as a supplement, frequency and duration of use, grouping strategies for the product, managing instruction with technology, courseware, teacher training and support, and satisfaction with the use of the software. Separate protocols were developed for math and reading, for treatment and control conditions, and for each school visit and time of year. The content of the interviews was derived from the study's conceptual framework based on prior research in the area of technology implementation and from vendor implementation models. The conceptual

framework used the model of instruction developed by Cohen and Ball (1999), elaborating that model with implementation issues from the educational technology literature, such as ongoing supports for teacher use of technology, location and amount of technology use, classroom management, and technology difficulties experienced. The vendor implementation models were obtained through review of vendor literature and from information provided in vendor presentations to the research team, supplemented with follow-up questions as needed to obtain a full set of implementation recommendations for every product.

In the interview protocol, each question item was followed by a list of possible teacher responses. For some items, an "other" category was included with space to record responses that were not one of the response options. After each response, interviewers were instructed to code the response that best matched the teacher's response. In the event that an interviewer was not clear how a response should be coded, interviewers were instructed to ask the teacher which of the response categories best fit the teacher's experience. Separate protocols were developed for math and reading, for treatment and control conditions, and for each school visit and time of year. The duration of the interview was typically 40 minutes for treatment teachers and 10-20 minutes for control teachers, depending on whether the control teacher was using a software product. Draft interview protocols were piloted in spring 2004 with four different teachers, and observers were trained using the protocols in summer 2004.

Processing of Observation and Interview Data

Monitoring and processing of observation data were supported by a web-based data collection monitoring system and an electronic scanning system for data entry. Observation and interview forms had a pre-printed label with an identification bar code that identified the site visitor, type of form, school, classroom, and place in the data collection sequence (first, second, or third visit). Forms were inspected for completeness and cleaned to ensure accurate reading by the electronic scanner. After forms had passed the initial quality control check, they were scanned into the monitoring system. Data entry for the forms and coded response items for teacher interviews was performed using an electronic scanning and archiving system. Forms that had been cleaned and approved for scanning were scanned into the electronic scanning system, and data records were formatted for importing to an analysis file. To verify that the error rate of the scanning process was less than 1 percent, each week during the data collection windows, 15 percent of forms were randomly selected and 100 percent verified by comparing entered values with the original hardcopy forms.

Observation and Interview Constructs

One item was constructed from the observation data, and two were constructed from teacher interview data for analysis purposes. The "Location of Use" variable was coded as 1 for use in a regular classroom (and 0 for use in a lab, library, or other site). The "Difficulties with Access to Product" variable was created to measure the extent to which teachers experienced difficulties gaining access to the product or finding enough time for students to use it as planned (because of technical problems with the school network, students being pulled out of class to receive special services, scheduling conflicts with use of computer lab,

or other reasons). Teachers were asked whether they experienced difficulties "getting students enough time or access on the assigned product." Responses were coded as "yes" or "no," and teachers were also asked to describe the nature of the difficulty. The scale used averaged responses from the first two visits, resulting in a scale with a range from 0 (no difficulties experienced prior to either visit) to 1 (difficulties experienced prior to each visit), which were multiplied by 100 to make the value analogous to a percentage. The "Adequacy of Supported Time" variable was created from an interview item about whether teachers felt there was enough "supported," or paid, time to prepare to use the product (0 = "supported time not enough" to 1 = "supported time was adequate"). The questions were asked during the third visit, after teachers already had opportunities to assess how much time they needed to prepare to use products.

Descriptive Data

Tables A.8a-d show means and standard deviations for all data items used in the estimation models. Both fall and spring overall scores and subtest scores are included (standard deviations of the control group spring test score are the basis of effect size calculations in the text). Some data items are defined for only treatment classrooms, and school characteristics are the same for treatment and control classrooms.

TABLE A.8(a)

Descriptive Statistics of Variables Used in Estimating Effects: First Grade
(Means and Standard Deviations)

Descriptive	Treatment Group	Control Group
Students		
Female (percent)	48.70 (49.07)	48.74 (48.51)
Age as of October 2004 (years)	6.63 (0.39)	6.67 (0.44)
Fall score on Stanford Early School Achievement Test (NCE)		
Sounds and letters	50.66 (19.94)	51.17 (20.15)
Sentence reading	49.94 (19.73)	51.30 (20.67)
Word reading	49.97 (19.55)	50.29 (20.60)
Total	50.10 (20.30)	50.91 (21.19)
Spring score on Stanford Achievement Test-9 (NCE)		
Sounds and letters	50.37 (19.83)	49.47 (19.78)
Sentence reading	49.61 (19.80)	50.34 (20.49)
Word reading	50.74 (19.90)	50.08 (20.68)
Total	49.77 (20.37)	49.47 (20.71)
Fall score on Test of Word Reading Efficiency		
Phonemic decoding	109.33 (8.27)	109.86 (8.31)
Sight word	105.90 (9.86)	106.43 (10.11)
Total word reading	109.13 (10.42)	109.75 (10.59)

Table A.8(a)

Descriptive	Treatment Group	Control Group
Spring score on Test of Word Reading Efficiency		
Phonemic decoding efficiency	109.32 (10.78)	109.53 (11.11)
Sight word efficiency	109.92 (11.09)	110.18 (11.79)
Total word reading efficiency	111.53 (12.53)	111.83 (13.19)
Teachers		
Female (percent)	98.88 (10.60)	97.10 (16.90)
Teaching experience (years)	11.65 (9.34)	11.59 (8.91)
Master's degree (percent)	38.20 (48.86)	55.07 (50.11)
Time spent using other products (hours a year)	4.79 (6.95)	9.07 (12.75)
Time spent using study products (hours a year)	47.74 (27.00)	n.a.
Study product used in the classroom (percent)	82.02 (38.62)	n.a.
Access problems (scale from 0 to 100)	53.37 (40.45)	n.a.
Adequate amount of preparation support time (percent)	67.42 (47.13)	n.a.
School has on-site computer specialist (percent)	74.16 (44.03)	n.a.
Participated in technology professional development last school year (percent)	52.81 (50.20)	n.a.
Schools		
Students receiving free/reduced-price lunch (percent)	48.61 (28.86)	
Student-teacher ratio	15.62 (2.09)	
Black students (percent)	29.68 (30.99)	

Table A.8(a)

Descriptive	
Hispanic students (percent)	22.67 (24.92)
Has IEP (percent)	9.88 (13.87)
In urban area (percent)	45.24 (50.38)

Table A.8(b)

Descriptive Statistics of Variables Used in Estimating Effects: Fourth Grade
(Means and Standard Deviations)

Descriptive	Treatment Group	Control Group
Students		
Female (percent)	48.23 (49.52)	51.58 (49.43)
Age as of October 2004 (years)	9.74 (0.63)	9.73 (0.57)
Fall score on Stanford Achievement Test (NCE)		
Reading comprehension	42.32 (21.00)	42.78 (19.37)
Reading vocabulary	40.78 (17.32)	40.51 (16.30)
Work study skills	42.19 (18.24)	43.07 (17.32)
Total	40.26 (18.47)	40.55 (16.63)
Spring score on Stanford Achievement Test (NCE)		
Reading comprehension	40.99 (20.04)	40.71 (18.51)
Reading vocabulary	43.12 (17.89)	43.07 (16.77)
Work study skills	42.83 (22.54)	43.15 (21.18)
Total	41.76 (19.38)	41.68 (17.65)
Teachers		
Female (percent)	80.95 (39.58)	89.09 (31.46)
Teaching experience (years)	9.19 (8.60)	10.10 (9.62)
Master's degree (percent)	26.98 (44.74)	30.91 (46.64)
Time spent using other products (hours a year)	6.13 (11.93)	6.97 (10.30)

Table A.8(b)

Descriptive	Treatment Group	Control Group
Time using study products (hours a year)	38.97 (13.75)	n.a.
Study product used in the classroom (percent)	71.43 (45.54)	n.a.
Access problems (scale from 0 to 100)	58.73 (38.67)	n.a.
Adequate amount of preparation support time (percent)	63.49 (48.53)	n.a.
School has on-site computer specialist (percent)	77.78 (41.91)	n.a.
Participated in technology professional development last school year (percent)	52.38 (50.34)	n.a.
Schools		
Students receiving free/reduced-price lunch (percent)	63.06 (20.00)	
Student-teacher ratio	16.59 (3.98)	
Black students (percent)	55.89 (38.67)	
Hispanic students (percent)	23.30 (30.06)	
Has IEP (percent)	12.06 (11.64)	
In urban area (percent)	55.81 (50.25)	

TABLE A.8(c)

Descriptive Statistics of Variables Used in Estimating Effects: Sixth Grade
(Means and Standard Deviations)

Descriptive	Treatment Group	Control Group
Students		
Female (percent)	51.11 (49.52)	51.95 (48.73)
Age as of October 2004 (years)	11.62 (0.48)	11.68 (0.53)
Fall score on Stanford Achievement Test (NCE)		
Procedures	48.53 (20.87)	50.44 (21.62)
Problem solving	48.60 (22.15)	50.25 (23.05)
Total	48.29 (20.80)	50.14 (22.03)
Spring score on Stanford Achievement Test (NCE)		
Procedures	51.46 (20.26)	50.30 (19.98)
Problem solving	51.75 (20.21)	51.08 (20.03)
Total	51.77 (20.08)	50.77 (19.86)
Teachers		
Female (percent)	63.83 (48.57)	82.35 (38.70)
Teaching experience (years)	10.33 (8.59)	11.07 (9.73)
Master's degree (percent)	29.79 (46.23)	35.29 (48.51)
Time spent using other products (hours a year)	0.51 (2.57)	2.44 (7.04)
Time spent using study products (hours a year)	50.59 (31.84)	n.a.
Study product used in the classroom (percent)	53.19 (50.44)	n.a.

Table A.8(c)

Descriptive	Treatment Group	Control Group
Access problems (scale from 0 to 100)	52.13 (41.65)	n.a.
Adequate amount of preparation support time (percent)	55.32 (50.25)	n.a.
School has on-site computer specialist (percent)	91.49 (28.21)	n.a.
Participated in technology professional development last school year (percent)	38.30 (49.14)	n.a.
Schools		
Students receiving free/reduced-price lunch (percent)	64.59 (22.90)	
Student-teacher ratio	17.10 (4.01)	
Black students (percent)	31.02 (34.82)	
Hispanic students (percent)	34.53 (36.70)	
Has IEP (percent)	10.61 (10.55)	
In urban area (percent)	32.14 (47.56)	

TABLE A.8(d)

Descriptive Statistics of Variables Used in Estimating Effects: Algebra
(Means and Standard Deviations)

Descriptive	Treatment Group	Control Group
Students		
Female (percent)	52.05	47.24
	(49.98)	(49.95)
Age as of October 2004 (years)	14.84	14.83
	(1.00)	(0.94)
Fall score on ETS Algebra Test (percent correct)		
Total	31.47	33.38
	(11.96)	(12.00)
Concepts	33.17	34.17
	(16.18)	(16.23)
Processes	29.30	30.90
	(15.80)	(17.15)
Skills	32.36	34.91
	(18.81)	(18.26)
Spring score on ETS Algebra Test (percent correct)		
Total	35.08	38.15
	(13.01)	(13.73)
Concepts	35.29	37.50
	(16.75)	(17.05)
Processes	32.48	36.16
	(17.14)	(18.05)
Skills	37.84	40.74
	(19.18)	(20.14)
Teachers		
Female (percent)	51.35	68.75
	(50.67)	(47.09)
Teaching experience (years)	12.36	10.34
	(9.46)	(9.70)
Master's degree (percent)	54.05	53.13
	(50.52)	(50.70)
Time spent using other products (hours a year)	3.80	2.63
	(17.60)	(7.32)

Table A.8(d)

Descriptive	Treatment Group	Control Group
Time spent using study products (hours a year)	45.50 (27.05)	n.a.
Study product used in the classroom (percent)	16.22 (37.37)	n.a.
Access problems (scale from 0 to 100)	52.70 (38.99)	n.a.
Adequate amount of preparation support time (percent)	62.16 (49.17)	n.a.
School has on-site computer specialist (percent)	67.57 (47.46)	n.a.
Participated in technology professional development last school year (percent)	35.14 (48.40)	n.a.
Schools		
Students receiving free/reduced-price lunch (percent)	51.17 (26.14)	
Student-teacher ratio	16.50 (3.73)	
Black students (percent)	42.07 (36.85)	
Hispanic students (percent)	14.79 (22.56)	
Has IEP (percent)	4.91 (5.23)	
In urban area (percent)	56.52 (50.69)	

Appendix B

Estimating Effects and Assessing Robustness

A. Estimating Product Effects

For the study, volunteering teachers in each school were randomly assigned either to implement a software product (treatment teachers) or not to implement it (control teachers). Random assignment of teachers means that the product's effect on student achievement can be estimated by comparing achievement of treatment and control classrooms. The nested structure of the sample was acknowledged by estimating effects using hierarchical linear models. Other analyses also used hierarchical models when the data were clustered by classroom or school.

A Three-Level Hierarchical Linear Model for Estimating Main Effects

The simplest estimator of product effects is the difference of the average spring test scores in treatment and control classrooms. The experimental design ensures that the presence of the product in treatment classrooms is the only difference in the two types of classrooms, on average, and therefore achievement differences can be attributed to the effects of the product.

However, the simple estimator is based on the number of classrooms in the study, which limits its statistical power for exploring relationships between product effects and moderating variables. A model based on the number of students allows the study to increase its statistical power and estimate relationships of effects and moderating variables more precisely.

The simple three-level hierarchical linear model used for estimating the main effects has student, classroom, and school components:

(A.1 Student) $$Y_{ijk} = \alpha_{0jk} + \pi X_{ijk} + \varepsilon_{ijk} .$$

(A.2 Classroom) $$\alpha_{0jk} = \beta_{0k} + \beta_1 T_{jk} + \varphi W_{jk} + \mu_j .$$

(A.3 School) $$\beta_{0k} = \delta_0 + \delta_1 Z_k + \nu_k .$$

where Y is the student's spring test score, X is a set of baseline student characteristics, T is an indicator of whether the classroom is assigned to the treatment group, W is a set of teacher characteristics, and Z is a set of school characteristics. The most important student characteristic in the set X is likely to be the fall test score, because studies often find that initial achievement explains much of the variation in later achievement. Note the cascading relationships of the coefficients, with the intercept of the student equation being modeled as a function of classroom characteristics and the intercept of the classroom equation modeled as a function of school characteristics.

Substituting equations (A.3 into A.2 and the combined equation into A.1) yields a mixed model of student achievement, termed the "main effects model":

$$\textit{(A.4 Main Effects)} \quad Y_{ijk} = \delta_0 + \beta_1 T_{jk} + \delta_1 Z_k + \pi X_{ijk} + \varphi W_{jk} + \xi_{ijk} ,$$

where the combined error term ξ is defined as:

$$\xi_{ijk} = \nu_k + \mu_j + \varepsilon_{ijk} .$$

The mixed model shows that individual student achievement is the sum of average achievement of students in control classrooms; a product effect for students in classrooms using a product; student, classroom, and school characteristics; and an error term with student, classroom, and school components.

Equation (A.4) was estimated with the package HLM 6.03. The package used full-information maximum likelihood techniques and also estimated robust (Huber-White) standard errors, which are reported in the text. As a check, an initial set of models was estimated with SAS Proc Mixed, SUDAAN, and Stata, and results were nearly identical.

Results for a full main effects model are shown in Table B.1. The table indicates the complete set of estimates for all coefficients in the model, listed by level (student, classroom, school). It also shows the residual variances at the three levels at the bottom of the table. Positive coefficients indicate a variable is correlated with an increase in the spring test score and negative coefficients indicate a variable is correlated with a decrease. The units of the coefficient are the same as the units of the test scores, normal curve equivalents for first, fourth, and sixth grades, and percent correct for algebra.

The results are conventional in several respects. One is that the fall score is a strong predictor of the spring score. Another is that having a larger percentage of minority students was correlated with having lower scores. Other findings vary depending on grade level. A significant amount of test score variance is related to classrooms and schools. For example, for first grade, 7 percent of the residual variance of scores (the amount of variance

Table B.1. Main Effects Hierarchical Linear Model Estimates: Outcome is Spring Test Score (with Standard Errors in Parentheses).

Variable Name	First Grade[a]	Fourth Grade	Sixth Grade	Algebra
Student Level				
Intercept	49.56	42.01	51.38	36.16
	(0.54)	(0.43)	(0.60)	(0.47)
Student is female	1.48	0.73	0.35	0.26
	(0.51)	(0.39)	(0.53)	(0.65)
Student age	-2.09	-1.92	-1.17	-0.89
	(0.50)	(0.42)	(0.45)	(0.44)
Fall test scores[a]	0.23	0.35	0.35	0.35
	(0.02)	(0.03)	(0.02)	(0.04)
	0.17	0.29	0.39	
	(0.02)	(0.03)	(0.02)	
	0.32	0.22		
	(0.03)	(0.02)		
	0.42			
	(0.05)			
Classroom Level				
Treatment classroom	0.73	0.41	1.43	-0.87
	(0.79)	(0.58)	(0.99)	(0.88)
Teacher is female	1.82	-0.65	0.31	1.74
	(1.39)	(0.63)	(1.52)	(1.22)
Years of teaching experience	0.01	0.04	0.06	-0.01
	(0.04)	(0.04)	(0.05)	(0.06)
Teacher has a master's degree	-0.55	0.66	-0.1	0.71
	(0.82)	(0.76)	(1.03)	(0.97)
School Level				
Students eligible for free lunch (percent)	-4.19	2.62	7.46	-7.59
	(3.29)	(2.32)	(6.31)	(4.85)
Student-teacher ratio	0.31	0.05	-0.07	-0.48
	(0.40)	(0.11)	(0.21)	(0.17)
Percent of Black students	-3.86	-9.01	-8.87	-5.89
	(2.49)	(2.00)	(3.78)	(2.91)
Percent of Hispanic students	3.93	-10.24	-5.36	-5.49
	(3.29)	(2.39)	(4.07)	(2.49)
Students in special education (percent)	-8.15	-1.3	1.64	-49.58
	(4.97)	(3.54)	(7.34)	(13.23)
School is in urban area	1.42	-0.70	-2.32	2.49
	(1.10)	(0.96)	(2.95)	(1.34)
Residual Variance				
Student level	**123.65**	**99.20**	**131.70**	**117.94**
Teacher level	**10.05**	**4.28**	**14.98**	**8.53**
School level	**6.15**	**3.53**	**3.06**	**0.30**

[a]For fall test score, subtest scores are entered separately. For first grade, the four fall scores are the three SESAT subscores (sounds and letters, word reading, and sentence reading), and the total fall TOWRE score. For fourth grade, the three subscores are entered (vocabulary, word study skills, and comprehension). For sixth grade, the two subscores are entered (procedures and problem-solving). For algebra, the total fall score is entered.

after adjusting for covariates) occurs at the classroom level and 4 percent occurs at the school level. The remaining 89 percent of score variance occurs at the student level. At other grade levels, classroom and school variances also are considerable, with classroom variance the larger of the two. The treatment effects reported in the text refer to the estimated coefficients of the "treatment classroom" indicator variable at the teacher level. For example, the estimate of 0.73 for the treatment effect in the column for first grade is the same as is shown in Table II.6 as the difference between treatment and control group average scores. The *p*-value shown in Table II.6 is the *p*-value of the estimated treatment coefficient.

B. Estimating Moderating Relationships

The hierarchical model supports estimating relationships between overall product effects and classroom and school characteristics. In the main-effects model above, classroom and school characteristics are predictors of student achievement. Some or all of these characteristics also may moderate product effects.

To allow for moderating variables, the model is modified by adding interactions between the treatment indicator and classroom characteristics to equation A.2:

(*A.2′ Classroom Interactions*) $$\alpha_{ojk} = \beta_{0k} + \beta_{1k} T_{jk} + \beta_{2k} T_{jk} \Phi_{jk} + \varphi W_{jk} + \mu_{jk} .$$

The product effect in equation A.2′ also has been modified by adding a school subscript to its estimator. The set of variables indicated by Φ can include the variables in the set *W* but also can include other classroom variables such as implementation variables. The school subscript is added to support the second modification of adding an equation that allows school characteristics to moderate the product effect:

(A.7 School Interactions) $$\beta_{1k} = \gamma_0 + \gamma_1 Z_k + \tau_k .$$

Combining the equations and collecting terms yields a mixed-model estimating equation in which the product effect is related to classroom and school conditions:

(A.8 Mixed Model with Interactions)]

$$Y_{ijk} = \delta_0 + [\gamma_0 + \gamma_1 Z_k + \beta_2 W_{jk}] T_{jk} + [\delta_1 Z_k + \varphi W_{jk} + \pi X_{ijk}] + \xi'_{ijk} ,$$

where the error term has the structure:

$$\xi'_{ijk} = \nu_k T_{jk} + \mu_{jk} + \varepsilon_{ijk} .$$

The estimator's structure in (A.8) has two components, shown in square brackets. The first is the product effect, which is related to school characteristics (*Z*) and classroom characteristics (*W*). The second is the effect of student, classroom, and school characteristics.

Statistically significant estimates of the set of coefficients represented by γ_1 and β_2 are evidence of moderating relationships between characteristics and the product effect.

The mixed model with interactions is one of many kinds of interaction models. In principle, all lower-level coefficients in hierarchical linear models can be modeled as functions of factors at higher levels. However, the complexity of the resulting estimates makes interpretation difficult and simplifying assumptions are needed to reduce complexity and make interpretations clearer.

In particular, the study made two simplifying assumptions. One is that student characteristics affect achievement directly (these relationships are represented by the π coefficients in the model) but that these coefficients do not interact with classroom characteristics. Stated another way, classroom characteristics affect all students in the class similarly. The practical implication of the assumption is that the model does not allow, for example, that achievement of male students may depend on the gender of the teacher or that low-achieving students fare better with teachers who have more experience.

The second simplifying assumption is that the magnitude of the interactions between the product effect and classroom characteristics is not related to school characteristics (there are no higher-level equations for the β_2 coefficients). The practical implication of the assumption is that the product effect may be related to whether a classroom has high or low average student achievement, for example, but the magnitude of the relationship is not in turn related to school characteristics such as poverty level.

A useful feature of the interaction model (A.8) is that it estimates moderating relationships for each characteristic that often is used to create subgroups, while adjusting for moderating relationships with other characteristics. For example, the moderating effect of school poverty is estimated after adjusting for the moderating effect of the school's racial and ethnic composition of students. The estimates can be understood as the effect of each characteristic, holding other characteristics constant.

However, the model's ability to estimate these relationships depends on the degree of variability in the data. If the data include only schools that are both high-poverty and with many minority students (in other words, if the two characteristics are highly correlated), the model may be unable to separate the characteristics, and the result may be statistically insignificant coefficients for poverty and for minority representation. Particularly for the sixth grade and algebra studies, which have fewer schools than the first grade and fourth grade reading studies, the amount of variation in the data is a constraint on estimating moderating relationships.

Table B.2 shows estimates from the full model for the overall score. The full set of estimates shown in the text includes moderating relationships with subtest scores as the outcome, but the structure of the models is the same. The full interactions model has many parameters, more than 40, and the two equations at the third level add to its complexity. The table shows the estimated coefficients in five groupings: three are estimates corresponding to the main model (shown in Table B.1) and two are estimates corresponding to the moderating variables, one at the classroom level and one at the school level. Not all

moderating variables could be included for the math grade levels. The models proved to be numerically unstable when the full set of variables was entered, probably because there were relatively few schools and not enough variability in the data to estimate the large number of moderators.

The third group of coefficients, labeled "Classroom level interactions with treatment," corresponds to the set of variables Φ in equation (A.2′) above. The set of teacher characteristics W is augmented by a set of implementation measures, including the amount of time products were used, whether teachers had problems gaining access to computers, and whether classes had technical difficulties during observations. Positive signs of the estimated coefficients indicate that the product effect was larger when the variable had a larger value, holding other variables constant. For example, the positive coefficient for "Years of Teaching Experience" in the first column means product effects were larger in first grade treatment classrooms when teachers had more years of experience.

The fifth group of coefficients, labeled "School level treatment equation," refers to equation (A.7) above. The same school variables entered in the main model are entered in this equation, if the estimation supports the full set. For the algebra study, the model was numerically unstable when the full set of school variables was included and only two variables were included.

D. Assessing Robustness

The experimental design and the longitudinal data structure support several impact estimators. This section compares impact estimators ranging from simple ones to the complex estimator presented in the text. Data from some school districts also included information from district tests. When the study was able to determine the test properties, it calculated estimates of product effects based on these tests. Not all districts provided test

Table B.2 Estimates of Moderating Relationships From Hierarchical Linear Model: Outcome is Spring Test Score (with standard errors).

Variable Name	First Grade	Fourth Grade	Sixth Grade	Algebra
Student Level				
Student is female	1.47	0.72	0.35	0.32
	0.51	0.40	0.54	0.65
Student age	-2.04	-1.89	-1.14	-0.74
	0.50	0.40	0.45	0.44
Fall test scores[a]	0.23	0.34	0.35	0.35
	0.02	0.03	0.02	0.04
	0.16	0.28	0.40	--
	0.02	0.03	0.02	
	0.33	0.21	--	--
	0.03	0.02		
	0.42	--	--	--
	0.05			
Classroom Level				
Teacher is female	3.96	-1.28	1.07	1.10
	2.50	1.03	1.52	1.43
Years of teaching experience	-0.09	0.09	-0.07	-0.03
	0.05	0.04	0.11	0.09
Teacher has a master's degree	-0.66	0.94	-0.09	0.64
	1.15	0.94	1.21	0.83
Annual hours of other product use (hundreds)	5.28	1.71	4.93	-2.83
	2.82	4.10	11.83	3.68
Classroom Level, Interactions with Treatment				
Annual hours of study product use (hundreds)	2.25	12.10	-3.60	-7.04
	2.15	2.99	2.52	2.96
Location of product use	3.08	2.64	-1.01	-0.03
	2.19	1.07	2.00	2.99
Whether teacher had access Problems	-1.01	0.08	-2.10	-3.78
	1.29	0.86	1.69	1.95
Whether teacher had adequate time to prepare to use product	2.20	0.26	2.04	-0.44
	0.92	0.75	1.36	2.34
Whether school has a computer Specialist	-0.19	0.23	0.48	0.61
	1.22	0.64	3.26	1.70
Years of teaching experience	0.15	-0.17	0.14	0.05
	0.08	0.07	0.13	0.11
Whether teacher had professional development last year on technology	-1.23	-1.13	--	--
	0.92	0.74		
Teacher is female	-1.56	0.27	--	--
	2.66	1.43		
Teacher has a master's degree	-0.52	-0.16	--	--
	1.52	1.34		

Table B.2 *(continued)*

Variable Name	First Grade	Fourth Grade	Sixth Grade	Algebra
School Level				
Intercept	49.72	41.94	51.33	36.11
	0.57	0.43	0.57	0.50
Percent of students eligible for free lunch	-6.78	2.78	7.42	-9.17
	3.28	2.09	6.06	5.69
Student-teacher ratio	0.22	0.05	-0.15	0.44
	0.38	0.13	0.23	0.25
Percent of Black students	-1.68	-7.89	-8.53	-4.86
	2.80	1.85	3.81	3.24
Percent of Hispanic students	7.32	-9.21	-3.73	-4.33
	3.65	2.32	4.18	2.97
Percent of students in special education	-8.38	-3.35	1.17	-44.48
	5.50	3.64	8.52	13.22
School is in urban area	2.28	-1.10	-2.21	2.54
	1.21	1.00	2.49	1.53
School Level Treatment Equation				
Intercept	-2.82	-4.18	-2.15	3.40
	3.71	2.04	3.61	2.75
Percent of students eligible for free lunch	-6.21	-1.69	-4.18	5.03
	4.19	2.40	8.46	4.48
Student-teacher ratio	-1.02	-0.11	-0.41	0.20
	0.32	0.11	0.52	0.46
Percent of Black students	0.88	-3.72	-1.17	--
	3.63	1.74	6.91	
Percent of Hispanic students	8.75	-4.67	-3.41	--
	4.42	1.99	6.32	
Percent of students in special education	5.21	-1.57	8.99	--
	4.41	2.77	11.10	
School is in urban area	-0.86	-1.73	5.12	--
	1.43	0.83	3.78	
Variance Characteristics				
Student level	123.56	99.00	131.88	117.66
Teacher level	5.84	1.18	11.84	7.03
School level	8.11	4.49	2.14	0.46

[a]For fall test score, subtest scores are entered separately. For first grade, the four fall scores are the three SESAT subscores (sounds and letters, word reading, and sentence reading), and the total fall TOWRE score. For fourth grade, the three subscores are entered (vocabulary, word study skills, and comprehension). For sixth grade, the two subscores are entered (procedures and problem-solving). For algebra, the total fall score is entered.

score data that could support separate estimates, especially for first graders (testing first graders is relatively uncommon) and high school students (districts' testing depends on student grade levels).

Other Estimators

Table B.3 shows estimates for overall scores in the four grade levels for four estimators and the study's main estimator (presented in the text). The first estimator, termed the "naïve difference" is the average spring score for students in treatment classrooms minus the average spring score for students in control classrooms. The estimator is straightforward to calculate and has an intuitive interpretation. However, it does not correctly estimate variances when students are nested in classrooms.

Table B.3 Alternative Estimators of Product Effects on Overall Test Score. (Standard Errors of the Estimated Effects in Parentheses)

	First Grade	Fourth Grade	Sixth Grade	Algebra
	SAT-9 Reading Score (NCE)	SAT-10 Reading Score (NCE)	SAT-10 Math Score (NCE)	ETS Algebra Exam (Percent Correct)
Estimator				
1. Naïve difference of student average scores	0.17 (0.82)	0.98 (0.78)	0.77 (0.72)	-1.89 (0.63)
2. Difference of classroom average scores	0.34 (2.04)	-0.09 (1.33)	0.44 (2.37)	-1.64 (1.09)
3. Difference of classroom differences	1.11 (1.06)	0.86 (0.64)	1.09 (1.11)	-1.03 (1.21)
4. Repeated cross section (HLM)	0.97 (0.78)	0.66 (0.55)	1.03 (1.04)	-0.55 (0.93)
5. Panel HLM (reported in text)	0.73 (0.79)	0.41 (0.58)	1.43 (0.99)	-1.70 (0.93)

Source: Author calculations. Other variables in the HLM models include student age, gender, and pretest scores (for the panel); teacher gender, experience, and whether they had a master's degree; classroom average pretest score (for the cross section); school race and ethnicity, percent of students in special education, and percent eligible for free lunch; and student, classroom, and school random effects.

The second estimator is the difference between treatment and control classrooms on the spring test score, with score differences first calculated for each school and the resulting school differences then averaged. The standard error of the average school difference is the estimator of sampling variance.

The third estimator is the previous estimator adjusted for the average classroom score on the fall test. This "differences of differences" estimator allows student mobility to relate to product effects. In principle, for schools with high mobility, the average student should have less exposure to the product because some will not have attended for the whole year. Potential product effects thereby would be attenuated.

The fourth estimator is from an HLM in which spring test scores are a function of fall test scores and other covariates. The fall test score is entered in the second level as the average fall score for the classroom in which the student was located in the spring. Because the model uses student test scores in the spring as the outcome but classroom average scores in the fall as the baseline, it does not assume the same students are tested in the fall and spring, as does the HLM estimator in the text. This "repeated cross sections" estimator is a variant of the "differences of differences" estimator.

Table B.3 shows that the estimator used does not affect the direction of the main findings. Regardless of the estimator, product effects are statistically insignificant and the point estimates are small. The magnitudes of the estimated effects differ, however, with the difference of classroom average posttest scores generally being the smallest and the difference of classroom differences (the posttest minus the pretest) the largest.

The relationship of the repeated cross section HLM estimator and the HLM estimator used in the text (the panel estimator) did not always have the expected direction. The HLM estimator in the text is based on students who completed a fall test *and* a spring test; the cross section HLM is based on the larger group of students who completed a spring test. Because the cross-section sample is likely to have lower average product use, estimates were expected to be smaller than the panel estimate to the extent that the product use mattered. However, for the first and fourth grade reading studies and the algebra study, the cross section estimates appear to be larger than the panel estimates, though differences were not tested for statistical significance. It is difficult to draw any conclusions from small differences between statistically insignificant estimates, however.

District Tests as the Outcome

Another approach for assessing robustness is to estimate effects using district tests that the study collected from records, and compare the results to those using the study's tests. District tests, it could be argued, may be more salient to schools and teachers than the study's tests, which have no accountability dimension. District tests also are administered as a full battery and may measure reading or math more reliably than the shorter test used by the study.

Two important caveats for this comparison relate to test availability and modeling strategies. District tests were not available in all districts and, in particular, were not available for many first graders and high school students because standardized testing is less common in these grades. Also, consecutive year tests are not always available to adjust for pretest levels, and tests themselves sometimes differ in consecutive years. Using district tests also required a different estimation approach because results for each district had to be estimated separately. Here a two-level hierarchical model is used because the number of schools in any one district was too small to support a three-level model.

Table B.4 shows the results of the comparison exercise. Estimated effects were converted to effect sizes using the standard deviation of the control group spring test score, and the effect sizes were compared to the study's results. For 14 districts, the study was able to obtain a district test that supported an estimate of product effect sizes. Because sample sizes are smaller than the study's overall sample size in the grade levels, the focus here mostly is on effect sizes rather than on statistical significance.

The results have the same sign in 10 of 14 districts. Results in the four districts that differed in sign showed a larger effect size on the district test in two districts and smaller effect size in two districts. For some districts, the effect size is larger for the study test, and for some districts it is larger for the district test. Overall, the simple average of effect sizes using district tests is 0.05, and the simple average of effect sizes using the study's test is slightly larger, 0.07.

The findings in Table B.4 support the robustness of the study's findings based on its test. Though not conclusive evidence, because district tests are available for less than half the districts in the study, using district tests do not yield qualitatively different findings about product effects.

Table B.4 Comparison of Effect Sizes Based on District Tests and the Study Test.

Grade	Outcome: District Test	Outcome: Study Test
First Grade		
A	-0.11	-0.08
B	-0.13	-0.02
Fourth Grade		
C	-0.01	0.20
D	-0.04	-0.03
E	-0.07	0.04
F	0.12	-0.52
G	0.04	0.05
H	-0.35	-0.17
Sixth Grade		
I	-0.02	-0.12
J	0.23	0.40
K	0.12	0.15
L	0.52	-0.25
Algebra		
M	0.13	0.22
N	0.06	0.04
Average of Measured Effect Sizes	**0.05**	**0.07**


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA