The ARC Center Tri-State Student Achievement Study

A Report of the ARC Center
at
The Consortium for Mathematics and Its Applications (COMAP)

Sheila Sconiers, Project Director
COMAP

Andy Isaacs
THE UNIVERSITY OF CHICAGO

Traci Higgins
TERC

James McBride, Statistician
THE UNIVERSITY OF CHICAGO

Catherine Randall Kelso
THE UNIVERSITY OF ILLINOIS AT CHICAGO

Funded by the National Science Foundation,
Proposal No. ESI-9729328

This study was prepared with the support of the National Science Foundation. However, any opinions, findings, conclusions, and/or recommendations herein are those of the authors and do not necessarily reflect the views of the NSF.

Table of Contents

Abstract 3

1. Introduction 5

1.1 The Curricula 5

1.2 Prior Research 5

1.3 The Study 7

2. Method 8

2.1 Selecting States 8

2.2 Surveying Schools 8

2.3 Identifying Eligible Schools 9

2.4 Matching Reform and Comparison School-Grade Combinations 9

2.5 Exclusions, Missing Data, and Weights 11

3. Results 11

3.1 Comparisons by State-Grade Combination 12

3.2 Comparisons by Race/Ethnicity 16

3.3 Comparisons by Income 18

3.4 Comparisons by Gender 20

4. Conclusion 20

5. References 21

Appendix A 23

Appendix B 29

Appendix C 33

Tables 1–14

In 2000–2001, the ARC Center, located at the Consortium for Mathematics and Its Applications (COMAP) [http://www.comap.com/elementary/projects/arc/], carried out a study of reform mathematics programs in elementary schools in Illinois, Massachusetts, and Washington. The study examined the performance of students using three elementary mathematics curricula—*Everyday Mathematics; Math Trailblazers;* and *Investigations in Number, Data, and Space*—on state-mandated standardized tests administered in spring 2000. The study included over 100,000 students, 51,340 students who had studied one of the three reform curricula for at least two years and 49,535 students from non-using comparison schools matched by reading level, socioeconomic status, and other variables. Small differences on the SES variables remaining between the reform schools and the matched comparison schools were further controlled by adjustments based on regression analyses. Usage of the reform curricula was verified by a telephone survey of schools and districts.

Results show that the average mathematics scores of students in the reform schools are significantly higher than the average scores of students in their matched comparison schools. The results hold across five different state-mandated tests, and across topics ranging from computation, measurement, geometry, and algebra to problem solving and making connections. The study compared the scores on all the topics tested at all the grade levels tested (grades 3–5) in each of the three states. Of 34 comparisons across five state-grade combinations, 28 favor the reform students, six show no statistically significant difference, and none favor the comparison students. The results also hold across all income and racial/ethnic subgroups, except for Hispanic students, where there are no significant differences between the scores.

As Ball and Cohen (1996) point out, there are good reasons to believe that changing curriculum materials can alter classroom instruction. Curriculum materials provide the activities that shape the daily interactions between teachers and students. If they are written with sufficient specificity, curriculum materials can help teachers translate research findings and authoritative recommendations into classroom reality. Because they can be easily disseminated, curriculum materials have the potential to help large numbers of teachers transform their classroom instruction.

In the early 1990s, the National Science Foundation (NSF) recognized this potential and funded groups from three institutions to create comprehensive elementary mathematics curricula that would be research-based and aligned with the vision for school mathematics in the National Council of Teachers of Mathematics (NCTM's) *Curriculum and Evaluation Standards for School Mathematics* (1989). The institutions were The University of Chicago, the University of Illinois at Chicago, and TERC in Cambridge, MA. The curricula those groups produced are *Everyday Mathematics; Math Trailblazers;* and *Investigations in Number, Data, and Space,* respectively.

1.1 The Curricula

While each of the programs is unique, the three teams of curriculum designers began with the same goal: to develop materials with broader, more rigorous content than traditional texts. At the same time, the projects sought to develop curricula with balanced approaches to the subject matter and to teaching. Curriculum development was based on detailed examinations of current practice and on recent research findings about children's mathematical learning. This common goal has resulted in curricula that claim to:

- Build on children's experiences;
- Teach basic arithmetic as well as geometry, data analysis, measurement, probability, and concepts of algebra;
- Challenge students with engaging and meaningful applications;
- Connect topics within mathematics and with other subjects;
- Encourage students to solve problems in many ways;
- Balance skills with concepts and problem solving;
- Include a variety of instructional approaches;
- Help teachers extend their understanding of mathematics and teaching; and
- Provide a variety of assessment instruments and procedures.

Development of the curricula began in the late 1980s. Each program was created one grade at a time to allow for extensive field testing to inform both revisions and the design of subsequent materials. *Everyday Mathematics* (EM) was published between 1989 and 1998. *Investigations in Number, Data, and Space* (IN) was published between 1994 and 1998, and *Math Trailblazers* (MT) was published in 1997 and 1998.

1.2 Prior Research

A growing body of research suggests that elementary mathematics instruction aligned with the NCTM Standards (1989, 2000) has a positive impact on student achievement. (Carpenter, Fennema, Peterson, Ching, & Loef, 1989; Cobb, Wood, Yackel, Nicholls, Wheatley, Trigatti, & Perlwitz, 1991; Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996; Fuson & Briars, 1990; Hiebert & Wearne, 1993, 1996; Smith, Lee, & Newman, 2001; Wood & Sellers, 1997.) Much of this early research was exploratory and small scale—focused on individual classrooms, schools, and districts. Student achievement was often assessed with instruments developed by the researchers rather than with standardized tests. Many of these studies were based on work with teachers and on instructional materials designed to supplement or replace parts of whatever curriculum was already in place. Although these research results are suggestive, they cannot be used to predict effects of comprehensive Standards-based curricula such as those targeted in this study.

Schoenfeld (2002) points out that Standards-based curricula are just entering a phase of large-scale implementation. In his examination of the available literature he found evidence that at least some of these curricula are producing gains in student achievement on measures of conceptual understanding and problem solving. He found no evidence that the curricula are having a negative impact on basic skills. Likewise, Hiebert (1999) found that instructional programs emphasizing conceptual understanding could produce the desired learning without sacrificing skill proficiency. Although the evidence is limited, both Schoenfeld and Hiebert found it encouraging.

Preliminary research does suggest that *Everyday Mathematics*, *Math Trailblazers,* and *Investigations* are having a positive impact on student learning in individual classrooms, schools, districts, and in some cases larger geographic regions such as cities and states. (Briars & Resnick, 2000; Carroll, 1997; Carroll & Isaacs, 2003; Carter, Beissinger, Cirulis, Gartzman, Kelso, & Wagreich, 2003; Flowers, 1998; Fuson, Carroll, & Drueck, 2000; Goodrow, 1998; Mokros, 2003; Mokros, Berle-Carman, Rubin, & O'Neil, 1996; Mokros, Berle-Carman, Rubin & Wright, 1994: Riordan & Noyce, 2001.) Under a variety of research conditions, including state-level comparisons of reform and traditionally-taught students, longitudinal studies, pre- and post-test comparisons, case studies, dissertation research, and publicly available school district evaluation reports, students using these curricula have been found to outperform students using other programs. (See Carroll & Isaacs, 2003; Carter, Beissinger, Cirulis, Gartzman, Kelso, & Wagreich, 2003; and Mokros, 2003 for reviews of this work.)

For example, Carter, Beissinger, Cirulis, Gartzman, Kelso, & Wagreich (2003) report on five studies examining the impact of *Math Trailblazers* on student achievement. Several of the studies examined standardized test performance of students prior to implementation of *Math Trailblazers* versus performance after several years of implementation. Other studies examined student achievement in schools that were using the curriculum compared with matched schools not using the curriculum. In the final study reported, a school within a school that had adopted the curriculum, was compared to the larger school, which had not. In all cases standardized tests scores were compared and students using the *Math Trailblazers* program outperformed comparison groups not using the program. Of special interest is a case study of student performance in a suburban district. Students in the only school in the district that used *Math Trailblazers* performed significantly better on the open-ended items of the *Stanford Achievement Test*, Ninth Edition (SAT9) than the other students in the district.

Similarly, a series of early studies suggests a positive impact of the *Investigations* curriculum on student learning. Three studies comparing students using *Investigations* with students using other curricula were reviewed in Mokros (2003; also see Flowers, 1998; Goodrow, 1998; Mokros, Berle-Carman, Rubin & Wright, 1994). The studies used a variety of testing techniques, examining traditional computation skills, higher level problem solving, and conceptual understanding using both paper/pencil and interview methods. In these studies, matched comparison groups and *Investigations* students showed similar gains across the school year on basic computation problems, but the comparison students were less successful than the *Investigations* students on tasks requiring complex problem solving or a deep understanding of arithmetic operations.

Likewise, Carroll & Isaacs (2003) reviewed a variety of studies examining the impact of *Everyday Mathematics* on student achievement. In the reported studies, students were tested on a variety of standardized instruments (the *Illinois Goals Assessment Program;* the *Comprehensive Testing Program*, 3rd Edition; and *the Metropolitan Achievement Test*, 7th Edition) and on well-known research-based tests (the *National Assessment of Educational Progress* and the *Cognitively Based Elementary Math Test* (Wood & Cobb, 1989)). Studies were conducted using longitudinal designs (e.g., Briars & Resnick, 2000; Fuson, Carroll, & Drueck, 2000), comparisons between matched schools (e.g., Carroll, 1997), and a comparison between cohorts exposed to long-term implementation versus no implementation of the curriculum within the same district were used (e.g., Briars & Resnick, 2000). Across these studies, students who had used the *Everyday Mathematics* program outperformed peers who had had little or no exposure to EM.

Several studies have contributed directly to the work of the ARC study. In these studies, large samples of students using the Standards-based curricula were compared to matched groups using state-mandated, standardized test scores as the dependent variable.

One such study evaluated the impact of the *Everyday Mathematics* curriculum on student achievement among third grade students in Illinois during the 1992–1993 school year. Carroll (1997) identified schools that were using the program as the core curriculum in all third grade classes. Twenty-six schools, all located in the Chicago metropolitan area, were identified as meeting this criterion. In 14 of the 26 schools, EM had been in place since the students were kindergartners. The performance of these students on the *Illinois Goals Assessment Program* (IGAP), the state assessment at that time, was compared to the performance of students in suburban Cook County and in the rest of the state. Carroll found that 25 of the 26 schools had mean scores significantly above the state mean and 20 of the 26 schools had mean scores significantly above the suburban Cook County mean. (None of the schools scored significantly below the state mean and one had a mean significantly below that of suburban Cook County.) Additionally, more than half of the students who had been using EM since kindergarten exceeded the state math goals, double the percentage of the students in the rest of the state.

Another study investigated the effects of *Everyday Mathematics* and *Connected Mathematics* (CMP; a Standards-based middle school curriculum developed with NSF support) on 4th and 8th grade students' mathematics scores on the 1999 *Massachusetts Comprehensive Assessment System* (MCAS) (Riordan & Noyce, 2001). Groups of early and late implementers of each of the curricula were compared to matched schools not using the programs.

It was found that students using EM or CMP outperformed their counterparts in terms of their overall scaled scores and on some subtests (e.g., number sense and geometry) (Riordan & Noyce, 2001). Effect sizes ranged from small to moderate depending on the outcome measure and duration of implementation. Generally, the longer the implementation, the greater the advantage for reform students, both overall and within subgroups of students. Evidence showed some narrowing of the performance gap between racial/ethnic groups and between advantaged and disadvantaged populations. For example, Black, Hispanic, and low SES reform students outperformed White and higher SES students in the comparison groups. Among early implementers, girls outperformed boys. Positive differences were consistent across performance quartiles suggesting that the two curricula were effective for students at the bottom, middle, and top of the achievement continuum.

Another study of similar design examined student achievement resulting from the use of two Standards-based middle grades curricula; MATH*Thematics* and *Connected Mathematics* (CMP) in Missouri (Reys, Reys, Lapan, Holliday, & Wasman, 2003). The first three districts in the state to begin implementing these curricula were compared to three districts matched in terms of prior test scores and free/reduced lunch rates. The comparison districts were also selected to mirror the 6–8 configuration of the schools in the reform districts and to reflect similar geographic locations. Student achievement was assessed by the 1999 *Missouri Assessment Program* (MAP) and by the *Terra Nova* test. In each paired comparison the reform district outperformed the comparison district on at least two subtests (data analysis and algebra) and no significant differences favored the comparison districts.

1.3 The Study

As Standards-based curricula reach multi-year, full-scale implementation in many schools and districts, more is learned about how these materials affect student achievement. Preliminary evidence from a number of studies using an array of research designs indicates that these curricula are improving student performance on a variety of measures—both experimental instruments and standardized tests—without sacrificing basic skills. However, most of the studies examined the outcomes of implementation efforts in schools and districts that were among the first to embrace the approach embodied in these curricula, the early adopters. Since these early adopters received a degree of support from publishers and developers that could not be replicated given the large number of new implementers, the need to assess the curricula under more typical conditions became apparent.

In 1999, EM, MT, and IN were used in about 10% of the nation's school districts, which served about three million students. Therefore, the programs were used widely enough and had sufficient maturity to warrant a large-scale study of their effects on student achievement. Of particular interest was how reform students fared on standardized tests. This resulted in a research design intended to answer the questions: *How does the achievement of students who use EM, MT, or IN compare to that of students using other curricula on state-mandated, standardized tests?* and *Does this finding hold across subtests and student subgroups?*

2. Method

This study combined publicly available state test data from Illinois, Massachusetts, and Washington with survey data from schools using EM, MT, and IN. The combined data set made it possible to compare the achievement of students studying these curricula with matched comparison students not using any of the three curricula.

2.1 Selecting States

The ARC Center study focused on Illinois, Massachusetts, and Washington for two reasons. First, the reform programs EM, MT, and IN were represented by substantial numbers of users in these states, and second, the five different standardized tests mandated in these states permitted analysis across a variety of instruments. The five tests were:

- *Illinois Standards Achievement Test* (ISAT), grade 3
- *Illinois Standards Achievement Test* (ISAT), grade 5
- *Massachusetts Comprehensive Assessment System* (MCAS), grade 4
- *Iowa Test of Basic Skills* (ITBS), grade 3 (Washington State)
- *Washington Assessment of Student Learning* (WASL), grade 4

With the exception of the ITBS, the state-mandated tests were constructed to measure student achievement as defined in published documents that outline the mathematics learning standards for each state. The ITBS is a national, norm-referenced test. The five tests include a broad range of content including number sense, computation, estimation, algebraic concepts, geometry, measurement, probability, data analysis, and problem solving. All the tests include multiple-choice items. The MCAS and WASL also include short-answer questions and extended-response items. See Appendix B for a detailed description of these assessment instruments.

2.2 Surveying Schools

To assure that only schools fully implementing EM, MT, or IN were in the reform group and that no users of these programs were in the comparison group, a survey instrument was developed and administered. Each project conducted a telephone survey of districts and schools in Illinois, Massachusetts, and Washington that were known to use, or were suspected of using, its curriculum. These districts and schools were identified principally through customer lists provided by the program publishers. A district or school on these lists was surveyed if the annual sales for that district or school exceeded a given dollar amount, which varied by program, and even by state for a given program. The dollar amount cutoffs for inclusion in the survey were set low enough to achieve a near census of all students in the three states using the curricula. In addition, each project maintains a database of individuals and school systems that have contacted them with implementation questions or requests. Each project checked these databases for schools known to be using their materials but not appearing on the customer lists. Such schools were also surveyed. Schools and districts designated to be surveyed included at least 90% of all students in the three states using the three curricula.

The coverage rate for a survey is the ratio of total students in all schools responding to the survey to total students in all schools designated to be surveyed. Coverage rates were generally at the 90% level or higher.

The survey collected 1999–2000 school year implementation information for the grades for which state achievement test data were also available. To gauge the extent and length of implementation, survey data were also collected for previous grades. Survey respondents included district math supervisors, principals, or other knowledgeable persons.

The primary reason for conducting the implementation survey was to verify usage of the reform curricula. Additional questions about implementation variables such as staff development and time allotted for mathematics instruction were included in the survey of reform schools; but without corresponding data for the comparison schools, the analysis was limited. See Appendix C for the survey instrument. The analysis below uses the following survey information, which was coded for each school-grade combination:

- Percentage of teachers fully using a reform program (A "fully using" teacher was defined as one who used a program for at least 75% of his or her mathematics instruction.)
- Number of years at full program implementation ("Full implementation" was defined as at least 75% of teachers using the program for at least 75% of their mathematics instruction.)

2.3 Identifying Eligible Schools

A grade in a school (a school-grade case) was considered eligible for inclusion in the analysis, provided:

1. The grade was one for which student test data is available (grades 3 and 5 in Illinois, grade 4 in Massachusetts, and grades 3 and 4 in Washington);
2. The school-grade reported full implementation of EM, MT, or IN during the 1999–2000 school year; and
3. The program had been implemented in the previous grade within the school for at least two years (1998–2000)—so that students in the given grade would have had at least a two-year exposure to the program. In schools that did not include the previous grade, this requirement was modified to require that the program had been implemented in the previous grade for at least two years (1998–2000) for all possible feeder schools to that school.

The coded implementation survey file contained 1,058 school-grade records for which student test data were available (criterion 1, above). Of these, 742 (70%) were classified as eligible and were subsequently matched to comparison schools. Failure to meet the two-year implementation requirement (criterion 3, above) was the most common reason for classifying a school-grade case as "ineligible." The distribution of all school-grade case records, by state and grade level, is shown in Table 1. The distribution of eligible reform school-grade cases, by program, state, and grade level, is shown in Table 2.

A total of 110 different school districts are represented by the 742 eligible reform school-grade cases: 52 districts in Illinois, 38 in Massachusetts, and 20 in Washington.

2.4 Matching Reform and Comparison School-Grade Combinations

A matching routine was carried out for each of the five state-grade combinations in order to identify comparison schools that had not implemented any one of the three reform programs but that were similar in how they would be expected to perform on the respective statewide test. The matching procedure selected one matched comparison school for each of the 742 eligible reform school-grade cases included in the analysis.

Within each state-grade combination, schools known to use, or suspected of using, any of the three reform curricula were excluded as possible matched comparison schools. All remaining schools appearing on that state's public education data files formed the pool of schools eligible for selection as comparison schools. The following schools were excluded as possible comparison schools:

- All schools that were identified to be surveyed, even if they were subsequently classified as ineligible, or if the survey was not completed.

- Schools that were not surveyed, but that appeared on one of the publishers' customer lists with more than trivial annual sales.
- Schools that appeared neither on the publishers' customer lists nor program databases as current users, but that were known to have used one of the reform programs in the recent past.

Separate school-level regression analyses for the five state-grade combinations identified the strongest predictors of the average school mathematics score for each state-grade test. Reading score and low-income variables (variously designated as "low income" in Illinois, "eligible for free or reduced price lunch" in Massachusetts, and "Title I status" in Washington) consistently accounted for the greatest percentage of total variance. These variables were given greater weight in the matching process. Other variables—such as percent White, school mobility rate, and percent with limited English proficiency (LEP)—accounted for little of the total variance, but were typically significant. These variables were given less weight in the matching process.

The actual matching routine was carried out separately for each of the five state-grade combinations. No single routine could be used across states because the school, district, and student data available and related to matching varied by state. Moreover, the matching ratios (number of available comparison schools: number of reform schools to be matched) varied considerably by state. For example, matching in Massachusetts (where the matching ratio was less than 6 to 1) did not allow the same flexibility in accommodating multiple, simultaneous matching variables as matching in Illinois (where the matching ratios were about 10 to 1).

Table 3 shows the number of excluded schools and matching ratio for each state-grade combination. For the combined state-grade combinations, approximately 9% of the total schools were "exclusions" that had not been surveyed. As expected, the exclusion rates were highest in Washington, where all three reform programs are well represented; and lowest in Illinois, where only one program is well represented.

The variables used in matching for the different state-grade combinations were as follows:

- Illinois:
  School averages for reading score, low-income percent, White percent, LEP percent, and mobility percent.
- Massachusetts:
  School averages for reading score, free/reduced lunch percent, and White percent.
- Washington:
  School averages for reading score, Title I Mathematics percent, and White percent.
  (The school variable Title IS was also used as a stratification variable in Washington: A reform school and its matched/comparison school were required to have the same Title I status.)

For each reform school-grade case, the matching routine identified a comparison school that resembled the reform school with respect to the matching variables. As a starting position for matching within any state-grade combination, the maximum difference allowed in school reading scores was set to 1 point. For Illinois and Massachusetts it became necessary to broaden this difference to 2 points. A maximum 2% difference in averages between reform and matched schools for the low-income variable (low-income percent in Illinois, free/reduced lunch percent in Massachusetts, and Title I Mathematics percent in Washington) was allowed. Similarly, a maximum 5 percent difference in averages for White percent was allowed. See Appendix A.1 for a complete description of the matching routine; see Table 12 for an inventory of the matching variables and their data sources.

Table 4 shows the matching variable averages for students in the 742 eligible reform school-grade cases and their comparison school-grades. There is generally close agreement between the reform and comparison averages for the matching variables, but differences do exist, and such differences could bias any subsequent comparisons of mathematics test scores. Therefore, the comparison-student averages for all test variables were adjusted before any tabulated comparisons were made. Adjustment ensured that any bias ensuing from the matching procedure was minimized. (See Appendix A.4 for a complete description of the adjustment procedures.)

2.5 Exclusions, Missing Data, and Weights

Before any differences in the performance of reform and comparison students were tabulated, procedures were carried out to deal with missing data, unequal school size, and inconsistencies in the student populations tested in the several states.

Illinois and Washington identify students with an Individualized Education Program (IEP); Massachusetts identifies “mathematically disabled” students; and Washington identifies “special education” students. All reform and comparison student records for IEP, “mathematically disabled,” and “special education” students were deleted from the analysis and excluded from the tabulated comparisons. These deleted records represent approximately 10% of all student records.

Fewer than 3% of the student records included missing or incomplete math test data or reading scores. All such records were deleted from the analysis and excluded from the tabulated comparisons. Approximately 3% of the student records included missing values for race/ethnicity. Such records were not deleted; instead, a school-level value for “White percent” was imputed as a surrogate for the student-level variable “White” to each student record for that school with missing data. See Appendix A.2, together with Table 5, for an expanded discussion of missing-data procedures.

Table 6 shows, for each state-grade combination, the number of student records for reform and comparison students that were in fact used for tabulated comparisons and all subsequent analysis. In all, more than 100,000 student records are represented, with approximately equal numbers of reform-student and comparison-student records.

The near equality in numbers of reform-student and comparison-student records shown in Table 6 does not apply, however, at the individual school level. The difference between the number of students in a given reform school-grade and its matched school-grade was highly variable and sometimes substantial. Weighting was therefore necessary, and case weights were constructed for all comparison-student records. (See Appendix A.3 for a description of the construction of case weights.) Use of case weights for all tabulations (Tables 4 and 7–11) ensured that comparison schools contributed to overall statistics with the same proportions as their reform-school counterparts.

3. Results

Differences between reform and comparison student scores were tabulated for each of the five state-grade combinations; these results were also pooled to yield overall comparisons. Differences disaggregated by race/ethnicity, income, and gender were also tabulated; these disaggragated comparisons pooled results from all five state-grade combinations.

The mathematics test variables used for all tabulations are student-level variables. They vary somewhat by state and even within state for Washington. The overall mathematics test score variables are “math” and “total.” “Math” is the scaled test score; “total” is the percent of total possible points on the test. Each of the variables “computation,” “measurement,” “geometry,” “prob/stat,” and “algebra” denotes the percent of total possible points for the corresponding strand of test items.

The Massachusetts test categorizes test items by type as “open-response,” “short answer,” or “multiple-choice.” The Washington State tests categorize test items into various skill sets: “problem solving,” “concepts and estimation,” “logical reasoning,” “communicating understanding,” and “making connections.” A test variable with any such name denotes the percent of total possible points for the corresponding category of test items.

Each set of tabulated comparisons for a state-grade combination compares averages for reform students and comparison students within that state-grade combination. These differences between averages were not calculated simply by subtracting the observed comparison-student average from the observed reform-student average for each test variable. The observed comparison-student average for each test variable was instead adjusted prior to subtraction. The adjustment procedure was based on regression analyses and ensured that any bias ensuing from imperfect matching of reform and comparison schools was minimized. (See discussion at the end of Section 2.4.) Appendix A.4 describes the adjustment procedure more fully.

Having calculated an adjusted difference of average scores between reform students and their comparison students, the effect size for that difference was then calculated by dividing the adjusted difference by the standard deviation of the comparison student scores. Effect sizes are reported in order to facilitate comparisons across states and grade levels and comparisons with prior research. For comparisons by race/ethnicity, income, and gender that combine results from the individual state-grade tabulations, effect sizes were calculated as weighted-average effect sizes, taken across the state-grade combinations. (See Appendix A.5 for details about the effect size calculations.)

For this study, an effect size can be thought of as the percentile standing of the average reform student relative to the average comparison student. An effect size of 0.10 (the approximate three-state weighted average effect size for both the "math" and "total" test scores) indicates that the mean of the reform-student group is at the 54th percentile of the comparison group. This, in turn, implies a change in percentile standing of 4 percentile points. Tables 7–10 all use the label "percentile change" to denote the change in percentile standing of the average reform student relative to the average comparison student, as determined by the effect size.

3.1 Comparisons by State-Grade Combination

Table 7 shows comparisons for all test variables, by state-grade combination. By definition, each of the test variables except "math" has a range of 100 points (0% to 100%), so that the average differences reported may be interpreted either as point differences out of 100 points or as percentage differences. "Math," however, is a scaled score whose range is 80 points in Illinois and Massachusetts, greater than 100 points in Washington (grade 3), and greater than 350 points in Washington (grade 4), so that average differences for this variable may not be interpreted as percentage differences.

The effect sizes for "math" and "total" are approximately the same for all state-grade combinations—as expected, since these are the overall test score variables. The combined state-grade effect sizes for "math" and "total" are virtually identical and correspond to a percentile change of about 4% favoring the reform students.

Regarding "math" and "total" as a single comparison within each state-grade combination, 34 different comparisons are represented in Table 7. Six of these comparisons show no statistical difference; 28 of the comparisons show a significant difference, and all favor the reform students. The combined state-grade effect sizes are highly significant ($p < 0.001$) for all mathematics strands; and they are fairly consistent across strands, with probability and statistics as the single exception.

The pattern of differences and effect sizes across state-grade combinations varies according to the mathematics strand.

- For measurement, the differences are all highly significant and the effect sizes are consistent.
- For algebra, the differences are all highly significant, but the effect sizes for Massachusetts and Washington are roughly double those for Illinois.
- For computation[1], all differences are highly significant except for Washington grade 3 ($p < 0.01$). The effect sizes for Illinois and Massachusetts are consistent, but the effect sizes for Washington are about one-third those for Illinois and Massachusetts. The combined state-grade effect size, however, is approximately the same as the combined effect size for "total."
- For geometry, three of the differences are highly significant but the difference in Massachusetts is not significant.
- For probability and statistics, only the Illinois grade 5 difference is highly significant, and all other differences are virtually zero.

The comparisons shown in Table 7 are differences and convey no information about the level of student performance. Table 11, however, does show the actual levels of student performance corresponding to the

[1] The strand "computation" is variously labeled: Estimation/Number/Sense/Computation (IL), Number Sense and Numeration (MA), Number Sense (WA), and Math Concepts and Math Computation (ITBS).

differences in Table 7. (Each difference in Table 7 may be recovered from Table 11 by subtracting the corresponding reform- and comparison-student averages.) For comparative purposes, Table 11 also shows the levels of performance for all non-reform students. "Non-reform students" include all students within a state-grade combination that do not attend any of the eligible reform schools, and that are not identified as IEP, mathematically-disabled, or special-education students[2]. In particular, non-reform students include all comparison students. The side-by-side bar graphs in Figures 1 and 2 summarize information from both Tables 7 and 11. The differences reported in Table 7 correspond to the differences in adjacent bar heights in Figures 1 and 2.

Figure 1: Averages for the overall test score variable "total," by state/grade and reform status.

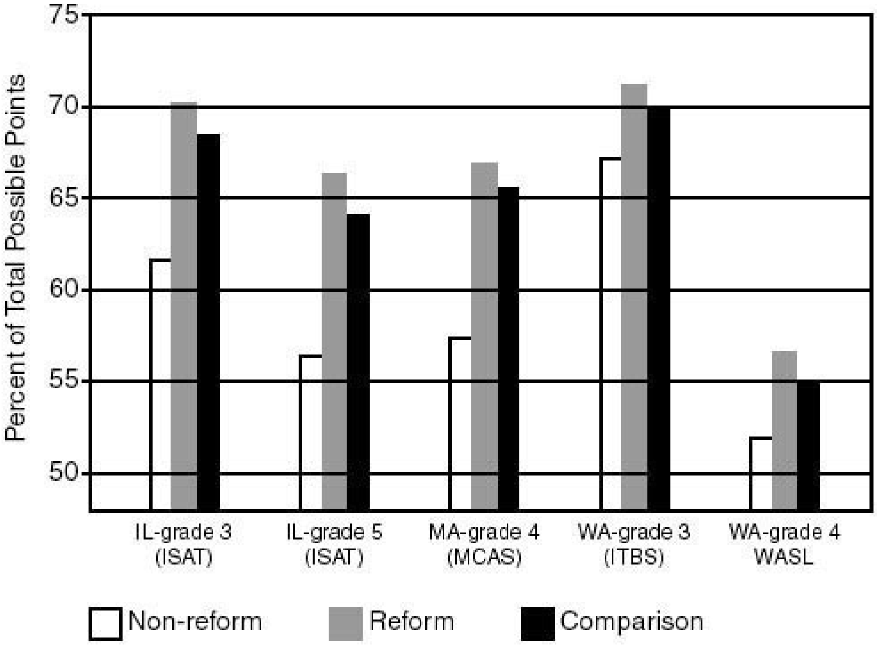


Figure 1 shows that the lag in average performance between non-reform and comparison students is substantially larger than the lag between comparison and reform students—across all state-grade combinations, but especially in Illinois and Massachusetts. Including averages for all non-reform students in Table 11 and Figure 1 highlights the impact of the matching procedure used to select comparison schools. The reform schools have, relative to other schools in their state, higher average reading scores, higher percentages of White students, and lower percentages of low-income students, all of which are associated with higher mathematics achievement. Direct comparison of reform- and non-reform-student performance would, therefore, largely reflect the differences in these variables between the two student groups and would not furnish valid measures of the effects of the reform curricula. The matching procedure, however, selected comparison schools with comparable values for these variables, and the differences between reform- and comparison-student performance do furnish valid measures of program effects. See Appendix A.6 for a discussion of the validity and measures taken to assure that regression-to-the-mean artifacts are not producing spurious results.

[2] The "non-reform students" in Table 11 are the aggregated students in all schools not classified as "eligible reform schools." (Table 2 inventories the "eligible reform schools," which are defined in Section 1.3.) IEP, mathematically disabled, and special education students are not included in any tabulations.

Figure 2a: Test variable averages for IL grade-3, by reform status.

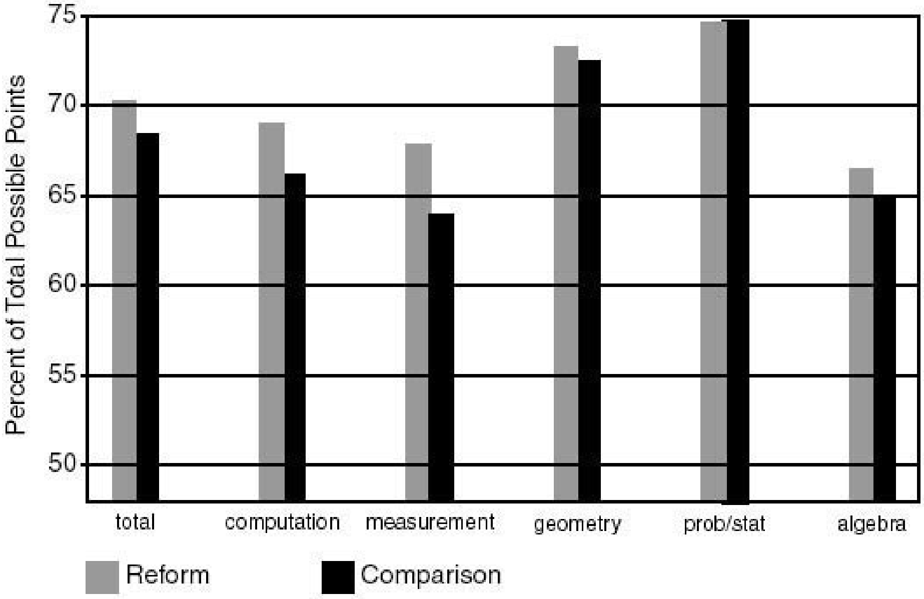


Figure 2b: Test variable averages for IL-grade 5, by reform status.

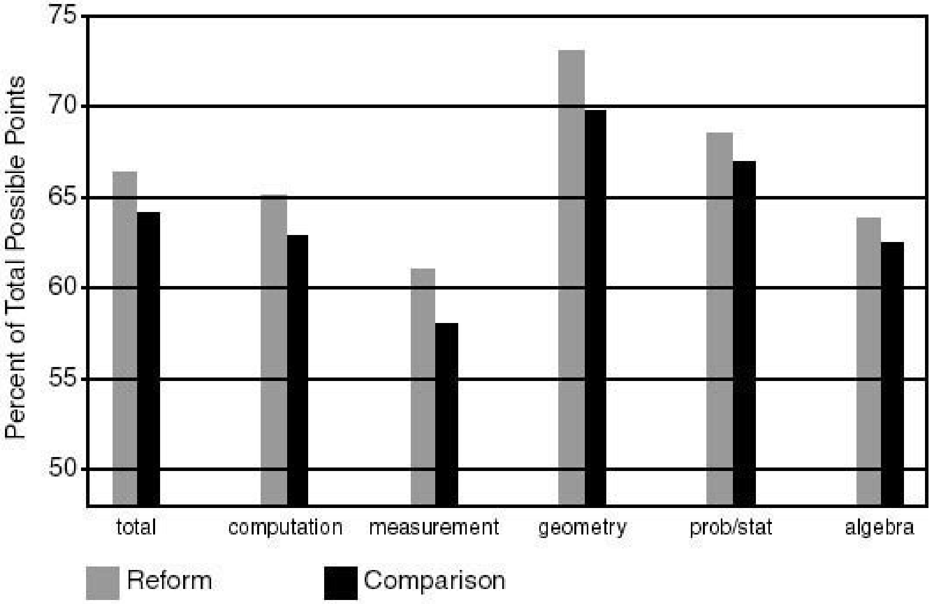


Figure 2c: Test variable averages for MA-grade 4, by reform status.

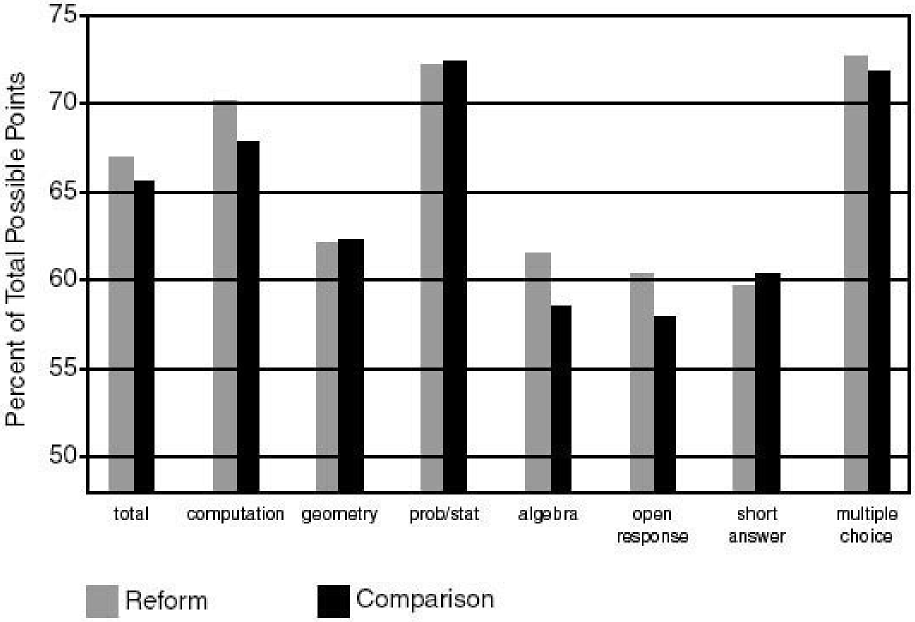


Figure 2d: Test variable averages for WA-grade 3, by reform status.

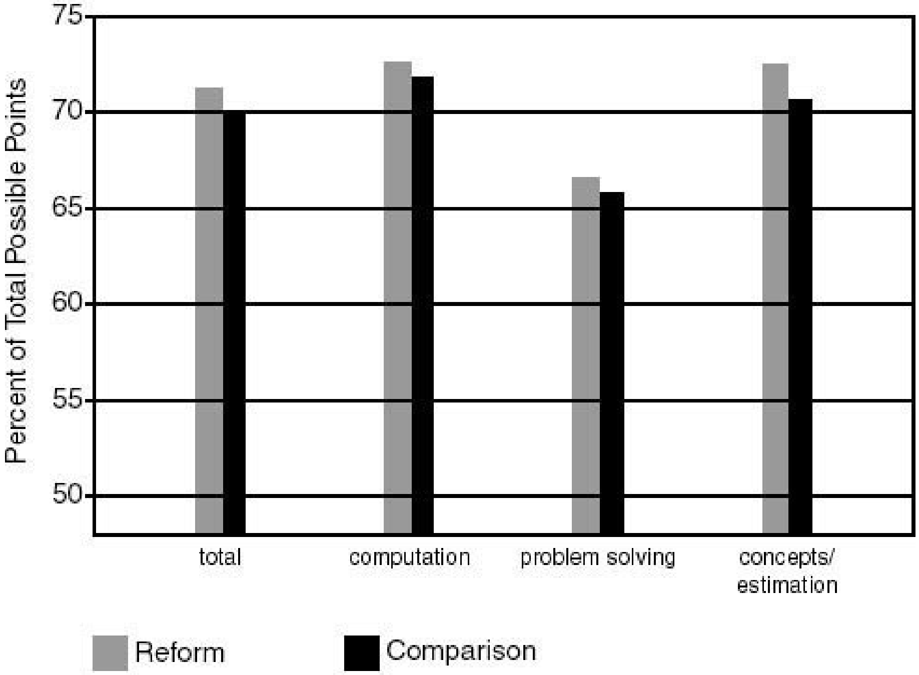
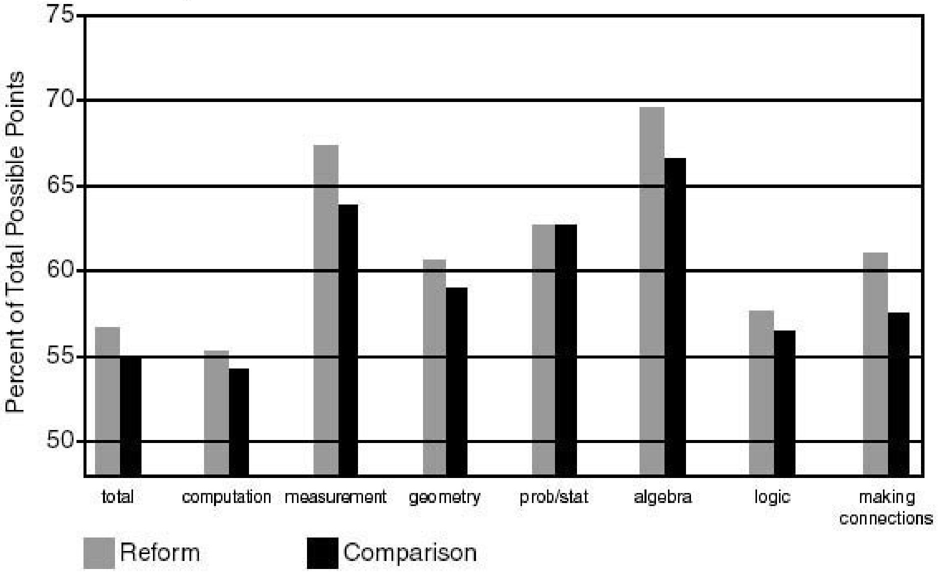


Figure 2e: Test variable averages for WA-grade 4, by reform status.



Averages for "problem solving" and "communicating" are not shown because they do not fit within the vertical scale.
For "problem-solving", the reform students' average is 38.82, and the comparison student's average is 36.31.
For "communicating", the averages are 45.03 and 45.06, respectively.

Figures 2a–2e show averages separately for each of the five state-grade combinations. Each graph shows averages for all test variables considered for that particular state and grade.[3] The graphs all use the same vertical scale to facilitate comparisons across state-grades. They are collectively useful for emphasizing the high variability in average scores across the different strands and other test item categories. Within each state-grade combination, there is typically a 10- to 15-point spread between the highest and lowest average strand score. For Washington-grade 4, there is a 30-point spread between the averages for algebra and problem solving.

3.2 Comparisons by Race/Ethnicity

Table 8 shows comparisons by race/ethnicity that combine results from the individual state-grade tabulations. Effect sizes are shown separately for Asians, Blacks, Hispanics, and Whites. Student data for those identified as Native American, "mixed," or "other" race categories, or for those whose race/ethnicity was missing and subsequently imputed, were included only in the calculation of the combined state-grade effect size and not in any comparisons by race/ethnicity.

The effect sizes shown by Table 8 are remarkably similar for Blacks and Whites, and for both races these effect sizes very nearly duplicate the combined state-grade effect sizes. The effect sizes for Asians are generally at the same level or higher than those for Blacks and Whites—they are at about the same level for computation and algebra, higher for measurement and probability and statistics, and much higher for geometry. With the exception of probability and statistics, virtually all of the effect sizes for Asians, Blacks, and Whites are highly significant and favor the reform students within each racial subgroup.

The results for Hispanics, however, are quite different. None of the effect sizes for "math," "total," computation, algebra, and probability and statistics are statistically significant for Hispanics. The effect sizes for measurement and geometry are both positive and significant; each, however, is smaller than the corresponding effect size for Asians, Blacks, or Whites.

[3] All differences are shown in the graphs in Figures 2a–2e, including differences that were not statistically significant.

The effect sizes for probability and statistics are exceptional within Table 8, just as they are within Table 7. They are small and generally favor the reform students, but are not statistically significant except for Whites. The combined state-grade effect size for probability and statistics is highly significant, but so small (0.025) that the result lacks practical significance.

The effect sizes shown in Table 8 are averages of the effect sizes for racial/ethnic subgroups taken across the state-grade combinations. But Table 8 conveys no information about the actual test variable averages and their differences that were used to calculate the effect sizes. Figure 3 complements Table 8 and shows averages for the overall test score variable “total” by racial/ethnic subgroups, within state-grade combinations. The differences in adjacent bar heights represent the differences in averages that were used to calculate effect sizes and construct the Table 8 effect sizes for “total.”

Figure 3: Averages for the overall test score variable "total," by state/grade, race/ethnicity, and reform status.

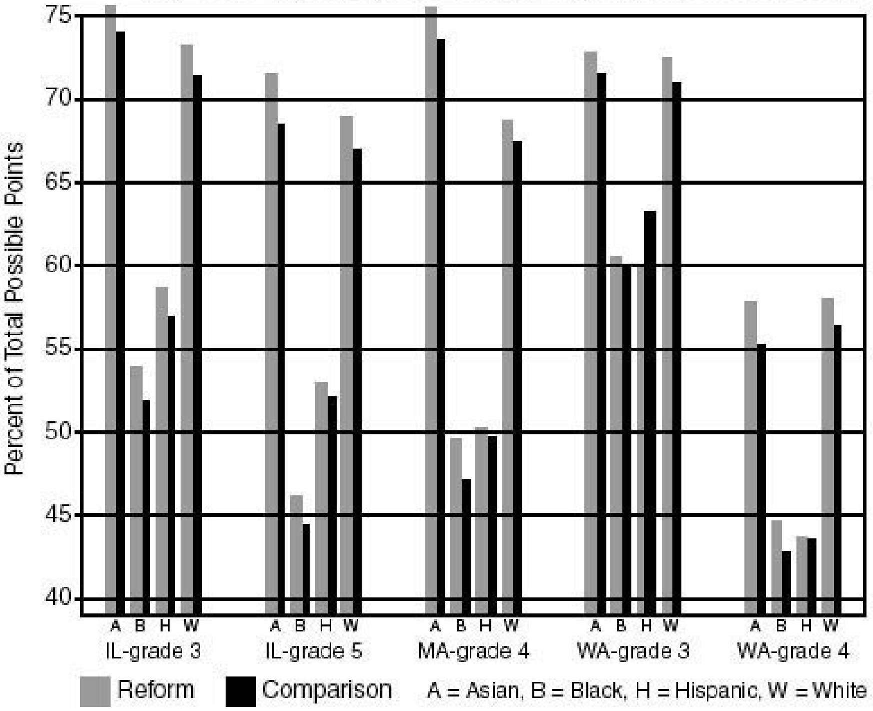


Figure 3 shows several consistent patterns:

- Of the 20 comparisons shown, 19 favor reform students. The only comparison favoring comparison students is for Hispanics within Washington-grade 3.
- The actual differences in averages between reform and comparison students are roughly similar within and across the state-grade combinations for Asians, Whites, and Blacks (excepting Washington-grade 3).
- Averages for Asians and Whites are substantially higher than averages for Blacks and Hispanics—by at least 10 points and by up to 25 points.
- Averages for Asians are marginally higher than those for Whites, and averages for Hispanics are marginally higher than those for Blacks.

3.3 Comparisons by Income

Table 9 shows comparisons by student family income, which combine results from the individual state-grade tabulations. Comparisons are reported by Title IS status for Washington State, and by socioeconomic status (SES) categories for combined Illinois/Massachusetts data.

Recall that in Washington State, the school-level variable Title IS was used as a stratification variable for matching: The comparison school for each reform school-grade case was required to have the same Title IS status as the reform school to be matched. The tabulations for Washington in Table 9 offer a direct comparison of effect sizes for lower-income Title IS students with those for higher-income non-Title IS students.

For Illinois and Massachusetts, a school-level variable called “SES” was defined for each reform school. The SES categories “low,” “middle,” and “top” were defined by using low-income school percent (in Illinois) and free or reduced lunch school percent (in Massachusetts) to assign each school to one of these three categories. Each comparison school was then categorized in the same way as the reform school it was matched to. The SES categories defined for reform schools include the following percentages of low-income (Illinois) and free or reduced lunch (Massachusetts) students: 47% in “low,” 12% in “middle,” and 2% in “top.” That is, 47% of the students in the “low” category schools are low-income, 12% of the students in the “middle” category schools are low-income, and 2% of the students in the “top” category schools are low-income. Virtually the same percentages hold for the SES categories defined for the comparison schools. Table 9 offers a direct comparison of effect sizes for the three SES categories of schools.

The effect sizes shown by Table 9 are quite similar across SES and Title IS categories for the overall test score variables “math” and “total.” All such effect sizes are highly significant and all favor the reform students. Effect sizes for students in low-SES and top-SES schools are at the same level, and marginally higher than those for students in middle-SES schools. Effect sizes for Title IS students are marginally higher than those for non-Title IS students, and marginally lower than those for low-SES students.

The pattern of effect sizes across SES categories varies according to the mathematics strand.

- For computation, effect size increases with SES status. The effect size for students in low-SES schools is two-thirds that for students in high-SES schools.
- For measurement, effect sizes for students in low- and high-SES schools are at the same level, and are larger than the effect size for students in middle-SES schools by a 4 to 3 margin.
- For geometry, the effect size for students in low-SES schools is three times that for students in high-SES schools. The effect size for students in middle-SES schools is zero (no difference between reform and comparison students).
- For algebra, all effect sizes are at about the same level.
- For probability and statistics, the effect sizes are exceptional, just as they were within Tables 7 and 8, and the previous discussion applies.

Except for probability and statistics, all of the effect sizes noted above are highly significant and all favor the reform students.

A similar analysis of the pattern across Title IS and non-Title IS categories according to strand is not productive except for the computation strand. The Title IS effect sizes for measurement, geometry, algebra, and probability and statistics are based on 507 student records, representing only six schools. While these effect sizes are reported in Table 9, they should not be considered as reliable. For the computation strand, the Title IS effect size is positive (favoring reform students), not significant (although nearly so), and actually larger than the corresponding effect size for non-Title IS students.

Figure 4: Averages for the overall test score variable "total," by state/grade, SES and Title IS status, and reform status.

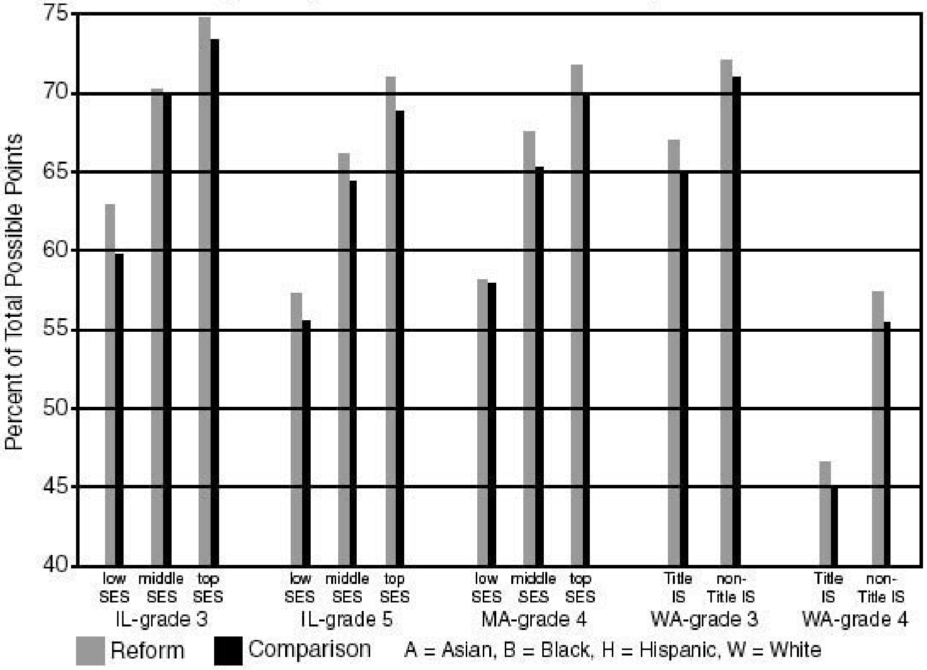


Figure 4 complements Table 9 in the same way that Figure 3 complements Table 8. Figure 4 shows two consistent patterns.

- Averages increase with income status, within each state-grade combination.
- Of the 13 comparisons represented, all differences in averages favor the reform students.

The pattern of differences in averages across income categories, however, varies according to the state-grade combination.

- For Illinois-grade 3, the difference for students in low-SES schools is double that for students in high-SES schools and the difference for students in middle-SES schools is negligible.
- For Illinois-grade 5, the differences are approximately equal for all SES categories.
- For Massachusetts-grade 4, the differences for students in middle- and high-SES schools are approximately equal and the difference for students in low-SES schools is negligible.
- For each Washington grade, the differences by Title IS status are roughly equal.

The low-income variable used in Illinois to define SES categories was the school-level variable “low-income school percent.” Individual student records could not be used to identify low-income students. Direct comparison between low-income students attending the reform and their matched schools was therefore impossible. The low-income variable used in Massachusetts to define SES categories was the student-level variable “free or reduced lunch.” (Student-level data were aggregated to create the school-level variable “free or reduced lunch school percent,” and this school variable was used to define SES categories.) This made it possible to directly compare students attending the reform and their matched schools by free/reduced lunch status. And tabulations of adjusted differences were made in the usual way.

- For free/reduced lunch students, none of the differences are significant. The difference for “total” is zero; differences for computation, algebra, and open-response items are positive and favor the reform students; and differences for the other four strands and item categories are negative and favor the comparison students.

- For non-free/reduced lunch students, the differences are all marginally higher than those shown for all Massachusetts reform students in Table 7, and have the same significance levels.

3.4 Comparisons by Gender

Table 10 furnishes comparisons by student gender that combine results from the individual state-grade tabulations. All of the effect sizes shown in the table are positive and favor the reform students, and all are highly significant, except those for probability and statistics, which are significant at the $p < 0.01$ level. While the effect sizes for females are very slightly larger for "math" and "total" and across all strands except for probability and statistics, such differences are insignificant.

4. Conclusion

This study examined achievement test data for a near census of students in three states using NSF-funded comprehensive elementary mathematics curricula. These students' test results were compared to those of students in non-using schools carefully matched by reading level, SES, and other variables. Possible bias due to imperfect matching was controlled by adjustments based on regression studies. The principal finding of the study is that the students in the NSF-funded reform curricula consistently outperformed the comparison students: All significant differences favored the reform students; no significant difference favored the comparison students. This result held across all tests, all grade levels, and all strands, regardless of SES, gender, and racial/ethnic identity. Use of these curricula results in higher test scores.

Two anomalous results deserve further study. One of these is the finding that for the most part reform students did not outperform the comparison students in probability and statistics. The few differences that were statistically significant did favor the reform students, but those differences were smaller than in other areas. Since all three reform curricula include considerable work with probability and data analysis, it may be that the failure of the reform students to outperform the comparison students can be attributed to a misalignment of the tests and the curricula. The tests in the study appear to assess primarily low-level skills in probability and statistics and may therefore fail to measure the learning that the curricula promote in these areas. A careful content analysis of the test instruments may resolve this issue.

The other anomalous result is that although Hispanic students using the reform curricula generally outperformed the comparison Hispanic students, the differences were small and not statistically significant. Understanding this result is complicated by the varying status of bilingual education in the several states and differences among the states in the rules that govern when LEP students are required to take state-mandated tests. Further study will require a careful look at the relationship among several factors including use of language in the curricula, inclusion practices, and SES. It may be, for example, that limited English proficiency among Hispanic students limits the benefit they receive from the classroom discussions that are integral to these programs.

There are at least two directions for further study. One is to examine implementation variables more carefully, including staff development, number of years of use, time allotted for mathematics instruction, use of supplementary materials, and so forth. Questions about these variables were included in the telephone survey of the reform schools, but without corresponding data for the comparison schools, the analysis was limited. A minimum requirement for such a study would be to survey the comparison schools to gather implementation data; a more ambitious approach would be to survey individual teachers in both reform and comparison schools. A second direction for further work would be to replicate the study in other states, including states with innovative assessment systems that emphasize problem solving, mathematical reasoning, and communication.

5. References

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curricular materials in teacher learning and instructional reform? *Educational Researcher*, *25* (*9*), pp. 6–8.

Briars, D. & Resnick, L. (2000). Standards, assessment—and what else? The essential elements of standards-based school improvement (CSE Technical Report 528). Los Angeles: Center for the Study of Evaluation, UCLA. http://www.cse.ucla.edu/CRESST/Reports/TECH528.pdf

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experiemental study. *American Educational Research Journal, 26 (4),* 499–551.

Carroll, W. M. (1997). Results of third-grade students in a reform curriculum on the Illinois state mathematics test. *Journal for Research in Mathematics Education, 28 (2),* 237–242.

Carroll, W. M. and Isaacs, A. (2003). Achievement of students using the University of Chicago School Mathematics Project's Everyday Mathematics. In Sharon L. Senk & Denisse R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Lawrence Erlbaum.

Carter, A., Beissinger, J., Cirulis, A., Gartzman, M., Kelso, C., and Wagreich, P. (2003). Student learning and achievement with Math Trailblazers. In Sharon L. Senk & Denisse R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Lawrence Erlbaum.

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., & Perlwitz, M. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education, 22 (1),* 3–29.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, V. R., Jacobs, & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction, *Journal for Research in Mathematics Education*, *27*, 403–434.

Flowers, J. (1998). A study of proportional reasoning as it relates to the development of multiplication concepts. Unpublished doctoral dissertation, University of Michigan, Ann Arbor. (Cited in Mokros, 2002).

Fuson K. C. & Briars, D. J. (1990). Using a base-ten blocks learning/teaching approach for first and second-grade place-value and multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *21*, 180–206.

Fuson, K., Carroll, W., and Drueck, J. (2000). Achievement results for second and third graders using the Standards-based curriculum Everyday Mathematics*. Journal for Research in Mathematics Education, 31 (3),* 277–295.

Goodrow, A. (1998). Children's construction of number sense in traditional, constructivist, and mixed classrooms. Unpublished doctoral dissertation, Tufts University, Medford, MA.

Hiebert, J. (1999). Relationships between research and the NCTM Standards. *Journal for Research in Mathematics Education, 30 (1),* 3–19.

Hiebert, J. & Wearne, D. (1993). Instructional tasks, classroom discourse, and student learning in second-grade arithmetic. *American Educational Research Journal*, *30*, 393–425.

Hiebert, J. & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction, 14*, 251–283.

Mokros. J. (2003)—Learning to reason numerically: The impact of Investigations. In Sharon L. Senk & Denisse R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn*? Mahwah, NJ: Lawrence Erlbaum.

Mokros, J., Berle-Carman, M., Rubin, A., & O'Neil, K. (1996, April). Learning operations: Invented strategies that work. Paper presented at the annual meeting of the American Educational Research Association, New York.

Mokros, J., Berle-Carman, M., Rubin, A., & Wright, T. (1994, December). Full year pilot grades 3 and 4: *Investigations in Number, Data, and Space*. Cambridge, MA: TERC (available from the author).

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and evaluation standards for school mathematics*. Reston, VA: Author.

Reys, R., Reys, B., Lapan, R., Holliday, G., Wasman, D. (2003). Assessing the impact of Stabdards-based middle grades mathematics curriculum materials on student achievement. *Journal for Research in Mathematics Education*, *34* (1), 74–95.

Riordan, J. E., & Noyce, P. E. (2001). The impact of two standards-based mathematics curricula on student achievement in Massachusetts. *Journal for Research in Mathematics Education*, *32* (4), pp. 368–398.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31 (1),* 13–25. (http://MathematicallySane.com)

Smith, J., Lee, V., & Newmann, F. (2001). Improving Chicago Schools: Instruction and achievement in Chicago elementary schools. Report of the Chicago Annenberg Research Project. Consortium on Chicago School Research: Chicago, IL.

Wood, T. & Cobb, P. (1989). The development of a cognitively-based elementary school mathematics test: Final report. West Lafayette, IN: Purdue University, School of Mathematics and Science Center.

Wood, T. & Sellers, P. (1997). Deepening the analysis: Longitudinal assessment of a problem-centered mathematics program. *Journal for Research in Mathematics Education, 28 (2),* 163–186.

Appendix A

A.1 The Matching Procedure

The matching routine identified, for each reform school-grade case, a comparison school that resembled the reform school with respect to the matching variables. Matching was carried out separately for each of the five state-grade combinations. The variables used in matching for the different combinations were as follows:

- Illinois:
  School averages for reading score, low-income percent, White percent, LEP percent, and mobility percent.
- Massachusetts:
  School averages for reading score, free/reduced lunch percent, and White percent.
- Washington:
  School averages for reading score, Title I Mathematics percent, and White percent.

In addition, for Washington State, the school variable "Title IS" identifies Title I schools (At least 40% of the students receive free/reduced lunch.) and was used as a stratification variable: A reform school and its matched/comparison school were required to have the same Title IS designation. In Washington, approximately 27% of all schools with third grade classrooms and 20% of all schools with fourth grade classrooms are Title IS schools. (The variable Title IS should not be confused with "Title I Mathematics," which is a student-level variable used to designate Title I students in mathematics.)

As a starting position for comparing the values of matching variables for two schools within any state-grade combination, the maximum difference allowed in school reading scores was set to 1 point; and a "score" variable (a linear combination of the standardized school differences in the low-income variable, White percent, mobility percent, and LEP percent) was defined, using a weight of 2 for low-income percent, and weights of 1 for the remaining variables. The low-income variable used varied according to state—"low-income percent" in Illinois, "free/reduced lunch percent" in Massachusetts, and "Title I Mathematics percent" in Washington.

Passing through the reform schools sequentially, each reform school was compared with all available comparison schools, and a "match" declared for the lowest value of "score"—ignoring any "score" for a comparison school that had been previously selected as a "match" for another reform school. For a given set of weightings for the matching variables in the "score" variable, the list order of reform schools determined the complete set of matches. A comparison school, once matched, could not be used again as a match. Thus, reform schools not near the top of the list could receive relatively high matching "scores," because all of their most compatible comparison schools had already been used as matches. To remedy this, the sequential matching procedure was repeated twice. Prior to each repetition, the 10% of all reform schools having just received the highest matching scores were advanced to the top of the list. The matches determined by the second repetition were regarded as the final matches.

A maximum difference in averages between reform and matched schools of 2% for the low-income variable (low-income percent in Illinois, free/reduced lunch percent in Massachusetts, and Title I Mathematics percent in Washington) was enforced. Similarly, a maximum difference in averages of 5% for White percent was enforced. If either of these maxima was exceeded, the maximum difference allowed in reading scores was broadened until such average differences were within their required ranges. (A cap of 2 points was set for the difference in school reading scores.) Once this situation prevailed, the matching program was rerun several times, with the weights of the components of the "score" variable systematically perturbed—in an attempt to reduce further the difference in averages between reform and matched schools for low-income percent. In all cases, for any alternative version of the "score" variable considered, the weight of each non-low-income variable was constrained to be less than the weight of the low-income variable. Additionally, the difference in number of students between a reform school and its match was minimized to the extent possible.

A.2 Exclusions and Missing Data Procedures

Table 5 shows, for each state-grade combination, the student record counts for all reform and comparison students at various stages of exclusion.

- Column A shows total record counts for data available from state files—for the combined set of reform and matched students, prior to any exclusions.
- Column B shows record counts after mathematically-disabled, IEP, and special-education students' records have been excluded.
- Column C shows record counts after all remaining records containing incomplete mathematics test data have been excluded.
- Column D shows the extent of missing variable information for the records in column C. At this stage, reading score and race/ethnicity are the only variables used by the matching routine that have any missing data for student records.

No attempt was made to impute any missing reading score or race/ethnicity for students in Massachusetts. Records with such missing data (1.2% of the total) were simply excluded for all subsequent analysis. Nor was any attempt made to impute any missing reading score in the remaining state-grade combinations. Student records with this value missing were excluded from all subsequent analysis.

Approximately 3% of student records included a missing value for race/ethnicity, and the pattern for this missing data was problematic in Illinois and Washington. Twelve different schools completely withheld student race data: 7 in Illinois-grade 3, 4 in Illinois-grade 5, and 1 in Washington-grade 3. Nearly all of the remaining schools had only a small percentage of missing race/ethnicity data. Excepting Massachusetts, the following imputation strategy was used, and no such student records were excluded from the subsequent analysis.

- For Illinois (grades 3 and 5), school-level race/ethnicity data was available from the School Report Card state data file. A school-level value for "White percent" was imputed as a surrogate for the student-level variable "White" to each student record for that school with missing data.
- For Washington (grades 3 and 4), the school-level value for "White percent" was calculated from the partial student data available for that school, and appended to student records as just described. For the one school with no available data, a school-level regression was run for Washington (grade 3) to model "White percent" as a function of the remaining matching variables. The predictive equation generated an imputed value that was appended to student records as just described.

A.3 Construction of Case Weights for Comparison Students

Case weights were constructed for all comparison-student records. Use of case weights for all tabulations (Tables 4 and 7–11) ensured that comparison schools contributed to overall statistics with the same proportions (number of cases) as their reform-school counterparts.

The case weight is the same for all students within a given comparison school-grade, but varies over the set of comparison school-grades. Each case weight is a ratio that may be described by using Table 5: If R denotes a reform school-grade case and M denotes its comparison school-grade, the case weight attached to each comparison-student record in school-grade M was constructed using the file whose counts appear in column E of Table 5:

- case weight = (# of R records in column E file) / (# of M records in column E file) .

A.4 Corrective Adjustment for Matching Differences

The matching routine described above identified, for each reform school-grade case, a comparison school that resembled the reform school with respect to the matching variables. Table 4 shows that there is generally close agreement between the reform-student and comparison-student averages for those matching variables. But differences do exist that could bias any comparisons. Therefore, comparison-student averages for the overall test score variables ("math" and "total") and all remaining test variables (such as "computation," "algebra," and "problem solving") were adjusted before any tabulated comparisons were made. Adjustment ensured that any bias ensuing from the matching procedure was minimized.

The adjustment procedure followed these steps:

1. Comparison-student averages were adjusted independently within the five different state-grade combinations. And they were adjusted independently for each test variable within a given state-grade combination.
   In cases of subgroup analysis that required comparisons by race/ethnicity, income, and gender, averages were also adjusted independently within each relevant subgroup.
   Example: For a subgroup analysis by gender, comparison-student averages were adjusted independently for males and females. And for each gender, adjustments for the different test variables (such as "total," "geometry," and "algebra") were executed independently. Finally, all adjustments within a given state-grade combination were executed independently of those for the other combinations.

2. Consider a given test variable for a given subgroup, within a given state-grade combination. Using comparison student data for that subgroup, a linear regression of the test variable (dependent variable) was run on all of the matching variables (independent variables) used by the matching routine for that state-grade combination.

3. Let Pred(y|X) denote the value of the resulting predictive equation, evaluated for a specific combination of matching (independent) variables X. More specifically, let
   Pred(y|reform X) = Pred(y|observed matching variable averages for reform students)
   Pred(y|comparison X) = Pred(y|observed matching variable averages for comparison students)

4. Let Obs(y) denote the observed average for the given subgroup of comparison students on the given test variable. Obs(y) was adjusted as follows:
   Adj(y) = Obs(y) + [Pred(y|reform X) – Pred(y|comparison X)]
   The adjusted difference between averages for this test variable and subgroup that was used for tabulations is:
   Observed test-variable average for the subgroup of reform students – Adj(y) .

The adjusted difference for any test variable and subgroup furnishes an estimate of the "treatment effect" for that variable and subgroup, where "treatment" is defined as use of one of the reform programs. The adjusted difference estimates are similar in form and generally quite close to the treatment estimates furnished by an analysis of covariance (ANCOVA) model with multiple covariates (the matching variables), where the regressions are constrained to be linear and parallel. The adjusted difference estimates are also generally quite close to the estimates furnished by an ANCOVA model that allows for non-parallel regressions, where each covariate is included in the model as a deviation from its overall mean (for combined reform and comparison students). See Thomas Cook and Donald Campbell, *Quasi-Experimentation*, Houghton Mifflin, 1979: p. 155.

A.5 Calculation of Effect Sizes and Their Standard Errors

Each calculation of effect size for a difference of average scores used the standard deviation of the comparison-student scores. The focus of this study is on change relative to the untreated group—i.e., relative to the set of comparison students. The standard deviation of the comparison-student scores is therefore more appropriate for effect size calculations than the average of the standard deviations of the reform- and comparison-student scores. In fact, for this study, the method used to calculate a reference-value standard deviation has negligible impact on effect size.

Consider ($\text{avg}X_R - \text{avg}X_C$), the unadjusted difference of average scores between reform and comparison students for the test variable X within a given state/grade combination or subgroup of that combination. The variance of this difference is

$\text{Var}(\text{avg}X_R - \text{avg}X_C) = \text{Var}(\text{avg}X_R) + \text{Var}(\text{avg}X_C)$

$= (1 - f)\sigma_R^2/n_R + \sigma_C^2/n_C$,

where σ_R^2 and σ_C^2 are the population variances, n_R and n_C are the sample sizes, and f equals the proportion of reform students sampled. The proportion f is approximately 0.81 within each state/grade combination. For example, in Illinois, surveyed schools include approximately 90% of all reform students in the state, the response rate was 94%, and 96% of student records have no missing data: thus

$f = 0.90x0.94x0.96 = 0.81$

For all test variables $\sigma_R^2 \approx \sigma_C^2$ and $n_R \approx n_C$ (unweighted), so that

$\text{Var}(\text{avg}X_R - \text{avg}X_C) \approx 1.19\sigma_C^2/n_R$

For the associated estimated effect size ($\text{avg}X_R - \text{avg}X_C)/s_C$, where s_C is the sample estimate of σ_C, the standard error is approximately $1.09/\text{sqrt}(n_R)$.

The effect sizes quoted in this report for a given state/grade combination have the form ($\text{avg}X_R - \text{avg}X_{CA})/s_C$, where X_{CA} is the adjusted comparison-student average described in Appendix A.4. More generally, weighted-average effect sizes taken across state/grade combinations have the form $\Sigma_i\ (n_{Ri}/N)\ (\text{avg}X_{Ri} - \text{avg}X_{CiA})/s_{Ci}$, where n_{Ri} equals the number of reform students sampled from state/grade i and $N = \Sigma_i\ n_i$. If all differences of average scores were <u>un</u>adjusted (i.e., if each X_{CiA} was replaced by X_{Ci} in the expression above), the general weighted-average effect size would have standard error ≈ $1.09/\text{sqrt}(N)$. For convenience, all significance levels quoted in this report are based on standard errors calculated via this simplifying assumption.

A bootstrapping analysis was used to generate realistic standard errors of effect size, and these were compared with estimates calculated under the simplifying assumption. The latter were invariably larger, usually exceeding the realistic standard errors by 25 to 50%. Thus, the actual p-values are consistently smaller than the p-values used to determine significance levels within this report. For example, some effect sizes reported as significant at the $p < 0.01$ level are very likely significant at the $p < 0.001$ level.

A.6 Validity of Results

Reliability of measures and regression artifacts

The selection of matched/comparison schools is described in Section 2.4 and in Appendix A.1. Measurement errors in any one of the covariates used for matching can bias the estimates of treatment differences and effect sizes, and this problem has been described as a "regression artifact." For example, assume that schools were matched solely on the basis of fallible (unreliable) reading test averages: School mathematics test averages would regress toward the respective group means, resulting in an overestimate of the treatment effect. The key feature here is reliability, which may be conceptualized either as "stability" or "high test-retest correlation." Matching on unreliable covariates may bias the estimates of treatment effects.

With the exception of average school reading score, all matching variables were measured as either school-wide percentages or school-grade percentages. In some cases these percentages were calculated by aggregating individual student data; in other cases, reported school-level data was used. (See Table 12.) All such measures may be considered to be extremely reliable, even when school-level data is used as a surrogate for grade-specific data. For example, in grades 3 and 5 for Illinois, two independent sources of race/ethnicity data are available—the School Report Card reports school-level data, and the student file can be aggregated to generate grade-specific data. The correlation between these measures exceeds 0.98; and the correlation actually increases when IEP students are removed prior to aggregation (as they were in the analysis.)

The four standardized reading tests represented in the study have approximately the same reliability for reading test scores: Individual student test-retest correlations are all about 0.90. The matching procedure, however, used aggregated units (schools) rather than individuals (students) to identify the control population. Because group means are more stable than individual scores, the test-retest correlations for school average reading scores are all about 0.99. The operative unit for matching is the school, and thus the operative reliability for the reading test covariate is 0.99.

A simulation study was used to quantify the bias in reported treatment effects due to regression artifacts. Bias was estimated for a range of assumed reliabilities for the covariates used in matching, and for situations in which one or multiple covariates were used to determine the matches. Table 13 shows a selection of typical results from the simulation study. Note the following:

- The estimated bias for all but one of the situations described in Table 13 is positive. Positive bias means that treatment effects are overestimated, and in a direction that favors the reform students.
- For realistic covariate reliabilities (0.99 for reading score, and 0.98 for the remaining covariates), the estimated bias is typically small—averaging 4 to 5%, and rarely exceeding 10%.
- Even for relaxed covariate reliabilities that are substantially smaller than those likely to hold for this study, the estimated bias is frequently less than 10% and rarely exceeds 20%

An alternative analysis

School reading score is a covariate that was used in matching and that was measured after treatment implementation. It is plausible that student reading scores may have improved as a result of using one of the reform programs, which would complicate the interpretation of treatment differences and effect sizes.

An alternative analysis was therefore considered. The matching procedure was repeated, but average school reading score was dropped as a required matching variable in each case.[1] Table 14 summarizes the treatment differences estimated under this revised matching procedure. For each state/grade combination:

- Differences in line (A) are the average differences reported in Table 7, under full matching that included school reading score. All differences have been adjusted, as described in Appendix A.4.
- Differences in line (B) are the average differences under revised matching that excluded school reading score as a matching variable. These differences have been adjusted as described in Appendix A.4—using the full set of matching variables, but not the reading scores.
- Differences in line (C) are the average differences under revised matching, now adjusted using all matching variables and the reading scores.

The weighted-average difference for the test variable total using lines (B) is about 8% higher than the weighted-average difference using lines (A). The differences in lines (B) completely ignore the reading score covariate—at both the matching and adjustment phase.

The weighted-average difference for the test variable total using lines (C) is about 20% lower than the weighted-average difference using lines (A). The differences in lines (C) are adjusted for all of the matching variables used, including the reading score covariate. Under the revised matching procedure, the averages for all matching variables used are virtually identical for reform and comparison students. The differences in average reading scores for reform and comparison students, however, are substantially greater than those obtained previously and shown in Table 6. (For example, the two largest differences are double the size of the largest difference obtained previously.) Consequently, the adjustments used to generate the differences in lines (C) are principally driven by these differences in average reading scores. Moreover, reform students have higher average reading scores in three of the state/grade combinations, and

[1] In Massachusetts, available data did not permit alternative matching.

negligibly lower scores in the fourth. It is plausible that reading scores have improved as a result of using one of the reform programs. And, if this is so, then the adjustments used to generate lines (C) will have overcompensated for the observed differences in reading scores between reform and comparison students, thereby underestimating the treatment effects.

Appendix B

The ARC Center study includes data from five state-mandated tests administered in the spring of 2000. Table 15 compares the five tests based on the number and type of questions, time limits, and use of calculators and other tools.

Table 15: State Tests Comparison

	MA **grade 4**	**IL** **grade 3**	**IL** **grade 5**	**WA** **grade 3**	**WA** **grade 4**
Name of Test	*Massachusetts Comprehensive Assessment System* (MCAS)	*Illinois Standards Achievement Test* (ISAT)	*Illinois Standards Achievement Test* (ISAT)	*Iowa Test of Basic Skills Form* M (ITBS)	*Washington Assessment of Student Learning* (WASL)
Number of Questions by Type	39 items (29 multiple-choice, 5 open-response, 5 short answer.)	70 multiple choice (2 extended-response questions were on the test, but not included in scoring.)	80 multiple choice (2 extended-response questions were on the test, but not included in scoring.)	90 multiple-choice	40 items (24 multiple-choice, 3 extended response, 13 short answer)[1]
Time Limit	No time limit; administered in 2 sessions	3 35-minute sessions (115 minutes)	3 35-minute sessions (115 minutes)	80 minutes	None
Calculators	None allowed	None allowed	Allowed on 100% of test	None allowed	Allowed on 50% of test
Other Tools	Tool kit supplied, geometric shapes and ruler	None	None	None	Rulers, protractors, pattern blocks, and other manipulatives[2]

The *Illinois Standards Achievement Tes*t (ISAT) is aligned with the *Illinois Learning Standards* that were published in 1997. The ISAT was first administered in 1999. An overall mathematics score is reported as well as results for eight standard sets:

- Estimation/Number Sense/Computation
- Algebraic Patterns/Variables
- Algebraic Relationships/Representations
- Geometric Concepts
- Geometric Relationships
- Measurement
- Data Organization/Analysis
- Probability

[1] Executive Summary
[2] Grade 4 Mathematics Test Specifications – October 2001

Information and sample items are available at www.isbe.state.il.us/assessment/isat.htm.

The *Washington Assessment of Student Learning* (WASL) is designed to measure the mathematics proficiency of students according to the state *Essential Academic Learning Requirements* (EALR). Administration of the test in grade 4 was voluntary in 1997 and required since 1998. A total math score is reported along with scores in the following content and process strands:

- Number Sense
- Measurement Concepts
- Geometric Sense
- Probability and Statistics
- Algebraic Sense
- Solving Problems
- Reasoning Logically
- Communicating Understanding
- Making Connections

Information and sample items are available at www.k12.wa.us

The *Iowa Test of Basic Skills* (ITBS) is a norm-referenced test. Since 1999, grade 3 students in Washington State have taken the ITBS (Form M). National norms for the test were established in 1995. The mathematics portion of the test consists of three sections:

- Math Concepts and Estimation
  The Math Concepts portion includes number properties and operations, algebra, geometry, measurement, and probability and statistics. The Estimation portion measures students' mental arithmetic and estimation skills.
- Math Problem Solving and Data Interpretation
  The Problem Solving and Data Interpretation test includes word problems and interpretation of tables and graphs.
- Math Computation
  Each problem in the Math Computation test requires the use of one arithmetic operation—addition, subtraction, multiplication, or division.

Information on the ITBS is available at www.riverpub.com/products/group/itbs and www.uiowa.edu/~itp/itbs.htm.

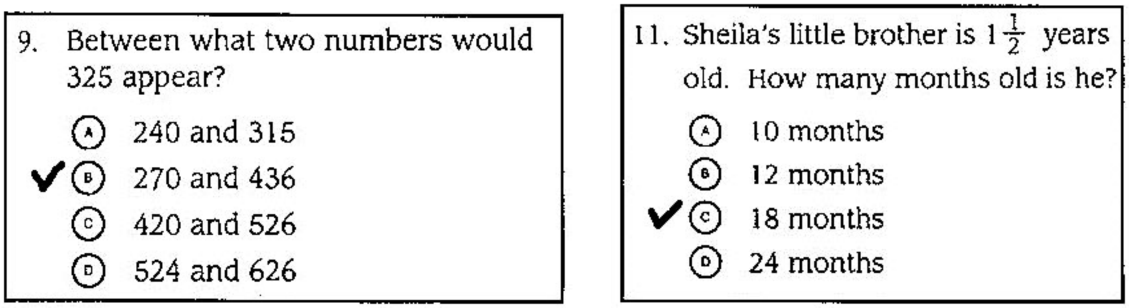
The *Massachusetts Comprehensive Assessment System* (MCAS) was first administered in 1998. Early versions, including the spring 2000 version, were based on the 1996 *Massachusetts Mathematics Curriculum Framework*. The following mathematics content strands were tested:

- Number Sense
- Patterns, Relation, and Functions
- Geometry and Measurement
- Statistics and Probability

The test items from the spring 2000 administration of the MCAS are available at http://www.doe.mass.edu/mcas/2000/release/.

Sample Items

The format of the items on the tests varies widely. As shown in Table 15, all of the tests include multiple-choice items. Figure a shows two sample items from the grade3 ISAT in the Estimation/Number Sense/Computation standard set. Figure b shows a released multiple-choice item from the grade 4 MCAS from the Number Sense reporting category and Fractions and Decimals substrand. Note that students are encouraged to use their toolkit to help solve the problem.



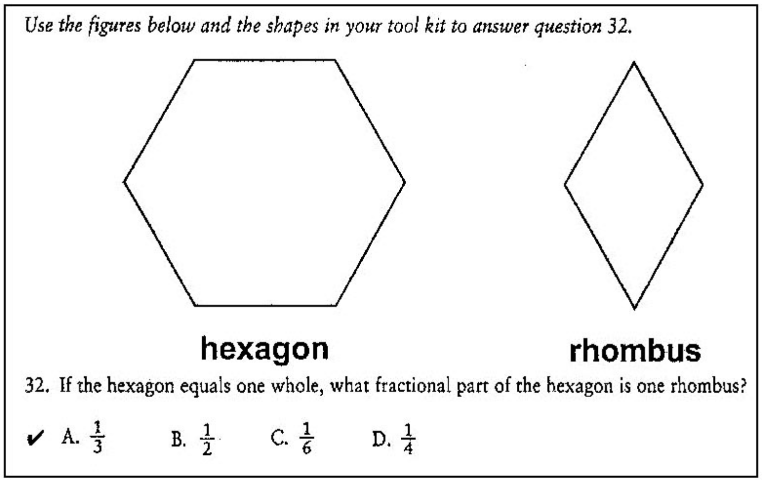
9. Between what two numbers would 325 appear?

Ⓐ 240 and 315
✓ Ⓑ 270 and 436
Ⓒ 420 and 526
Ⓓ 524 and 626

11. Sheila's little brother is $1\frac{1}{2}$ years old. How many months old is he?

Ⓐ 10 months
Ⓑ 12 months
✓ Ⓒ 18 months
Ⓓ 24 months

Figure a. Sample multiple-choice items from the grade 3 ISAT



Use the figures below and the shapes in your tool kit to answer question 32.

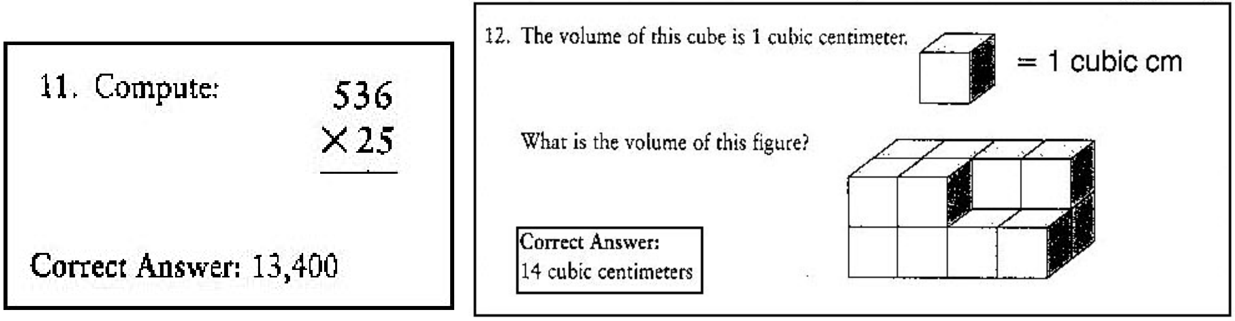
hexagon
rhombus

32. If the hexagon equals one whole, what fractional part of the hexagon is one rhombus?

✓ A. $\frac{1}{3}$ B. $\frac{1}{2}$ C. $\frac{1}{6}$ D. $\frac{1}{4}$

Figure b. Released multiple-choice item from the grade 4 MCAS

The state tests in Washington and Massachusetts included short-answer questions. Figure c shows two short-answer questions from the grade 4 MCAS. One is from the Number Sense reporting category and the Number Computation substrand and the other is from the Geometry and Measurement reporting category and the Measurement substrand.



11. Compute: 536 × 25

Correct Answer: 13,400

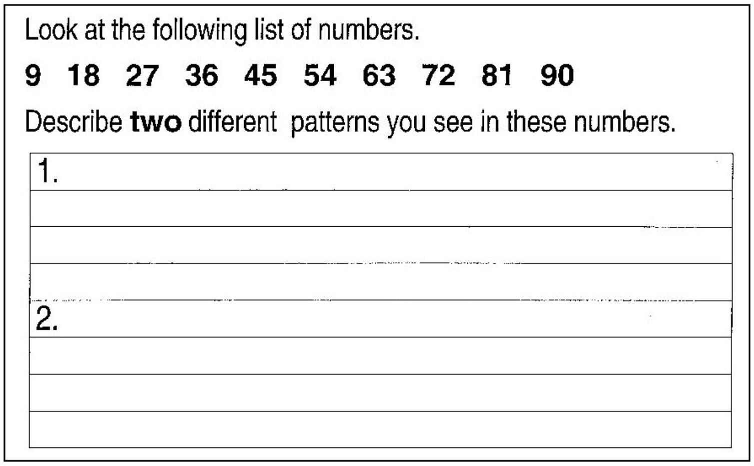
12. The volume of this cube is 1 cubic centimeter.

What is the volume of this figure?

Correct Answer:
14 cubic centimeters

Figure c. Two short-answer questions from the grade 4 MCAS

In 2000, Washington and Massachusetts included results from open-response items in student scores. Figure d is a sample open-response item from the grade 4 WASL. Figure e is the corresponding scoring rubric.



Look at the following list of numbers.

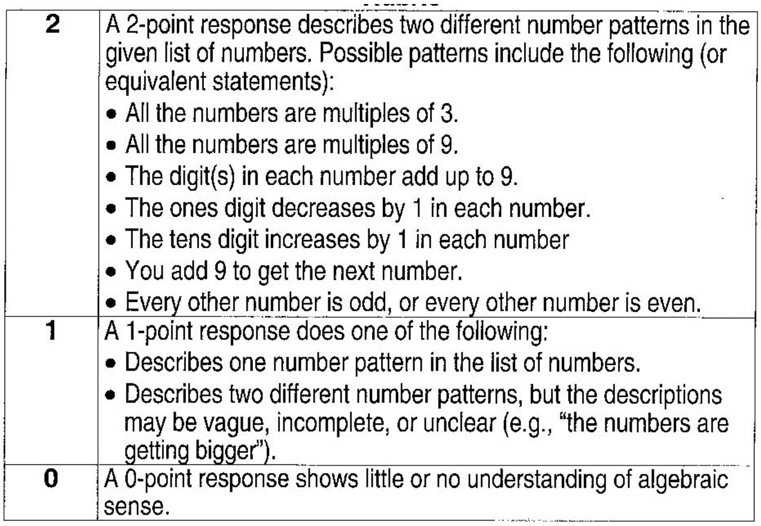
9 18 27 36 45 54 63 72 81 90

Describe **two** different patterns you see in these numbers.

1.

2.

Figure d. Sample open-response item from the grade 4 WASL



2	A 2-point response describes two different number patterns in the given list of numbers. Possible patterns include the following (or equivalent statements): • All the numbers are multiples of 3. • All the numbers are multiples of 9. • The digit(s) in each number add up to 9. • The ones digit decreases by 1 in each number. • The tens digit increases by 1 in each number • You add 9 to get the next number. • Every other number is odd, or every other number is even.
1	A 1-point response does one of the following: • Describes one number pattern in the list of numbers. • Describes two different number patterns, but the descriptions may be vague, incomplete, or unclear (e.g., "the numbers are getting bigger").
0	A 0-point response shows little or no understanding of algebraic sense.

Figure e. Scoring rubric for open-response item from the grade 4 WASL

Appendix C

This Implementation Survey was administered in print and by follow-up phone interview to each school in the study using one of the reform curricula.

To:		**From:**	
		Phone:	
		Fax:	
Phone:		**e-mail**	
Fax:		**Date:**	
		Pages:	

The (name of project) at the (name of institution), with funding from the National Science Foundation and in collaboration with TERC, COMAP, The University of Chicago, and the University of Illinois at Chicago, is carrying out a study of student achievement with (name of curriculum) and two other reform-oriented curricula. As part of our study, we are surveying schools using these curricula.

The survey includes eight questions. We need this information about each school using (name of curriculum). Please look over the questions and answer them as accurately as possible. Thank you.

School: ________________________________

Name and position of person completing survey:

____

Phone number and best time to call:

____

1. What was the primary mathematics program (the instructional materials) used by this school in each grade for mathematics during the 1999–2000 school year?

Grade	Primary Mathematics Program
2	
3	
4	
5	

2. What percentage of teachers for this school used (name of curriculum) or at least 75% of their math instruction during the 1999–2000 school year?

Grade	% of teachers fully using
2	
3	
4	
5	

3. As of June 2000, for this school at each grade, how many years had (name of curriculum) been fully implemented? (Use ‘At least 75% of teachers using the curriculum for at least 75% of their mathematics instruction’ as the definition of ‘full implementation.’)

Grade	Years at full?
2	
3	
4	
5	

4. As of June 2000, at each grade level, what was the total number of hours on average of staff development related to (name of curriculum) for a typical teacher using it (since 1994)?

Grade	# of hours of staff development (circle a range)
2	0–6 7–30 31–99 100+
3	0–6 7–30 31–99 100+
4	0–6 7–30 31–99 100+
5	0–6 7–30 31–99 100+

5. How many actual minutes of mathematics instruction were there per day during the 1999–2000 school year?

Grade	Minutes per day
2	
3	
4	
5	

6. On average, for what percent of math time did teachers use (name of curriculum) during the 1999–2000 school year?

Grade	% of time
2	
3	
4	
5	

7. On average, during the 1999–2000 school year how many units (or lessons, or modules, as appropriate) did teachers complete?

Grade	# units
2	
3	
4	
5	

8. What supplementary mathematics materials (test prep, math facts practice, problem solving, etc.) were used during the 1999–2000 school year?

Grade	What supplementary materials were used?
2	
3	
4	
5	

Table 1: Number of school/grade case records coded for the implementation survey, by state and grade level

	Grade				
State	**2**	**3**	**4**	**5**	**Total**
Illinois	289	290*	275	254*	1,108
Massachusetts	93	159	158*	81	491
Washington	177	178*	178*	125	658
Total	559	627	611	460	2,257

* Student achievement data is available for these five state/grade combinations, which include a total of 1,058 school/grade case records.

Table 2: Number of eligible reform school/grade cases, by state, grade level, and reform program

	Reform Program			
State-Grade Level	**Everyday Mathematics**	**Investigations**	**Math Trailblazers**	**Total**
Illinois-Grade 3	203	0	13	216
Illinois-Grade 5	168	0	6	174
Massachusetts-Grade 4	64	63	0	127
Washington-Grade 3	64	15	34	113
Washington-Grade 4	63	15	34	112
Total	562	93	87	742

Table 3: Exclusions and matching ratio, by state and grade level.

		A	B	C	D	E
State	**Grade level**	**Total schools at grade level**	**Exclusions***	**Total schools available for matching**	**Number of reform schools to be matched**	**Matching ratio**
Illinois	3	2,308	187	1,905	216	8.8 : 1
	5	2,184	191	1,819	174	10.5 : 1
Massachusetts	4	1,051	173	751	127	5.9 : 1
Washington	3	1,150	248	789	113	7.0 : 1
	4	1,149	254	783	112	7.0 : 1
Total		7,842	1,053	6,047	742	

* Includes declared exclusions, including school/grades that were not surveyed, together with surveyed school/grade cases subseque
There are 737 declared excluded school/grades that were not surveyed, representing 9.4% of the total school/grades in Column A

The count in Column A is equal to the sum of the counts in Columns B, C, and D.

The matching ratio in Column E is equal to the ratio of the counts in Columns C and D.

Table 4: Matching variable averages for students in eligible reform school/grades and their comparison school/grades*.

State-Grade Level		**Matching Variable**				
			Average percentage of:			
Illinois-Grade 3		reading score	white	low-income	mobility	LEP
	Reform	165.75	74%	18%	13%	6%
	Comparison	165.85	77%	18%	13%	5%
Illinois-Grade 5		reading score	white	low-income	mobility	LEP
	Reform	164.46	76%	17%	11%	5%
	Comparison	164.2	81%	18%	12%	4%
Massachusetts-Grade 4		reading score	white	free/reduced lunch	mobility**	LEP**
	Reform	236.99	77%	12%	16%	2%
	Comparison	236.28	80%	14%	11%	1%
Washington-Grade 3		reading score	white	TitleI Math	TitleIS	
	Reform	192.07	80%	2%	16%	
	Comparison	191.97	82%	2%	16%	
Washington-Grade 4		reading score	white	TitleI Math	Title1S	
	Reform	412.46	80%	3%	6%	
	Comparison	412.36	82%	3%	6%	

* Averages for comparison students in matched schools are weighted averages. (See Appendix A, Section 3.)

** Variable was tracked but not used in matching .

Table 5: Student record counts at various stages of exclusion, by state and grade level

		A	B	C	D		E*
					missing data		
State	Grade Level	reform and comparison school records on state file prior to exclusions	records after excluding math-disabled/IEP/special education students	records after excluding students with incomplete math test data	reading score	race/ ethnicity	records in final file used for tabulations and constructing case weights
Illinois	3rd	32,923 8.7% are IEP	30,052	29,490	498	1,249	28,992 14,875 reform 14,117 comparison
	5th	30,807 9.9% are IEP	27,763	27,555	74	1,195	27,481 13,820 reform 13,661 comparison
Massachusetts	4th	17,822 15.8% are math-disabled	15,010	14,575	84	94	14,397 6,879 reform 7,518 comparison
Washington	3rd	16,868 9.1% are spec-edn/IEP	15,333	14,794	125	458	14,669 7,813 reform 6,856 comparison
	4th	17,702 11.1% are spec-edn	15,737	15,407	71	274	15,336 7,953 reform 7,383 comparison
Total		116,122	103,895	101,821	852	3,270	100,875 51,340 reform 49,535 comparison

* Excepting Massachusetts, the total count in Column E is equal to the count in Column C minus the count for reading score in Column D. For Massachusetts, the total count in Column E is equal to the count in Column C minus the sum of the counts in Column D.

Table 6: Number of student records used for tabulated comparisons, by state, grade level, school status, and curriculum.

			Reform Program			
State	**Grade Level**	**School Status**	**Everyday Math**	**Investigations**	**Math Trailblazers**	**Total**
Illinois	3rd	Reform	13,840	0	1,035	14,875
		Comparison	13,216	0	901	14,117
	5th	Reform	12,988	0	832	13,820
		Comparison	13,098	0	563	13,661
Massachusetts	4th	Reform	3,962	2,917	0	6,879
		Comparison	4,181	3,337	0	7,518
Washington	3rd	Reform	4,412	916	2,485	7,813
		Comparison	3,923	783	2,150	6,856
	4th	Reform	4,499	920	2,534	7,953
		Comparison	4,063	907	2,413	7,383
Totals		Reform	39,701	4,753	6,886	51,340
		Comparison	38,481	5,027	6,027	49,535
			78,182	9,780	12,913	100,875

Table 7: Average differences and effect sizes, by state/grade combination.

		math	total	computation	measurement	geometry	prob/stat	algebra				
IL grade3	difference	1.39***	1.82***	2.78***	3.84***	0.76***	0.06	1.44***				
(n=14,875)	effect size	0.098	0.099	0.141	0.164	0.038	0.003	0.073				
IL grade5	difference	1.82***	2.20***	2.29***	3.02***	3.26***	1.55***	1.44***				
(n=13,820)	effect size	0.121	0.116	0.117	0.132	0.165	0.079	0.067				
									open response	short answer	multiple choice	
MA grade 4	difference	1.34***	1.33***	2.36***		-0.19	-0.16	3.07***	2.46***	-0.62	0.89***	
(n=6,879)	effect size	0.087	0.078	0.127		-0.010	-0.008	0.137	0.119	-0.024	0.053	
									problem solving	concepts/ estimation		
WA grade3	difference	1.34***	1.27***	0.74**					0.76**	1.86***		
(n=7,813)	effect size	0.073	0.078	0.039					0.036	0.108		
										logic	commun- icating	making connections
WA grade4	difference	3.02***	1.77***	1.02***	3.43***	1.61***	0.00	2.99***	2.51***	1.17***	-0.03	3.55***
(n=7,953)	effect size	0.093	0.093	0.041	0.120	0.078	0.000	0.112	0.090	0.040	-0.001	0.116
Combined	effect size	0.098***	0.097***	0.102***	0.142***	0.078***	0.025***	0.088***				
(n=51,340)	percentile change	+3.92%	+3.88%	+4.08%	+5.68%	+3.12%	+1.00%	+3.52%				

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate strand portion of test.

The record counts in column one are the numbers of reform-student records used for tabulations. For any given tabulation, the weighted number of comparison-student records used is equal to the number of reform-student records used.

Two-sided significance levels are defined as follows: *** is $p < 0.001$, ** is $p < 0.01$, * is $p < 0.025$.

Table 8: Average effect sizes and percentile changes, by student race/ethnicity

		math	total	computation	measurement	geometry	prob/stat	algebra
Asian	effect size	0.106***	0.115***	0.097***	0.175***	0.162***	0.043	0.086***
(n=3,071)	percentile change	+4.24%	+4.60%	+3.88%	+7.00%	+6.48%	+1.72%	+3.44%
Black	effect size	0.092***	0.101***	0.109***	0.129***	0.081***	0.029	0.087***
(n=3,509)	percentile change	+3.68%	+4.04%	+4.36%	+5.16%	+3.24%	+1.16%	+3.48%
Hispanic	effect size	0.021	0.031	0.017	0.094***	0.049*	-0.005	0.035
(n=3,002)	percentile change	+0.84%	+1.24%	+0.68%	+3.76%	+1.96%	-0.20%	+1.40%
White	effect size	0.100***	0.100***	0.106***	0.144***	0.070***	0.020***	0.091***
(n=37,609)	percentile change	+4.00%	+4.00%	+4.24%	+5.76%	+2.80%	+0.80%	+3.64%
Combined	effect size	0.098***	0.097***	0.102***	0.142***	0.078***	0.025***	0.088***
(n=51,340)	percentile change	+3.92%	+3.88%	+4.08%	+5.68%	+3.12%	+1.00%	+3.52%

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate stand portion of test

comparison-student

records used is roughly equal to the number of reform-student records used.

subsequently imputed.

The tabulations for geometry, prob/stat and algebra are based on 5-10% fewer student records, because these strands are not separately scored in the WA (grade 3) test.

the WA (grade 3) test.

Two-sided significance levels are defined as follows: *** is $p < 0.001$, ** is $p < 0.01$, * is $p < 0.025$

Table 9: Average effect sizes and percentile changes, by school SES and Title1S status

			math	total	computation	measurement	geometry	prob/stat	algebra
Illinois and Massachusetts	SES low	effect size	0.114***	0.102***	0.103***	0.144***	0.152***	0.027*	0.072***
	(n=9,723)	percentile change	+4.56%	+4.08%	+4.12%	+5.76%	+6.08%	+1.08%	+2.88%
	SES middle	effect size	0.077***	0.078***	0.124***	0.110***	0.000	0.004	0.081***
	(n=8,476)	percentile change	+3.08%	+3.12%	+4.96%	+4.40%	+0.00%	+0.16%	+3.24%
	SES top	effect size	0.101***	0.108***	0.151***	0.150***	0.046***	0.039***	0.081***
	(n=17,375)	percentile change	+4.04%	+4.32%	+6.04%	+6.00%	+1.84%	+1.56%	+3.24%
Washington	TitleIS	effect size	0.096***	0.094***	0.046	0.065 #	0.032 #	-0.026 #	0.087 #
	(n=1,793)	percentile change	+3.84%	+3.76%	+1.84%	+2.60%	+1.28%	-1.04%	+3.48%
	non-TitleIS	effect size	0.083***	0.085***	0.039***	0.125***	0.082***	0.003	0.116***
	(n=13,973)	percentile change	+3.32%	+3.40%	+1.56%	+5.00%	+3.28%	+0.12%	+4.64%
Combined	(n=51,340)	effect size	0.098***	0.097***	0.102***	0.142***	0.078***	0.025***	0.088***
		percentile change	+3.92%	+3.88%	+4.08%	+5.68%	+3.12%	+1.00%	+3.52%

Math is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate stand portion of test.

The record counts in column two are the numbers of reform-student records used for tabulations. For any given tabulation, the weighted number of comparison-student records used is equal to the number of reform-student records used.

For IL and MA, the tabulations for measurement are based on approximately 20% fewer student records because this strand is not scored separately by MA.

# Statistics are based on only n=507 grade 4 reform-student records and an equal number of comparison-student records..
For WA, the tabulations for measurement, geometry, prob/stat, and algebra are based only on grade 4 student records. Thus, for these strands, the tabulations use only 507 reform-student records for TitleIS schools, and 7,446 reform-student records for non-TitleIS schools.

Two-sided significance levels are defined as follows: *** is p < 0.001, ** is p < 0.01, * is p < 0.025

Table 10: Average effect sizes and percentile changes, by student gender

		math	total	computation	measurement	geometry	prob/stat	algebra
Female	effect size	0.097***	0.097***	0.105***	0.144***	0.081***	0.023***	0.089***
(n=25,628)	percentile change	+3.88%	+3.88%	+4.20%	+5.76%	+3.24%	+0.92%	+3.56%
Male	effect size	0.096***	0.094***	0.098***	0.140***	0.074***	0.024***	0.084***
(n=24,933)	percentile change	+3.84%	+3.76%	+3.92%	+5.60%	+2.96%	+0.96%	+3.36%
Combined	effect size	0.098***	0.097***	0.102***	0.142***	0.078***	0.025***	0.088***
(n=51,340)	percentile change	+3.92%	+3.88%	+4.08%	+5.68%	+3.12%	+1.00%	+3.52%

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate stand portion of test

The record counts in column one are the numbers of reform-student records used for tabulations. For any given tabulation, the weighted number of comparison-student records used is roughly equal to the number of reform-student records used.
missing.

The tabulations for geometry, prob/stat and algebra are based on 15% fewer student records, because these strands are not separately scored in the WA (grade 3) test.

The tabulations for measurement are based on 30% fewer student records, because this strand is scored separately by neither the MA (grade 4) test nor the WA (grade 3) test.

Two-sided significance levels are defined as follows: *** is p < 0.001, ** is p < 0.01, * is p < 0.025

Table 11: Averages, by state/grade combination and reform status

State- Grade Level	Reform Status	Student Records	math	total	computation	measurement	geometry	prob/stat	algebra				
IL grade3	reform	14,875	167.18	70.22	68.95	67.78	73.26	74.64	66.44				
	comparison	14,117	165.79	68.40	66.17	63.94	72.50	74.58	65.00				
	non-reform	115,053	160.84	61.62	59.47	56.42	65.58	67.27	59.38				
IL grade5	reform	13,820	170.10	66.29	65.08	61.03	73.01	68.49	63.85				
	comparison	13,661	168.28	64.09	62.79	58.01	69.75	66.94	62.41				
	non-reform	116,316	162.19	56.34	55.03	50.60	62.33	58.82	54.56				
										open response	short answer	multiple choice	
MA grade 4	reform	6,879	244.13	66.88	70.13		62.05	72.16	61.50	60.32	59.65	72.65	
	comparison	7,518	242.79	65.55	67.77		62.24	72.32	58.43	57.86	60.27	71.76	
	non-reform	58,716	236.52	57.37	*		*	*	*	*	*	*	
										problem solving	concepts/ estimation		
WA grade3	reform	7,813	194.00	71.17	72.53					66.52	72.42		
	comparison	6,856	192.66	69.90	71.79					65.74	70.56		
	non-reform	59,021	189.58	67.18	69.74					61.34	67.83		
											logic	commun- icating	making connections
WA grade4	reform	7,953	401.60	56.62	55.22	67.26	60.58	62.64	69.50	38.82	57.62	45.03	61.02
	comparison	7,383	398.58	54.85	54.20	63.83	58.97	62.64	66.51	36.31	56.45	45.06	57.47
	non-reform	60,656	393.57	51.93	49.93	58.80	55.17	59.25	61.60	32.22	51.34	40.27	52.86

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate strand portion of test.
Averages for non-reform students exclude all IEP, mathematically-disabled, and special-education students.

* Data for individual strands and test item categories was available only for reform and comparison schools.

Table 12: Variables* used in matching and for subgroup analyses, by state/grade source.

	Illinois		Massachusetts	Washington	
Variable	**Grade 3 (ISAT)**	**Grade 5 (ISAT)**	**Grade 4 (MCAS)**	**Grade 3 (ITBS)**	**Grade 4 (WASL)**
reading score	student	student	student	student	student
gender	student	student	student	student	student
race	student	student	student	student	student
white (calculated from race)	student	student	student	student	student
LEP (Limited English proficiency status)	school	school	student		
mobility (school mobility rate)	school	school	school		
low family income status:					
low-income	school	school			
free/reduced lunch			student		
TitleI Mathematics				student	student
TitleIS				school	school
SES (calculated using low-income and free/reduced lunch [MA])	school	school	school		

* A variable is classified as a "student" variable if individual student-level data was available.
A variable is classified as a "school" variable if only school-level data was available.

For the matching procedure used to identify comparison schools, only school-level data were used. Student-level data for reading score, white, free/reduced lunch (in MA), LEP (in MA), and TitleI Mathematics (in WA) were all aggregated within schools to create school average percents that were used in matching.

Table 13. Estimated bias due to regression artifacts.
(Typical results based on IL-grade 5 data.)

		Assumed covariate reliabilities: read=0.99, lowincom=0.98, white=0.98	Assumed covariate reliabilities: read=0.97, lowincom=0.92, white=0.92
Test variable modeled	Covariates used	Estimated bias	Estimated bias
math	reading score	+4.2%	+8.2%
math	low-income %	+1.2%	+21.9%
math	reading score low-income %	+4.1%	+5.6%
math*	reading score low-income % white %	+6.2%	+3.0%
computation	reading score low-income %	+2.1%	+3.8%
algebra	reading score low-income %	+6.3%	+13.4%
algebra	reading score low-income % white %	-3.2%	+15.4%

* Math was also modeled using these same covariates, but with unrealistically low assumed reliabilities: read=0.90, lowincom=0.90, and white=0.90. The estimated bias was +38.3%.

Table 14: Average differences, by matching and adjustment procedure.

	Matching/ Adjustment procedure	Differences										
		math	total	computation	measurement	geometry	prob/stat	algebra				
IL grade3	(A)	1.39***	1.82***	2.78***	3.84***	0.76***	0.06	1.44***				
(n=14,875)	(B)	1.24***	1.64***	2.57***	3.67***	0.46**	-0.16	1.46***				
	(C)	1.27***	1.65***	2.56***	3.73***	0.51**	-0.13	1.41***				
IL grade5	(A)	1.82***	2.20***	2.29***	3.02***	3.26***	1.55***	1.44***				
(n=13,820)	(B)	1.81***	2.22***	2.38***	3.10***	2.91***	1.50***	1.66***				
	(C)	1.34***	1.66***	1.81***	2.55***	2.35***	1.06***	1.02***				
									problem solving	concepts/ estimation		
WA grade3	(A)	1.34***	1.27***	0.74**					0.76**	1.86***		
(n=7,813)	(B)	1.84***	1.68***	0.76***					1.66***	2.27***		
	(C)	0.96**	0.91***	0.16					0.68**	1.51***		
										logic	commun-icating	making connections
WA grade4	(A)	3.02***	1.77***	1.02***	3.43***	1.61***	0.00	2.99***	2.51***	1.17***	-0.03	3.55***
(n=7,953)	(B)	3.86***	2.28***	1.84***	3.73***	1.73***	-0.21	4.36***	2.82***	1.89***	0.73*	3.67***
	(C)	2.35***	1.39***	0.99**	2.80***	1.14***	-0.80*	3.42***	1.76***	0.86*	-0.13	2.71***

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate strand portion of test.
The record counts in column one are the numbers of reform-student records used for tabulations. For any given tabulation, the weighted number of comparison-student records used is equal to the number of reform-student records used.

"Matching/adjustment procedure" (column 2) is defined as follows:
Lines (A) correspond to full matching and adjustment using all covariates, including reading score. Lines (B) correspond to revised matching and adjustment; reading score is not used in either phase. Lines (C) correspond to revised matching; reading score is used, however, in the adjustment phase.

Two-sided significance levels are defined as follows: *** is p < 0.001, ** is p < 0.01, * is p < 0.025 .