U. S. DEPARTMENT OF EDUCATION

Effectiveness Of Reading And Mathematics Software Products: Findings From The First Student Cohort

Report to Congress
Executive Summary

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
Executive Summary
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 2007-4006
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.

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).

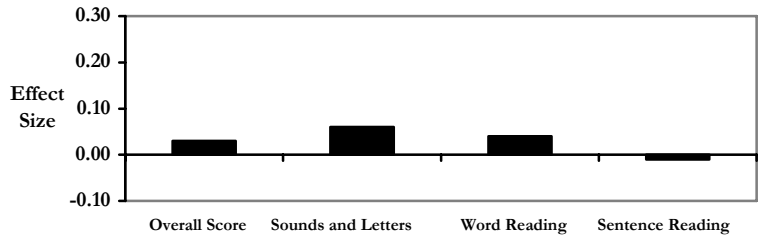
Products provided instruction and demonstration in tutorial modules, allowed students to apply skills in practice modules, and tested students on their ability to apply skills in assessment modules. (The tutorial-practice-assessment modular structure was common for products at other grade levels as well.) 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.

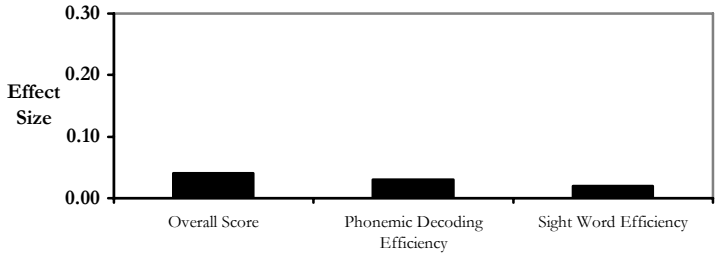
First grade reading products did not affect test scores by amounts that were statistically different from zero. Figure 1 shows observed score differences on the SAT-9 reading test, and Figure 2 shows observed score differences on the Test of Word Reading Efficiency. The differences are shown in "effect 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).

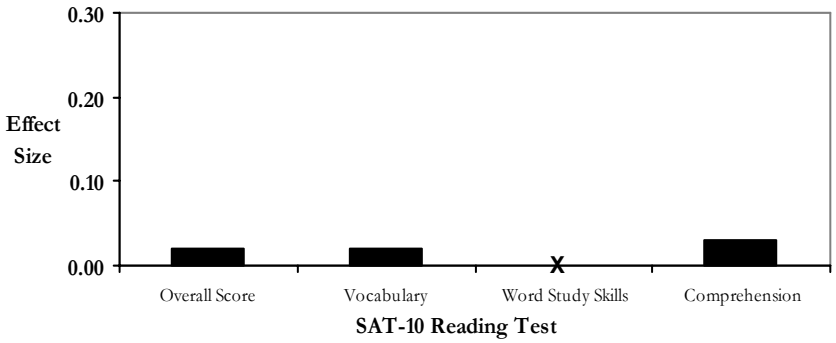
Three of the four products provided tutorials, practice, and assessment geared to specific reading skills, one as a core reading curriculum and two as supplements to the core curriculum. The fourth product offered teachers access to hundreds of digital resources such as text passages, video clips, images, 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.

Figure 3
Effects of Fourth Grade Reading Products on SAT-10 Reading Score



Note: None of the effects is significantly different from zero at the 0.05 level.
X: Estimate is too small to show.

Effects of Sixth Grade Math Products

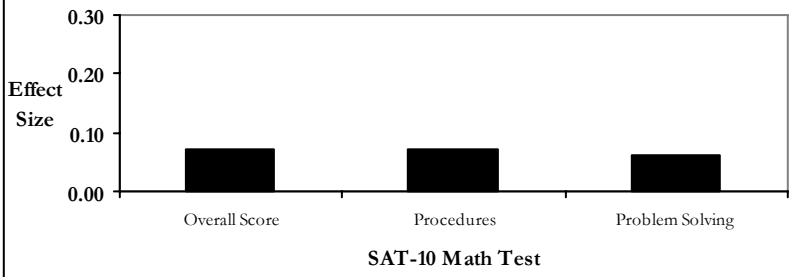
The sixth grade study included three products that were implemented in 10 districts and 28 schools. The sample included 81 teachers and 3,136 students. The three products were Larson Pre-Algebra (published by Houghton-Mifflin), Achieve Now (published by Plato), and iLearn Math (published by iLearn).

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.

Figure 4
Effects of Sixth Grade Math Products on SAT-10 Math Score



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

Effects of Algebra Products

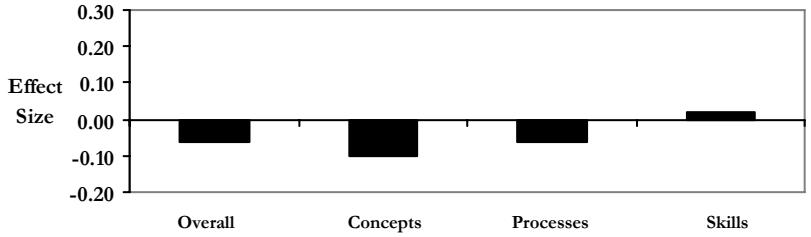
The algebra study included three products that were implemented in 10 districts and 23 schools. The sample included 69 classrooms and 1,404 students. The three products were Cognitive Tutor Algebra (published by Carnegie Learning), Plato Algebra (published by Plato), and Larson Algebra (published by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

Figure 5
Effects of Algebra Products on ETS Algebra Exam



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

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.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA