ISSUES & ANSWERS

REL 2007–No. 039

REL
NORTHEAST
& ISLANDS
Regional Educational Laboratory
At Education Development
Center, Inc.

Measuring how benchmark assessments affect student achievement

ies NATIONAL CENTER FOR
EDUCATION EVALUATION
AND REGIONAL ASSISTANCE

Institute of Education Sciences

U.S. Department of Education

ISSUES & ANSWERS

REL 2007–No. 039

REL NORTHEAST & ISLANDS
Regional Educational Laboratory
At Education Development
Center, Inc.

Measuring how benchmark assessments affect student achievement

December 2007

Prepared by

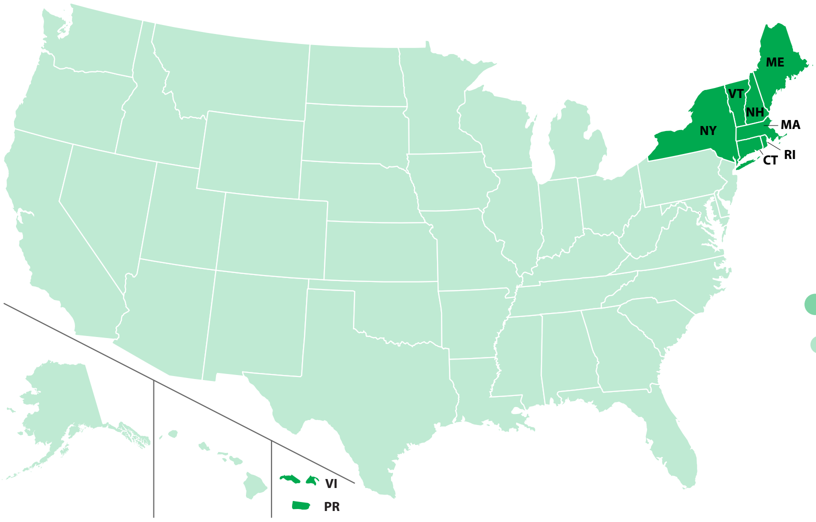
Susan Henderson
WestEd

Anthony Petrosino
WestEd

Sarah Guckenburg
WestEd

Stephen Hamilton
WestEd

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education



REL
NORTHEAST
& ISLANDS
Regional Educational Laboratory
At Education Development
Center, Inc.

Issues & Answers is an ongoing series of reports from short-term Fast Response Projects conducted by the regional educational laboratories on current education issues of importance at local, state, and regional levels. Fast Response Project topics change to reflect new issues, as identified through lab outreach and requests for assistance from policymakers and educators at state and local levels and from communities, businesses, parents, families, and youth. All Issues & Answers reports meet Institute of Education Sciences standards for scientifically valid research.

December 2007

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-06-CO-0025 by Regional Educational Laboratory Northeast and Islands administered by Education Development Center, Inc. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

This report is in the public domain. While permission to reprint this publication is not necessary, it should be cited as:

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (Issues & Answers Report, REL 2007–No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

This report is available on the regional educational laboratory web site at http://ies.ed.gov/ncee/edlabs.

Summary

Measuring how benchmark assessments affect student achievement

This report examines a Massachusetts pilot program for quarterly benchmark exams in middle-school mathematics, finding that program schools do not show greater gains in student achievement after a year. But that finding might reflect limited data rather than ineffective benchmark assessments.

Benchmark assessments are used in many districts throughout the nation to raise student, school, and district achievement and to meet the requirements of the No Child Left Behind Act of 2001. This report details a study using a quasi-experimental design to examine whether schools using quarterly benchmark exams in middle-school mathematics under a Massachusetts pilot program show greater gains in student achievement than schools not in the program.

To measure the effects of benchmark assessments, the study matched 44 comparison schools to the 22 schools in the Massachusetts pilot program on pre-implementation test scores and other variables. It examined descriptive statistics on the data and performed interrupted time series analysis to test causal inferences.

The study found no immediate statistically significant or substantively important difference between the program and comparison schools. That finding might, however, reflect limitations in the data rather than the ineffectiveness of benchmark assessments.

First, data are lacking on what benchmark assessment practices comparison schools may be using, because the study examined the impact of a particular structured benchmarking program. More than 70 percent of districts are doing some type of formative assessment, so it is possible that at least some of the comparison schools implemented their own version of benchmarking. Second, the study was “underpowered.” That means that a small but important treatment effect for benchmarking could have gone undetected because there were only 22 program schools and 44 comparison schools. Third, with only one year of post-implementation data, it may be too early to observe any impact from the intervention in the program schools.

Although the study did not find any immediate difference between schools employing benchmark assessments and those not doing so, it provides initial empirical data to inform state and local education agencies.

The report urges that researchers and policymakers continue to track achievement data in the program and comparison schools, to reassess the initial findings in future years, and to provide additional data to local and state decisionmakers about the impact of this benchmark assessment practice.

Using student-level data rather than school-level data might help researchers examine the impact of benchmark assessments on important No Child Left Behind subgroups (such as minority students or students with disabilities). Some nontrivial effects for subgroups might be masked by comparing school mean scores. (At the onset of the study, only school-level data were available to researchers.)

Another useful follow-up would be disaggregating the school achievement data by mathematics content strand to see if there are any effects in particular standards. Because the quarterly assessments are broken out by mathematics content strand, doing so would connect logically with the benchmark assessment strategy. This refined data analysis might be more sensitive to the intervention and might also be linked to information provided to the Massachusetts Department of Education about which content strands schools focused on in their benchmark assessments.

Conversations with education decisionmakers support what seems to be common sense. Higher mathematics scores will come not because benchmarks exist but because of how a school's teachers and leaders use the assessment data. This kind of follow-up research, though difficult, is imperative to better understand the impact of benchmark assessments. A possible approach is to examine initial district progress reports for insight into school buy-in to the initiative, quality of leadership, challenges to implementation, particular standards that participating districts focus on, and how schools use the benchmark assessment data.

December 2007

TABLE OF CONTENTS

Boxes

Figures

Tables

This report examines a Massachusetts pilot program for quarterly benchmark exams in middle-school mathematics, finding that program schools do not show greater gains in student achievement after a year. But that finding might reflect limited data rather than ineffective benchmark assessments.

OVERVIEW

Benchmark assessments are used in many districts throughout the United States to raise student, school, and district achievement and to meet the requirements of the No Child Left Behind Act of 2001 (see box 1 on key terms). This report details a study using a quasi-experimental design to examine whether schools using quarterly benchmark exams in middle-school mathematics under a Massachusetts pilot program show greater gains in student achievement than schools not in the program.

To measure the effects of benchmark assessments, the study matched 44 comparison schools to the 22 program schools in the Massachusetts pilot program on pre-implementation test scores and other variables. It examined descriptive statistics on the data and performed interrupted time series analysis to test causal inferences.

The study found no immediate statistically significant or substantively important difference between the program and comparison schools a year after the pilot began. That finding might, however, reflect limitations in the data rather than the ineffectiveness of benchmark assessments.

DATA ON THE EFFECTIVENESS OF BENCHMARK ASSESSMENTS ARE LIMITED

Benchmark assessments align with state standards, are generally administered three or four times a year, and provide educators and administrators with immediate student-level data connected to individual standards and content strands (Herman & Baker, 2005; Olson, 2005). Benchmark assessments are generally regarded as a promising practice. A U.S. Department of Education report (2007) notes that "regardless of their specific mathematics programs, No Child Left Behind Blue Ribbon Schools . . . [all] emphasize alignment of the school's mathematics curriculum with state standards and conduct frequent benchmark assessments to determine student mastery of the standards."

By providing timely information to educators about student growth on standards, such

BOX 1

Key terms used in the report

Benchmark assessment. A benchmark assessment is an interim assessment created by districts that can be used both formatively and summatively. It provides local accountability data on identified learning standards for district review after a defined instructional period and provides teachers with student outcome data to inform instructional practice and intervention before annual state summative assessments. In addition, a benchmark assessment allows educators to monitor the progress of students against the state standards and to predict performance on state exams.

Content strand. The Massachusetts Curriculum Frameworks contain five content strands that are assessed through the Massachusetts Comprehensive Assessment System: number sense and operation; patterns, relations, and algebra; geometry; measurement; and data analysis, statistics, and probability.

Effect size. An effect size of 0.40 means that the experimental group is performing, on average, about 0.40 of a standard deviation better than the comparison group (Valentine and Cooper, 2003). An effect size of 0.40 represents a roughly 20 percent improvement over the comparison group.

Formative assessment. In this study a formative assessment is an assessment whose data are used to inform instructional practice within a cycle of learning for the students assessed. In September 2007 the Formative Assessment for Students and Teachers study group of the Council of Chief State School Officers' Assessment for Learning further refined the definition of formative assessment as "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (see http://www.ccsso.org/projects/scass/Projects/Formative_Assessment_for_Students_and_Teachers/).

Interrupted time series analysis. An interrupted time series analysis is a series of observations made on one or more variables over time before and after the implementation of a program or treatment (Shadish, Cook, & Campbell, 2002).

Quasi-experimental design. A quasi-experimental design is an experimental design where units of study are not assigned to conditions randomly (Shadish, Cook, & Campbell, 2002).

Scaled scores. Scaled scores are constructed by converting students' raw scores (say, the number of questions correct) on a test to yield comparable results across students, test versions, or time.

Statistical power. Statistical power refers to the ability of the statistical test to detect a true treatment effect, if one exists. Although there are other design features that can influence the statistical power of a test, researchers are generally most concerned with sample size, because it is the component they have the most control over and can normally plan for.

Summative assessment. A summative assessment is designed to show the extent to which students understand the skills, objectives, and content of a program of study. The assessments are administered after the opportunity to learn subject matter has ended, such as at the end of a course, semester, or grade.

Underpowered study. A study is considered underpowered if, all else being equal, it lacks a sufficient sample size to "detect" a small but nontrivial treatment effect. An underpowered study would lead researchers to report that such a small but nontrivial difference was not statistically significant.

assessments allow instructional practices to be modified to better meet student needs. Benchmark assessments fill a gap left by annual state tests, which often provide data only months after they are administered and whose purpose is largely summative (Herman & Baker, 2005; Olson, 2005). A 2005 *Education Week* survey of superintendents found that approximately 70 percent reported using benchmark assessments in their districts (Olson). But there is little empirical evidence to determine whether and to what extent these aligned benchmark assessments affect student outcomes. This report provides evidence on the impact of a Massachusetts Department of Education benchmark assessment initiative targeting high-poverty middle schools.

Studies of benchmark assessments' effects on student outcomes are few. But the substantial literature on the effects of formative assessments more generally points consistently to the positive effects of formative assessment on student learning (Black & Wiliam, 1998a, 1998b; Bloom, 1984). Reviewing 250 studies of classroom formative assessments, Black and Wiliam (1998a, 1998b) find that formative assessments, broadly defined, are positively correlated with student learning, boosting performance 20–40 percent over that of comparison groups (with effect sizes from 0.40 to 0.70).[1] Black and Wiliam note that these positive effects are even larger for low-achieving students than for the general student population. Other studies indicate that formative assessments can support students and teachers in identifying learning goals and the instructional strategies to achieve them (Boston, 2002). Whether these trends hold for benchmark assessments, however, has yet to be shown.

Making this report particularly timely are the widespread interest in the Northeast and Islands Region in formative assessment and systems to support it and the piloting of a benchmark assessment approach to mathematics in Massachusetts middle schools. State education agencies in New York, Vermont, and Connecticut are also working with federal assessment and accountability centers and regional comprehensive centers to pilot formative and benchmark assessment practices in select districts. And the large financial investment required for the data management systems to support this comprehensive approach underscores the need for independent data to inform state and district investment decisions.

The 2005 Massachusetts Comprehensive School Reform and the Technology Enhancement Competitive grant programs include priorities for participating schools and districts to develop and use benchmark assessments. As a result, eight Massachusetts school districts use a data management system supported by Assessment Technologies Incorporated to develop their own grade-level benchmark assessments in mathematics for about 10,000 middle-school students in 25 schools. The decision of the Massachusetts Department of Education to support the development of mathematics benchmark assessments in a limited number of middle schools provided an opportunity to study the effects on student achievement.

There are few studies of benchmark assessments' effects on student outcomes, but the substantial literature on the effects of formative assessments more generally points consistently to the positive effects on student learning

This report details a study on whether schools using quarterly benchmark exams in middle-school mathematics under the Massachusetts pilot program show greater gains in student achievement after one year than schools not in the program. The study looked at 44 comparison schools and 22 program schools using quarterly benchmark assessments aligned with Massachusetts Curriculum Frameworks Standards for mathematics in grade 8, with student achievement measured by the Massachusetts Comprehensive Assessment System (MCAS).

FEW EFFECTS FROM BENCHMARK ASSESSMENTS ARE EVIDENT AFTER ONE PROGRAM YEAR

The study was designed to determine whether there was any immediate, discernible effect on eighth-grade mathematics achievement from using benchmark assessments in middle schools receiving the Comprehensive School Reform grants. An advantage of the study's achievement data was that they went beyond a single pretest year and included scores from five prior annual administrations of the MCAS, yielding five pre-implementation years for eighth-grade mathematics scores. A disadvantage of the data was that they contain only one post-test year.[2] Even so, the data could show whether there was any perceptible, immediate increase or decrease in scores due to the implementation of benchmark assessments.

BOX 2

Methodology

A quasi-experimental design, with program and matched comparison schools, was used to examine whether schools using quarterly benchmark exams in middle-school mathematics under the Massachusetts pilot program showed greater gains in student achievement in mathematics performance after one year than schools not in the program. Analyses were based on (mostly) publicly available,[1] school achievement and demographic data maintained by the Massachusetts Department of Education. The primary outcome measure was eighth-grade mathematics achievement, as assessed by the Massachusetts Comprehensive Assessment System (MCAS).

Defining the program

The study defined benchmark assessments as assessments that align with the Massachusetts Curriculum Frameworks Standards, are administered quarterly at the school level, and yield student-level data—immediately available to school educators and administrators—aligned with individual standards and content strands. For the benchmark assessment initiative examined in the report, the Massachusetts Department of Education selected high-poverty middle schools under pressure to significantly improve their students' mathematics achievement, choosing 25 schools in eight districts to participate in the pilot initiative.

Constructing the study database and describing the variables

Data were collected from student- or school-level achievement and demographic data maintained by the Massachusetts Department of Education.[2] The outcome variable was scaled eighth-grade MCAS mathematics scores over 2001–06. The MCAS, which fulfills the requirements of the No Child Left Behind Act of 2001 requiring annual assessments in reading and mathematics for students in grades 3–8 and in high school, tests all public school students in Massachusetts.

Other variables gathered for the study included the school name, location, grade structure, and enrollment; the race and ethnicity of students; and the proportion of limited English proficiency and low-income students.

Creating a comparison group

Only a well implemented randomization procedure controls for both known and unknown factors that could influence or bias the findings. But because the grants to implement the benchmark assessments were already distributed and the program was already administered to schools, random assignment was not possible. So, it was necessary to use other procedures to create a counterfactual—a set of schools that did not receive the program.

The study used covariate matching to create a set of comparison schools that was as similar as possible to the program schools (in the aggregate) on the chosen factors, meaning that any findings, whether positive or negative, would be unlikely to have been influenced by those factors. These variables included enrollment, percentage of students classified as low income, percentage of students classified as English language learners, and percentage of students categorized in different ethnic groups. Also included were each school's eighth-grade baseline (or pretest) mathematics score (based on an average of its 2004/05 eighth-grade mathematics scores) and the type of location it served.

Prior research guided the selection of the variables used as covariates in the matching. Bloom (2003) suggests that pretest scores are perhaps the most important variable to use in a matching procedure. There is also substantial research that identifies large gaps in academic achievement for racial minorities (Jencks & Phillips, 1998), low-income students (Hannaway, 2005), and English language learners (Abedi & Gandara, 2006). Although the research on the relationship between school size and academic achievement is somewhat conflicting (Cotton, 1996), the variability in school size resulted in total enrollment in the middle school being included in the matching procedure.

The eligibility pool for the comparison matches included the 389 Massachusetts middle schools that did not receive the Comprehensive School Reform grants. Statistical procedures were used to identify the two best matches for each program school from the eligibility pool. The covariate matching resulted in a final sample of 22 program schools and 44 comparison schools that were nearly identical on pretest academic scores. The project design achieved balance on nearly all school-level social and demographic characteristics, except that there were larger shares of African American and Pacific Islander students in program schools. These differences were controlled for statistically in the outcome analysis, with no change in the results (see appendix D).

Analyzing the data

After matching, descriptive statistics were used to examine the mean scores for all five pre-implementation years and one post-implementation year for the program and comparison schools. A comparative interrupted time series analysis was also used to more rigorously assess whether there was a statistically significant difference between the program and comparison schools in changes in mathematics performance (see Bloom, 2003; Cook & Campbell, 1979). The interrupted time series design was meant to determine whether there was any change in the trend because of the "interruption" (program implementation).

Notes

1. The 2001–03 achievement data were not publicly available and had to be requested from the Massachusetts Department of Education.
2. Student level data for 2001–03 had to be aggregated at the school level. The 2001–03 achievement data were provided by the Massachusetts Department of Education at the student level, but were not linked to the student level demographic data for the same years.

To attribute changes to benchmark assessments, more information than pretest and post-test scores from program schools was needed. Did similar schools, not implementing the benchmarking practice, fare better or worse than the 22 program schools? It could be that the program schools improved slightly, but that similar schools not implementing the benchmark assessment practice did much better or much worse. So achievement data were also examined from comparison schools—a set of similar Massachusetts schools that did not implement the program (for details on methodology, see box 2 and appendixes A and B; for details on selecting comparison schools, see box 2 and appendix C).

Researchers developed a set of 44 comparison middle schools in Massachusetts that were very similar to the program schools (in the aggregate) on a number of variables. Most important, the comparison schools were nearly identical to the program schools on the pre-implementation scores. The 44 comparison schools thus provided an opportunity to track the movement of eighth-grade mathematics scores over the period in the absence of the program.

. . . using descriptive statistics

Scaled scores for program and comparison schools from 2001 to 2006 did not show a large change in eighth-grade MCAS scores for either program or comparison schools (table 1).[3] Note that scaled scores for both groups were distributed in the MCAS "needs improvement" category for all years—further evidence to support the validity of the matching procedure.[4]

TABLE 1
Scaled eighth-grade mathematics scores for program and comparison schools in the Massachusetts Comprehensive Assessment System, 2001–06

Year	Program schools	Comparison schools
2001	224.80	226.31
2002	223.21	223.28
2003	224.81	224.09
2004	226.10	225.32
2005	225.62	225.23
2006	226.98	226.18

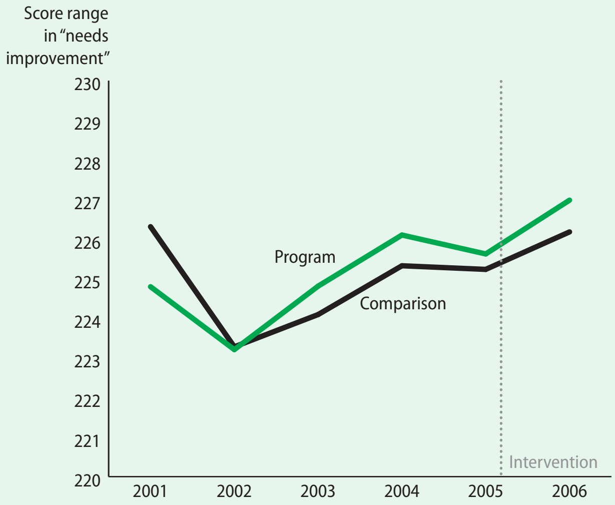
Note: Scaled scores are constructed by converting students' raw scores (say, the number of questions answered corrrectly) on a test to yield comparable results across students, test versions, or time. Scores for both groups are distributed in the Massachusetts Comprehensive Assessment System "needs improvement" category for all years.

Source: Authors' analysis based on data described in text.

There appeared to be a very slight uptick in eighth-grade mathematics outcomes after the intervention in 2006. There was, however, a similar increase in 2004, before the intervention. And trends were similar for the program and comparison groups. In both, there was a very slight increase on the outcome measure, but similar increases occurred before the 2006 intervention (figure 1). So, the descriptive statistics showed no perceptible difference between the 22 program schools and the 44 comparison schools on their 2006 eighth-grade mathematics outcomes.

FIGURE 1

Scaled eighth-grade mathematics scores for program and comparison schools in the Massachusetts Comprehensive Assessment System, 2001–06

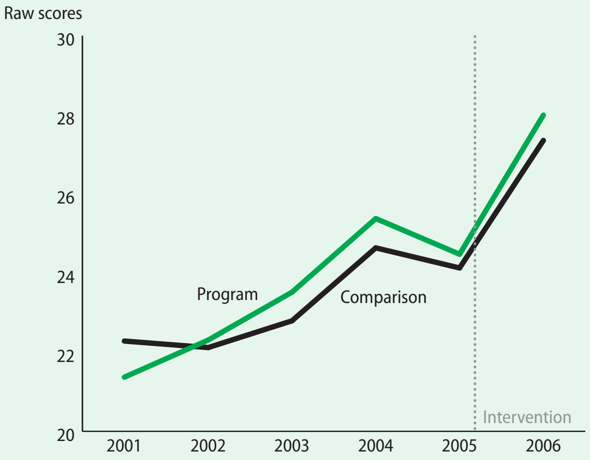


Note: Scaled scores are constructed by converting students' raw scores (say, the number of questions correct) on a test in order to yield comparable results across students, test versions, or time.

Source: Authors' analysis based on data described in text.

FIGURE 2

Raw eighth-grade mathematics scores for program and comparison schools in the Massachusetts Comprehensive Assessment System, 2001–06



Source: Authors' analysis based on data described in text.

The study also examined raw scores because it is possible that scaling test scores could mask effects over time. The range in raw scores was larger, and scores trended sharply higher in 2006. But again both program and comparison schools showed a similar trend, more sharply upward than that of the scaled scores (figure 2).

. . . or interrupted time series analysis

Relying on such strategies alone was not adequate to rigorously assess the impact of benchmark assessment. To assess the differences between program and comparison schools in changes in mathematics performance, the study used interrupted time series analysis, which established the pre-intervention trend in student performance and analyzed the post-intervention data to determine whether there was a departure from that trend (Bloom, 2003; see appendix D for details). Five years of annual pre-implementation data and a year of post-implementation data formed the time series. The program schools' implementation of the benchmark assessment practice in 2006 was the intervention, or "interruption."

There was a small but statistically significant increase in the program schools in 2006. The program schools had slightly higher mean eighth-grade mathematics scores than what would have been expected without the program. But this small, statistically significant increase also occurred in the comparison schools, where mean mathematics scores were slightly above the predicted trend.

Difference-in-difference analysis underscored the similarity between the groups. The program effect was about 0.38 of a mathematics test point (see appendix D, table D4), but it was not statistically significant. The most likely interpretation is that the achievement of both groups was slightly increasing and that the difference between them could have been due to chance rather than to any program effect. So, though both groups of schools saw similar, (slightly) higher than expected increases in their eighth-grade mathematics scaled scores in 2006, the small increase for the program schools cannot be attributed to the benchmark assessments.

WHY WEREN'T EFFECTS EVIDENT AFTER THE FIRST PROGRAM YEAR?

The study found no statistically significant or substantively important difference between schools in their first year implementing quarterly benchmark exams in middle-school mathematics and those not employing the practice. Why? The finding might be because of limitations in the data rather than the ineffectiveness of benchmark assessments.

First, data are lacking on what benchmark assessment practices comparison schools may be using, because the study examined the impact of a particular structured benchmarking program. More than 70 percent of districts are doing some type of formative assessment (Olson, 2005), so it is possible that at least some of the comparison schools implemented their own version of benchmarking. Given the prevalence of formative assessments under the No Child Left Behind Act, it is highly unlikely that a project with strictly controlled conditions could be implemented (that is, with schools using no formative assessment at all as the comparison group).

Second, the study was underpowered. That means that a small but important treatment effect for benchmarking could have gone undetected because there were only 22 program schools and 44 comparison schools.[5] Unfortunately, the sample size for program schools could not be increased because only 25 schools in the eight districts initially received the state grants (three schools were later dropped). Increasing the comparison school sample alone (from 44 to 66, for example) would have brought little additional power.

Third, with only one year of post-implementation data, it may be too early to observe any impact from intervention in the program schools.

HOW TO BETTER UNDERSTAND THE EFFECTS OF BENCHMARK ASSESSMENTS

Although the study did not find any immediate difference between schools employing benchmark assessments and those not doing so, the report provides initial empirical data to inform state and local education agencies.

To understand the longer-term effects of benchmark assessments, it would be useful to continue to track achievement data in the program and comparison schools to reassess the initial findings beyond a single post-intervention year and to provide additional data to local and state decisionmakers about the impact of this benchmark assessment practice.

To understand the longer-term effects of benchmark assessments, it would be useful to continue to track achievement data in the program and comparison schools to reassess the initial findings beyond a single post-intervention year

Using student-level data rather than school-level data might also help researchers examine the impact of benchmark assessment on important No Child Left Behind subgroups (such as minority students or students with disabilities). By comparing school mean scores, as in this study, some nontrivial effects for subgroups may be masked. At the onset of the study, only school-level data were available to researchers, but since then working relationships have been arranged with state education agencies for specific regional educational laboratory projects to use student-level data.

Another useful follow-up would be disaggregating the school achievement data by mathematics content strand to see if there are any effects on particular standards. As the quarterly assessments are broken out by mathematics content strand, doing so would connect logically with the benchmark assessment strategy. Such an approach could determine whether the intervention has affected particular subscales of mathematics in the Massachusetts Curriculum Frameworks. This more refined outcome data may be more sensitive to the intervention and might also provide information to the Massachusetts Department of Education about which content strands schools focused on in their benchmark assessments.

Higher mathematics scores will come not because benchmarks exist but because of how the benchmark assessment data are used by a school's teachers and leaders

Conversations with education decisionmakers support what seems to be common sense. Higher mathematics scores will come not because benchmarks exist but because of how the benchmark assessment data are used by a school's teachers and leaders. This kind of follow-up research is imperative to better understand the impact of benchmark assessment.

But the data sources to identify successful implementation in a fast-response project can be elusive. A possible solution is to examine initial district progress reports to the Massachusetts grant program. These data may provide insight into school buy-in to the initiative, quality of leadership, challenges to implementation, particular standards that participating districts focus on, and how schools use the benchmark assessment data. Researchers may ask whether and how the teachers and administrators used the benchmark data in instruction and whether and how intervention strategies were implemented for students not performing well on the benchmark exams. Based on the availability and quality of the data, the methodology for determining the impact of the intervention could be further refined.

APPENDIX A METHODOLOGY

This appendix includes definitions of benchmark assessments and of the Massachusetts pilot program, an overview of the construction of the study database, the methodology for creating comparison groups, and a description of the data analysis strategy. Because implementation of benchmark testing was at the school level, the unit of analysis was the school. Choosing that unit also boosted the statistical power of the study because there were 25 program schools in the original design rather than eight program districts.[6]

A quasi-experimental design, with program and matched comparison schools, was used to examine whether schools using quarterly benchmark exams in middle-school mathematics under the Massachusetts pilot program showed greater gains in student achievement after a year than schools not in the program. The comparisons were between program schools and comparison schools on post-intervention changes in mathematics performance. All the analyses were based on (mostly) publicly available,[7] school-level achievement and demographic data maintained by the Massachusetts Department of Education. The primary outcome measure was eighth-grade mathematics achievement, as assessed by the Massachusetts Comprehensive Assessment System (MCAS).

Defining the program

The study defined benchmark assessments as assessments that align with the Massachusetts Curriculum Frameworks Standards, are administered quarterly at the school level, and yield student-level data—quickly available to school-level educators and administrators—connected to individual standards and content strands.

The study examined a Massachusetts Department of Education program targeting middle schools. Because what constitutes a middle school differs from town to town, the study defined middle schools as those that include seventh and eighth grades. Other configurations (say, grades K–8, 5–9, 6–8, 6–9, 7–8, 7–9, or 7–12) were acceptable, provided that seventh and eighth grades were included.[8]

For its benchmark assessment initiative, the Massachusetts Department of Education selected high-poverty middle schools under pressure to significantly improve their students' mathematics achievement. To select schools, the Massachusetts Department of Education issued a request for proposals. The department prioritized funding for districts (or consortia of districts) with four or more schools in need of improvement, corrective action, or restructuring under the current adequate yearly progress status model. The "four or more schools" criterion was sometimes relaxed during selection. Applications were given priority based on the state's No Child Left Behind performance rating system:

- Category 1 schools were rated "critically low" in mathematics.
- Category 2 schools were rated "very low" in mathematics and did not meet improvement expectations for students in the aggregate.
- Category 3 schools were rated "very low" in mathematics and did not meet improvement expectations for student subgroups.
- Category 4 schools were rated "very low" in mathematics and did meet improvement expectations for all students.
- Category 5 schools were rated "low" in mathematics and did not meet improvement expectations for students in the aggregate.

The Massachusetts Department of Education selected 25 schools representing eight districts to participate in the pilot initiative. The selection of program schools targeted high-poverty schools having the most difficulty in meeting goals for student mathematics performance, introducing a selection bias into the project. Unless important variables were controlled for by design and analysis (for example, poverty and pretest or baseline mathematics scores), any results would be confounded by pre-existing differences between schools. In the study, balance was achieved between the program and comparison schools on poverty, pretest mathematics scores, and other school-level social and demographic variables. But because the study was based on a quasi-experimental design (without random assignment to conditions), it could not assess whether the participant and comparison groups were balanced on unobserved factors.

Constructing the study database

A master database was developed in SPSS to house all the necessary data. Data were collected from student- or school-level achievement and demographic data maintained by the Massachusetts Department of Education.[9] The outcome variable was scaled eighth-grade MCAS mathematics scores for 2001–06.

The MCAS was implemented in response to the Massachusetts Education Reform Act of 1993 and fulfills the requirements of the federal No Child Left Behind Act of 2001, which requires annual assessments in reading and mathematics for students in grades 3–8 and in high school. The MCAS tests all public school students in Massachusetts, including students with disabilities and those with limited English proficiency. The MCAS is administered annually and measures student performance on the learning strands in the Massachusetts Curriculum Frameworks (see appendix E). In mathematics these strands include number sense and operations; patterns, relations, and algebra; geometry; measurement; and data analysis, statistics, and probability.

According to the Massachusetts Department of Education (2007), the purpose of the MCAS is to help educators, students, and parents to:

- Follow student progress.
- Identify strengths, weaknesses, and gaps in curriculum and instruction.
- Fine-tune curriculum alignment with statewide standards.
- Gather diagnostic information that can be used to improve student performance.
- Identify students who may need additional support services or remediation.

The MCAS mathematics assessment contains multiple choice, short-answer, and open response questions. Results are reported for individual students and districts by four performance levels: advanced, proficient needs improvement, and warning. Each category corresponds to a scaled score range (see appendix E, table E1). Although the scaled score was the primary outcome variable of interest, the corresponding raw score was also collected to determine if scaled scores might have masked program effects.

The MCAS mathematics portion, comprising two 60-minute sections, is administered in May in grades 3–8. Students completing eighth grade take the MCAS in the spring of the eighth-grade year. Preliminary results from the spring administration become available to districts the next August. Eighth graders who enter in the 2007/08 school year, for example, take MCAS mathematics in May 2008, and their preliminary results become available in August 2008.

Other variables gathered for the study included the school name, grade structure, and enrollment, the race and ethnicity of students, and the proportion of limited English proficiency and low-income students. Demographic data were transformed from total numbers to the percentage of students in a category enrolled at the school (for example, those defined as low income). Supplementary geographic location data were added from the National Center for Educational Statistics, Common Core of Data to identify school location (urban, rural, and so on). A variable was also created to designate each school as a program or comparison school based on the results of the matching procedure. See appendix B for the specific steps in constructing the database.

Creating a comparison group

Only a well implemented randomization procedure controls for both known and unknown factors that could influence or bias findings. But because the grants to implement benchmark assessments were already distributed and the program was already assigned to schools, random assignment to conditions was not possible. So, it was necessary to use other procedures to create a counterfactual—a set of schools that did not receive the program.

The study used covariate matching to create a set of comparison schools (appendix C details the matching procedure). Using covariates in the matching process is a way to control for the influence of specific factors on the results. In other words, the comparison schools would be as similar as possible to the program schools (in the aggregate) on these factors, meaning that any findings, whether positive or negative, would be unlikely to have been influenced by these factors. The variables used in the matching procedure included a composite index of school-level social and demographic variables: enrollment, percentage of students classified as low income, percentage of students classified as English language learners, and percentage of students categorized in different ethnic groups. Also included in the matching procedure were each school's eighth-grade baseline (or pretest) mathematics score (based on an average of its 2004/05 eighth-grade mathematics scores) and the type of geographic location the school served (classified according to the National Center for Education Statistics' Common Core of Data survey).

Prior research guided the selection of the variables used as covariates in the matching. Bloom (2003) suggests that pretest scores are perhaps the most important variable to use in a matching procedure. Pretest–post-test correlations on tests like the MCAS can be very high, and it is important that the comparison group and program group are as similar as possible on pretest scores. By taking into account the 2004/05 average eighth-grade mathematics scores (also known as the Composite Performance Index), the report tried to ensure that the comparison schools are comparable on baseline mathematics scores.

There is substantial research that identifies large gaps in academic achievement for racial minorities (Jencks & Phillips, 1998), low-income students (Hannaway, 2005), and English language learners (Abedi & Gandara, 2006). Unless these influences were controlled for, any observed differences might have been due to the program or comparison schools having a higher share of students in these categories rather than to benchmarking. Although the research on the relationship between school size and academic achievement is somewhat conflicting (Cotton, 1996), the variability in school size led the report to introduce into the matching procedure the total enrollment in the middle school.

The eligibility pool for the comparison matches included the 389 Massachusetts middle schools that did not receive the Comprehensive School Reform grants. Statistical procedures were used to identify the two best matches for each program school from the eligibility pool. The covariate matching resulted in a final sample of 22 program schools and 44 comparison schools that were nearly identical on pretest academic scores.[10] In addition, the project design achieved balance on nearly all school-level social and demographic characteristics, except that there were larger shares of African American and Pacific Islander students in program schools. These differences were controlled for statistically in the outcome analysis, with no change in the results (see appendix D).

Analyzing the data

After matching, descriptive statistics were used to examine the mean scores for all five pre-implementation years and one post-implementation year for the program and comparison schools. A comparative interrupted time series analysis was also used to more rigorously assess whether there was a statistically significant difference between the program and comparison schools in changes in mathematics performance (see Bloom, 2003; Cook & Campbell, 1979). The interrupted time series design was meant to determine whether there was any change in the trend because of the "interruption" (program implementation).

The method for short interrupted time series in Bloom (2003) was the analysis strategy. Bloom argues that the approach can "measure the impact of a reform as the subsequent deviation from the past pattern of student performance for a specific grade" (p. 5). The method establishes the trend in student performance over time and analyzes the post-intervention data to determine whether there was a departure from that trend. This is a tricky business, and trend departures can often be statistically significant. It is important to rule out other alternative explanations for any departure from the trend, such as change in principals, other school reform efforts, and so on. Although Bloom outlines the method for use in evaluating effects on a set of program schools alone, having a well matched group of comparison schools strengthens causal inferences.

To project post-implementation mathematics achievement for each school, both linear baseline trend models and baseline mean models (see Bloom, 2003) were estimated using scaled and raw test score data collected over five years before the intervention. Estimates of implementation effects then come from differences-in-differences in observed and predicted post-implementation test scores between program and comparison schools.

APPENDIX B
CONSTRUCTION OF THE STUDY DATABASE

The following outlines the specific steps taken to construct the study database:

1. Identify all the middle schools in Massachusetts.
2. Identify the 25 program schools using benchmark assessments in mathematics.
3. Collect the following variables from the Massachusetts Department of Education web site on each of the schools—to proceed to the covariate matching exercise that will identify two matched comparison schools for each program school:
    a. School name.
        i. Source: http://profiles.doe.mass.edu/enrollmentbygrade.aspx.
    b. CSR implementation.
    c. School locale (urban, rural, and so on).
        i. Source: http://nces.ed.gov/ccd/districtsearch/.
    d. Does the school have a 6th grade?
        i. Source: http://nces.ed.gov/ccd/districtsearch/.
    e. Does the school have an eighth grade?
        i. Source: http://nces.ed.gov/ccd/districtsearch/.
    f. Total enrollment.
        i. Source: http://profiles.doe.mass.edu/enrollmentbygrade.aspx.
    g. Race/ethnicity of student population.
        i. Source: http://profiles.doe.mass.edu/enrollmentbyracegender.aspx?mode=school&orderBy=&year=2006.
    h. Limited English proficiency.
        i. Number of students.
            1. Source: http://profiles.doe.mass.edu/selectedpopulations.aspx?mode=school&orderBy=&year=2006.
        ii. Percentage of limited English proficiency students
            1. Number of limited English proficiency students / total enrollment
    i. Low income
        i. Number of low-income students.
            1. Source: http://profiles.doe.mass.edu/selectedpopulations.aspx?mode=school&orderBy=&year=2006.
        ii. Percentage of low-income students.
            1. Number of low-income students /total enrollment.
    j. Mathematics baseline proficiency index.
        i. Source: http://www.doe.mass.edu/sda/ayp/cycleII.

Seven program schools had missing data for the mathematics baseline proficiency index, which was serving as the measure for academic performance for each school in the matching equation. Therefore, it was substituted with the 2005 mathematics Composite Proficiency Index (CPI) score to get an accurate academic measure for each school. The 2005 mathematics CPI score was taken from "1999–2006 AYP History Data for Schools," which can be found at http://www.doe.mass.edu/sda/ayp/cycleIV/.

4. Charter and alternative schools were deleted from the master file because they would not have been eligible for the initial program and because their populations differ significantly in many cases from those of regular schools.

5. After the covariate matching was performed on this database, a new variable, STUDY GROUP, was created to determine if a school is defined as a program school, a comparison school, or an "other" school.

6. The following additional variables on achievement scores were collected for the schools that were either program or comparison schools (English scaled and raw scores were also collected and added to the database):

a. 2001 mathematics achievement mean scaled score and mean raw score.

b. 2002 mathematics achievement mean scaled score and mean raw score.

c. 2003 mathematics achievement mean scaled score and mean raw score.

d. 2004 mathematics achievement mean scaled score and mean raw score.

e. 2005 mathematics achievement mean scaled score and mean raw score.

f. 2006 mathematics achievement mean scaled score and mean raw score.

APPENDIX C
IDENTIFICATION OF COMPARISON SCHOOLS

Random assignment to study the impact of benchmark assessment was not possible because selected districts and schools had already been awarded the Comprehensive School Reform grants. When members of the group have already been assigned, the research team must design procedures for developing a satisfactory comparison group. Fortunately, researchers have been developing such methods for matching or equating individuals, groups, schools, or other units for comparison in an evaluation study for many years. And as one might imagine, there are many such statistical approaches to creating comparison groups. All such approaches, however, have one limitation that a well implemented randomized study does not: the failure to control for the influence of "unobserved" (unmeasured or unknown) variables.

Of the many statistical equating techniques, one of the more reliable and frequently used is covariate matching. How does "covariate matching" work?

> Let's say we have unit "1" already assigned to the program and wish to match another unit from a pool of observations to "1" to begin creating a comparison group that did not receive the program. When using covariate matching, the research team would first identify the known and measured factors that would be influential on the outcome (in this instance, academic performance) regardless of the program. Influential in this case means that less or more of that characteristic has been found—independent of the program—to influence scores on the outcome measure. These are known as *covariates.* To reduce or remove their influence on the outcome measure, one would select the unit that is closest to "1" on those covariate scores or characteristics. Let's say that "X" is the closest in the eligibility pool to "1." By matching "X" to "1," there would be two units that are very similar on the covariates. If this is done for each program unit, theoretically the influence of the covariates will be removed (or considerably reduced), the differences between the groups on important known factors will be ameliorated, and a potential explanation for observed results besides program effectiveness or ineffectiveness (that the groups were different before the program on important covariates) will be seriously countered (Rubin, 1980).

For the current project, covariate matching is used to create a set of comparison schools for the study.[11] The original proposal was to match comparison schools using three factors: the Socio-Demographic Composite Index (SCI); the school's adjusted baseline academic performance, holding constant the school's SCI; and the type of geographic location the school sits in (urban, suburban, rural, and so on). Using two comparison schools for each program school was eventually chosen to counter the possibility of idiosyncratic matching.

The first order of business was to create the SCI. The SCI is simply the "predicted" mathematics score (using the school's 2005 average baseline mathematics score), using a multivariate regression analysis, for each school based on a series of social and demographic covariates. In other words, it is a prediction of students' 2005 mathematics score using important covariates such as school enrollment, percentage of low-income students, percentage of English language learners, and percentage of minority/ethnic groups.[12] In short, multivariate regression was used to predict what the average mathematics score for the school is, given knowledge about the school's characteristics (how many kids are enrolled, the percentage of low-income students, the percentage of English language learners, and the percentage of minority students).[13]

One of the advantages in using multivariate regression is that the factors comprising the SCI are weighted proportionately to how much of the

TABLE C1

Percentage of low-income students and Socio-Demographic Composite Index for five selected schools

School	Percentage of low-income students	Socio-Demographic Composite Index
School 1	0	86
School 2	10	77
School 3	25	68
School 4	75	54
School 5	95	47

Source: Authors' analysis based on data described in text.

TABLE C2

2005 Composite Performance Index mathematics score, Socio-Demographic Composite Index, and "adjusted" academic score for five selected schools

School	2005 Composite Performance Index mathematics	Socio-Demographic Composite Index	"Adjusted" score
School 1	90	86	4
School 2	80	77	3
School 3	59	68	–10
School 4	51	54	–3
School 5	50	47	2

Source: Authors' analysis based on data described in text.

2005 mathematics score they predict. For example, if poverty is a more substantial factor in school success than enrollment, the regression will give more weight to the percentage of low-income students that a school has than the school's total enrollment. This is exactly what happened with the results. Table C1 provides the SCI for five middle schools with varying percentages of low-income students. As the percentage of low-income students increases, SCI decreases.

The other covariate used in the matching procedure is the school's adjusted academic score. This is simply the actual score minus the predicted score. In other words, if the SCI is subtracted from the 2005 CPI mathematics score, the result is the adjusted value that was used as the other major covariate in the matching procedure. Table C2 shows this relationship (numbers do not add up perfectly because of rounding).

The multivariate regression was conducted in both Stata and SPSS and, as might be expected, there was perfect agreement on the results.

What next? Finding similar schools using Mahalanobis Distance measures

Variables like SCI and the adjusted mathematics score form a "multidimensional space" in which each school can now be plotted.[14] The middle of this multidimensional space is known as a "centroid."[15] The Mahalanobis Distance can be computed as the distance of a case or observation (such as a school) from the centroid in this multidimensional space. Schools with similar Mahalanobis Distance measures are considered to be "close" in this multidimensional space, and therefore more similar. One way to think about how Mahalanobis Distance measures are computed is shown in the figure at http://www.jennessent.com/images/graph_illustration_small_4.gift.

Using a specialized software program (Stata), Mahalanobis Distance measures were computed for each of the 410 eligible middle schools in the state.[16] Table C3 provides the SCI, the adjusted mathematics achievement score, and the Mahalanobis Distance measure for five schools. If school 1 was a program school and the remaining 4 schools formed the pool for comparison schools, the best match according to the analysis would be school 2, because the Mahalanobis Distance measure for school 2 is more similar to school 1 than any of the other potential schools.

Note that the study plan required that two schools had to be matched to each program school. The 23 schools remaining in the program group required 46 comparison schools. To complete the matching process, a list was printed of the 23 program schools with the Mahalanobis Distance measure and its population geographic area code (that is, the

TABLE C3
Socio-Demographic Composite Index, "adjusted" academic score, and Mahalanobis Distance score for five selected schools

School	Socio-Demographic Composite Index	"Adjusted" score	Mahalanobis Distance score
School 1	86	4	49.18
School 2	77	3	38.54
School 3	68	–10	32.30
School 4	54	–3	19.15
School 5	47	2	14.87

Source: Authors' analysis based on data described in text.

type of geographic location the school serves, such as small urban or suburban), as schools were classified according to the National Center for Education Statistics Common Core of Data. A similar list was printed of the 387 remaining middle schools that comprised the potential schools for comparison matching. The matching was conducted by simply selecting the two schools with the closest Mahalanobis Distance measures and with the exact or very similar population geographic area code. This is also known as "nearest neighbor matching." The initial matching resulted in 46 comparison schools matched to 23 program schools. The findings are presented in table C4.

Because there are a variety of matching methods, and some variations on using Mahalanobis Distance measures for matching, replication of the findings was initiated.[17]

David Kantor developed a different procedure for using Mahalanobis Distances to form a comparison group in Stata called Mahapick.[18] Rather than compute Mahalanobis Distance measures from the centroid of a multidimensional space, as in the earlier procedure, Mahapick creates a measure based on the distance from a treated observation to every other observation. It then chooses the best matches for that treated observation and makes a record of these matches. It then drops the measure and goes on to repeat the process on the next treated observation.

Because Mahapick uses a different method and produces a different Mahalanobis Distance score, the goal was not to confirm whether the scores were identical. The goal was to see if a similar set of schools was constructed using a different matching method.[19]

One problem with using Mahapick is that the computation does not produce exclusive matches. The procedure selected 12 duplicate matches. Nine schools, however, were the results of exact matches in both procedures. Nearly half of the schools were selected by both (22 of 46), and this number might have been higher had Mahapick selected exclusive matches, that is, if it had not matched one comparison school to more than one program school.

Both methods produced a large number of mid-sized urban schools with Mahalanobis Distance scores that clustered very closely together. This is not surprising, as the Mahalanobis Distance measure scores (using the initial Stata procedure) for the program schools clustered between 12.25 and 31.56. This meant that the comparison group schools would likely be drawn from schools in the eligible pool whose distance measure scores also fell into this range. Of the 387 schools in the eligibility pool, 166 had Mahalanobis Distance measure scores of 12.25–31.56 (43 percent).

Because the two procedures did not produce the exact 46 comparison schools, a combination of the results from the initial Stata procedure and Mahapick was used to select the next iteration. Putting the nine exact matches produced by both procedures aside, 37 comparison schools were left to identify. The selected matches for each program school provided by the initial Stata procedure and Mahapick were examined. Once two comparison schools for a program school were selected, they were removed from consideration. In those cases in which there were more than two schools identified by the two matching procedures, the one with the higher adjusted 2005 mathematics score was selected. This decision is a conservative one and initially presents a bias favoring the comparison group. The results of using both procedures to perform the matching are provided in table C6.

Following this matching procedure, the pretest achievement data files from 2001–05 were added. During this process, it became known that 1 of the original 25 schools, because it is a reconfigured school, did not have pretest data for eighth graders. This school was dropped from the program sample (along with the corresponding two comparison schools). In addition, another school from the original matching set was dropped because it too did not have eighth-grade pretest data. It was replaced with another comparable school.

How well did the matching procedure work?

To test the effectiveness of the matching procedure, the 22 program schools and 44 comparison schools were compared across the variables in the school-level dataset. Table C4 presents the results from that comparison. In summary, the equating or matching process resulted in certain variables favoring the comparison schools (for example, higher baseline mathematics and English/language arts scores). Some of this might be due to the matching procedure as comparison schools with a higher 2005 adjusted mathematics score were selected when there was more than one possible school to pick from for a match. Two of the variables were statistically significant (2005 baseline mathematics scores and percentage of African American students).

Both of these differences were troubling, especially given that they were included in the covariate matching process. To further investigate whether there were systemic differences between the program and comparison schools on the pretest

TABLE C4

Comparison of means and medians of initial program and comparison schools

	Program schools (N=22)		Comparison schools (N=44)	
Characteristic	Mean	Median	Mean	Median
2005 Mathematics Composite Performance Index	53.10*	53.80*	60.56*	59.85*
2005 English language arts Composite Performance Index	69.51	69.50	74.17	76.90
Enrollment	620.23	635.00	562.95	578.50
Low-income students (percent)	61.32	64.60	54.71	51.70
English language learners (percent)	16.74	13.90	11.23	7.80
African American (percent)	7.45*	6.10*	14.99*	10.05*
Hispanic (percent)	29.81	23.00	26.40	14.10
Asian (percent)	9.60	4.00	6.60	5.20
White (percent)	51.80	55.20	49.80	47.60
Native American (percent)	0.26	0.20	0.41	0.25
Hawaiian/Pacific Islander (percent)	0.11	0.00	0.08	0.00
Multi-race non-Hispanic (percent)	1.00	0.70	1.73	1.30
Race-ethnicity composite (percent)	48.20	44.90	50.20	52.40
Highly qualified teachers (percent)	90.10	91.90	89.40	94.70
School location	Number	Percent	Number	Percent
Mid-size city	17	77.30	34	77.30
Urban fringe of large city	3	13.60	6	13.60
Urban fringe of mid-size city	2	9.10	4	9.10

* Statistically significant at the 0.05 level using a t-test (two-tailed).

Note: Race-ethnicity composite is the sum African American, Hispanic, Asian, Native American, Hawaiian, Pacific Islander, and multirace non-Hispanic.

Source: Authors' analysis based on data described in text.

achievement years, analyses of scaled eighth-grade mathematics scores for each year of pretest data available were examined. The differences between the two groups for each year of pretest data were statistically significant and favored the comparison schools. Table C5 presents the results of the t-tests. Rerunning the t-tests using raw scores did not change the results.

Resolving matching problems

Revisiting the "conservative tie-breaker." Note that when there were multiple schools eligible for the matching, the school that had higher achievement scores (using the 2005 CPI baseline measure) was selected. This might have explained the lack of equivalence on the pretest baseline mathematics scores, with the procedure inflating these pretest scores. Because achievement scores are highly correlated and the 2005 CPI actually represents an average of the 2004 and 2005 MCAS mathematics scores, it was reasonable to assume that this conservative decision was responsible for the lack of equivalence across all years.

Surprisingly, however, when the data were closely examined, the equivalence problem did not exist for only schools with the "conservative tie-breaker" decision. The higher pretest scores for the comparison schools were consistent across most of the program schools, including those where such choices were not made. This led to further investigations.

Revisiting Mahapick. Mahapick does not permit exclusive matches, so running the procedure results in the same schools getting selected as comparisons for more than one program school. This happened with eleven schools. One attempt

TABLE C5

T-test for differences in pretest mathematics scores between initial program and comparison schools, 2001–05

T-statistic	t-statistic	Difference	Significance (two-tailed)	Mean difference	Standard error difference	95 percent confidence interval of the difference	
Mathematics 2001 grade 8 scaled score							
Equal variances assumed	–2.161	50.000	0.036	–3.72213	1.72237	–7.18162	–0.26265
Equal variances not assumed	–2.480	44.996	0.017	–3.72213	1.50087	–6.74506	–0.69921
Mathematics 2002 grade 8 scaled score							
Equal variances assumed	–2.431	51.000	0.019	–4.33164	1.78158	–7.90832	–0.75496
Equal variances not assumed	–2.918	48.594	0.005	–4.33164	1.48465	–7.31579	–1.34749
Mathematics 2003 grade 8 scaled score							
Equal variances assumed	–2.043	55.000	0.046	–3.11489	1.52486	–6.17077	–0.05900
Equal variances not assumed	–2.362	47.665	0.022	–3.11489	1.31898	–5.76736	–0.46241
Mathematics 2004 grade 8 scaled score							
Equal variances assumed	–2.070	59.000	0.043	–3.04558	1.47100	–5.98904	–0.10212
Equal variances not assumed	–2.369	48.800	0.022	–3.04558	1.28561	–5.62938	–0.46179
Mathematics 2005 grade 8 scaled score							
Equal variances assumed	–2.342	64.000	0.022	–3.24972	1.38767	–6.02191	–0.47753
Equal variances not assumed	–2.723	60.799	0.008	–3.24972	1.19360	–5.63664	–0.86281
Mathematics 2006 grade 8 scaled score							
Equal variances assumed	–1.945	64.000	0.056	–2.91159	1.49668	–5.90155	0.07837
Equal variances not assumed	–2.103	51.800	0.040	–2.91159	1.38451	–5.69006	–0.13312

Source: Authors' analysis based on data described in text.

to remedy this problem: Mahapick creates the best match for each program observation (in this case, the program school), beginning with the very first case. By removing the two comparison schools after each match is made, the problem of non-exclusive selections is eliminated.

Mahapick was therefore used in this way, running 22 separate analyses, or one separate analysis for each program school. As the two comparison matches were made, they were eliminated from the next run, and so on, until the Mahapick procedure selected the 44 unique comparison schools that were needed. Although the results of this procedure produced a set of schools that were closer on the pretest achievement measures (differences were not significant), the measures were still higher for comparison schools than program schools during 2001 and 2002 and close to significant (table C6).

Adding the 2005 CPI mathematics baseline score as an additional "sort" variable. Finally, it was determined that the best way to create equivalence on the pretest achievement measures was to redo the sort and match again. This time, instead of sorting solely on the Mahalanobis Distance measure score and geographic location, the 2005 CPI baseline mathematics score was included. Printouts of both program and potentially eligible comparison schools sorted on these three variables were prepared. The priority was to ensure that schools were as close as possible on the 2005 CPI baseline mathematics score and distance measure within each geographic location category. The use of the 2005 CPI baseline mathematics score together with the distance measure resulted in a new set of 44 comparison schools. Note that one new comparison school did not have any pretest achievement data and was replaced with a similar school. Although there was considerable overlap

TABLE C6
T-test for differences in pretest scaled mathematics scores, 2001–05 (Mahapick sample)

T-statistic	t-statistic	Difference	Significance (two-tailed)	Mean difference	Standard error difference	95 percent confidence interval of the difference	
Mathematics 2001 grade 8 scaled score							
Equal variances assumed	–1.428	46.000	0.160	–2.55277	1.78826	–6.15235	1.04681
Equal variances not assumed	–1.616	44.688	0.113	–2.55277	1.57942	–5.73450	0.62895
Mathematics 2002 grade 8 scaled score							
Equal variances assumed	–1.514	47.000	0.137	–2.41042	1.59251	–5.61413	0.79329
Equal variances not assumed	–1.704	44.217	0.095	–2.41042	1.41459	–5.26095	0.44011
Mathematics 2003 grade 8 scaled score							
Equal variances assumed	–0.803	50.000	0.426	–1.15321	1.43681	–4.03912	1.73270
Equal variances not assumed	–0.884	44.829	0.382	–1.15321	1.30506	–3.78200	1.47559
Mathematics 2004 grade 8 scaled score							
Equal variances assumed	–0.167	58.000	0.868	–0.23275	1.39500	–3.02515	2.55966
Equal variances not assumed	–0.186	46.255	0.853	–.23275	1.25220	–2.75292	2.28743
Mathematics 2005 grade 8 scaled score							
Equal variances assumed	–0.303	63.000	0.763	–0.34596	1.14138	–2.62682	1.93491
Equal variances not assumed	–0.326	51.785	0.745	–0.34596	1.05996	–2.47313	1.78122
Mathematics 2006 grade 8 scaled score							
Equal variances assumed	–0.261	64.000	0.795	–0.36091	1.38324	–3.12426	2.40244
Equal variances not assumed	–0.272	47.280	0.786	–0.36091	1.32468	–3.02541	2.30359

Source: Authors' analysis based on data described in text.

with the earlier listings of schools, the t-tests of equivalence showed a near perfect balance on the pretest achievement measures (table C7).

The one variable that remained statistically significant was the difference on the percentage of African American students enrolled in the school. One possible reason for this imbalance is that the state grants were provided to a rather constricted range of middle schools in Massachusetts. These were mostly small city urban schools with diverse populations, and the average 2005 CPI mathematics score was clustered to the lower or middle part of the distribution. With all these factors in play, there was a limited pool of comparison schools for achieving perfect balance on all pre-existing variables. Taylor (1983) notes that matching on some variables may result in mismatching on others.

The final set of matches represents the most rigorous set of comparison schools that could have been selected given the limited eligibility pool. Although the imbalance on the percentage of African American students enrolled at the schools remains troubling,[20] the variable ("AFAM") was introduced as a covariate in the final time series analysis. It makes no difference in the results (see appendix D).

TABLE C7
Comparison of means and medians of final program and comparison schools

	Program schools (N=22)		Comparison schools (N=44)	
Characteristic	Mean	Median	Mean	Median
2005 Mathematics Composite Performance Index	53.10	53.80	52.82	54.25
2005 English language arts Composite Performance Index	77.15	77.05	73.86	74.05
Enrollment	620.23	635.00	547.73	577.50
Low-income students (percent)	61.32	64.60	62.92	65.50
English language learners (percent)	16.74	13.90	11.02	9.90
African American (percent)	7.45*	6.10*	15.36*	11.30*
Hispanic (percent)	29.81	23.00	33.93	26.65
Asian (percent)	9.58	4.00	5.24	2.85
White (percent)	51.77	55.15	43.48	37.95
Native American (percent)	0.26	0.20	0.36	0.25
Hawaiian/Pacific Islander (percent)	0.11*	0.00*	0.02*	0.00*
Multi-race non-Hispanic (percent)	1.00	0.70	1.60	1.25
Race-ethnicity composite (percent)	48.22	44.90	56.52	62.05
Highly qualified teachers (percent)	90.08	91.90	86.42	88.55
School location	Number	Percent	Number	Percent
Mid-size city	17	77.3	34	77.3
Urban fringe of large city	3	13.6	6	13.6
Urban fringe of mid-size city	2	9.1	4	9.1

* Statistically significant at the 0.05 level using a t-test (two-tailed).

Note: Race-ethnicity composite is the sum African American, Hispanic, Asian, Native American, Hawaiian, Pacific Islander, and multirace non-Hispanic.

Source: Authors' analysis based on data described in text.

APPENDIX D
INTERRUPTED TIME SERIES ANALYSIS

Conventional interrupted time series analysis generally requires multiple data points before and after an intervention (or "interruption") and the use of administrative or other data that is regularly and uniformly collected over time. It is common, for example, in interrupted time series analyses of the effects of laws or policies on crime for researchers to use monthly or weekly crime data to create more pretest and post-test points. All things being equal, the more data points in a time series, the more stable the analysis.

In education some commonly used achievement outcomes (such as standardized mathematics test scores) are usually administered once per year. Thus, the multiple pretests and post-tests that are common to conventional time series analyses may not be available when evaluating the impact of school innovations on student achievement. In this report, only five years of annual mathematics test score data and one post-test administration were available. Clearly, with only six data points, it would not be possible to conduct conventional time series analysis.

Bloom's (2003) method for "short interrupted time series," outlined in an evaluation of Accelerated Schools by MDRC, served as the analysis strategy. In short, Bloom (2003) argues that his approach can ". . . measure the impact of a reform as the subsequent deviation from the past pattern of student performance for a specific grade" (p.5). Bloom's method establishes the trend in student performance over time and then analyzes the post-program data to determine if there is a departure from that trend. As noted in the report, this is a tricky business, and trend departures can often be statistically significant. Although Bloom (2003) outlines his approach for use in evaluating the impact on a set of program schools alone, he recognizes the importance of having a well matched comparison group of schools to strengthen causal inferences.

Note, however, that Bloom's (2003) paper describes an evaluation that had five full years of student-level test score data and five full years of post-test student-level data. In this report, available at this time are only school-level means for one post-intervention year. Bloom (2003) may argue that this is not a fair test, as one year does not allow the school reform to be implemented to its fullest strength. Nonetheless, this should be viewed as a valuable foundational effort in the Regional Educational Laboratory's research on the impact of benchmark assessment.

Reconstructing the database

The first order of business was to convert the database from one in which each row represented all the data for each school (66 rows of data) to one in which each row represented a different year of either pretest or post-test information. For example, the 44 comparison group schools represented 230 unique rows of data; the 22 program schools represented 115 distinct rows of pretest or post-test information. The database, after reconstruction, consisted of 345 total rows of data (rather than 66). Variables were also renamed and reordered in the database, to ease analysis.[21]

A series of models analogous to Bloom's (2003) recommendations was then run to determine, using more advanced statistical analysis, whether there was any observed program impact on eighth-grade mathematics outcomes. Bloom's paper provides three potential time series models to take into account when constructing statistical analyses, and each has different implications for how the analysis is done. Bloom argues that the type of statistical model must take into account the type of trend line for the data. The three models include the linear trend model (in which the outcome variable increases incrementally over time), the baseline mean model (in which the outcome variable appears to be a flat line over time with no discernible increase or decrease), and the nonlinear baseline trend model (in which the outcome scores may be moving in a curvilinear or other pattern). The outcome data from the pretests clearly showed that the most applicable model was likely the baseline mean model, but given that there was a slight increase over time (from 2001 to 2006), the linear trend model could not be ruled out. So the analyses were run using both models.

For each of the two models, analyses were run to determine if there was an effect in 2006 for program and comparison schools separately—a difference-in-difference effect (or effect between program and comparison schools)—and then covariates were introduced to determine if any of the estimates changed for time or for program impact when variables such as percentage of African American students enrolled at the schools were introduced.[22] Variables used in the analysis are described in table D1.

First, using a "baseline mean model" as described by Bloom (2003), the researchers investigated if there was a perceptible immediate change from 2001–05 to 2006. This was done for comparison schools and program schools separately. When looking at program schools alone in table D2, there appears to be a significant increase in 2006. This increase ("Y2006") represents a 1.86 test point improvement over what would have been expected in the absence of the program.

It would have been possible to conclude that benchmark assessment had a statistically significant and positive impact on the implementation year mathematics outcomes from the results in table D2. But table D3 highlights the importance of the comparison group. The results for 44 comparison schools also show a significant increase in 2006. The increase is modest (1.48 test points) but also statistically significant. Thus, both program and comparison schools experienced significant and positive upward trends in 2006 that departed from past performance.

The difference-in-difference test, which is the most critical because it provides a direct comparison between the program and comparison schools, shows no significant difference, as highlighted in table D4. There is a significant increase in 2006, as expected,

TABLE D1
Variables used in the analysis

Variable	Description
Afam	Percent of students enrolled in the school who are African American.
Asian	Percent of students enrolled in the school who are Asian.
Hisp	Percent of students enrolled in the school who are Hispanic.
Hqtper	Percent of highly qualified teachers at the school.
Itreat_1	Effect from being in the program.
Intercept	Mean scores.
IY2006_1	Score increase in 2006, combining treatment and comparison schools.
Iy20xtre~1	Interaction term between the year 2006 and whether a school was in the program.
Lepper	Percent of students in the school classified as "limited English proficiency."
Lipper	Percent of students in the school classified as "low income."
Totenrl	Number of students enrolled at the school.
White	Percent of students enrolled in the school who are White.
Y2006	Score increase in 2006.

TABLE D2
Baseline mean model, program schools only (N=22 schools, 115 observations)

Variable	Coefficient	Standard error	Probability
Intercept	225.11	0.822	0.000
Y2006	1.86	0.556	0.001

Source: Authors' analysis based on data described in text.

TABLE D3
Baseline mean model, comparison schools only (N=44 schools, 230 observations)

Variable	Coefficient	Standard error	Probability
Intercept	224.69	0.781	0.000
Y2006	1.48	0.57	0.009

Source: Authors' analysis based on data described in text.

TABLE D4

Baseline mean model, difference-in-difference estimate (N=66 schools, 345 observations)

Variable	Coefficient	Standard error	Probability
Intercept	224.690	0.722	0.000
IY2006_1	1.480	0.517	0.004
Itreat_1	0.421	1.250	0.340
Iy20xtre_~1	0.379	0.899	0.420

Source: Authors' analysis based on data described in text.

as this analysis combines the program and comparison schools ("IY2006_1"). Whether a school was in the program or comparison group did not appear to have any impact ("Itreat_1"). But the key variable is the interaction between year 2006 and whether a school was in the program group, as represented by "Iy20xtre~1." The program effect is about 0.38 of a mathematics test point, but it is not significant and could have occurred by chance alone. The most accurate interpretation is that both groups are slightly increasing, but the difference between them is negligible. Therefore, any observable increase cannot be attributed to the program.

Although covariates should have been controlled by the matching procedure, analyses in appendix C showed that there were some differences on racial/ethnic variables. These and other covariates were introduced into the difference-in-difference analyses to see if the estimate for program effects would change. As the reader can see from table D5, the introduction of a number of variables into the regression did not change the estimate for program impact.

The same analyses described above were repeated, but the "linear trend model" outlined in Bloom (2003) was now assumed instead of the baseline mean model. Assuming different models simply means that different statistical formulae were used to conduct the analyses. Table D6 presents the data for program schools alone. The table shows that when time is controlled in the analysis, the statistically significant effect for Y2006 (for the program separately) disappears.

TABLE D5

Baseline mean model, difference-in-difference estimate, with covariates (N=66 schools, 345 observations)

Variable	Coefficient	Standard error	Probability
Intercept	242.140	22.290	0.000
IY2006_1	1.540	0.518	0.003
Itreat_1	−0.159	0.948	0.860
Iy20xtre_~1	0.416	0.900	0.640
Afam	−0.067	0.236	0.770
Asian	−0.043	0.248	0.860
Hisp	−0.087	0.228	0.700
White	−0.145	0.230	0.520
Totenrl	0.001	0.001	0.440
Lepper	0.049	0.070	0.480
Liper	−0.236	0.036	0.000
Hqtper	0.076	0.039	0.050

Source: Authors' analysis based on data described in text.

TABLE D6

Linear trend model, program schools only (N=22 schools, 115 observations)

Variable	Coefficient	Standard error	Probability
Intercept	224.370	0.825	0.000
Y2006	0.975	0.741	0.180
Time	0.325	0.174	0.060

Source: Authors' analysis based on data described in text.

TABLE D7

Linear trend model, comparison schools only (N=44 schools, 230 observations)

Variable	Coefficient	Standard error	Probability
Intercept	224.260	0.887	0.000
Y2006	0.981	0.749	0.190
Time	0.187	0.180	0.300

Source: Authors' analysis based on data described in text.

The analysis for comparison schools alone was repeated in table D7. Again, the statistically significant findings, assuming the baseline mean model, drop when assuming the linear trend model.

TABLE D8

Linear trend model, difference-in-difference estimate (N=66 schools, 345 observations)

Variable	Coefficient	Standard error	Probability
Intercept	224.150	0.787	0.000
Time	0.234	0.132	0.070
IY2006_1	0.855	0.626	0.170
Itreat_1	0.432	1.260	0.730
Iy20xtre_~1	0.368	0.894	0.410

Source: Authors' analysis based on data described in text.

Assuming the linear trend model, the difference-in-difference estimates in table D8 nearly replicated the results in the baseline mean model. Again, the program impact is 0.37 of a scaled point on the mathematics test, but this difference is again not significant and could easily have occurred by chance.

In table D9 covariates were again introduced into the difference-in-difference analysis. The results are similar to the baseline mean model except that the "time" variable is also introduced into the analysis.

TABLE D9

Linear trend model, difference-in-difference estimate, with covariates (N=66 schools, 345 observations)

Variable	Coefficient	Standard error	Probability
Intercept	241.430	21.470	0.000
Time	0.274	0.133	0.040
IY2006_1	0.804	0.632	0.200
Itreat_1	−0.187	0.915	0.830
Iy20xtre_~1	0.410	0.902	0.640
Afam	−0.069	0.228	0.760
Asian	−0.050	0.239	0.830
Hisp	−0.091	0.219	0.670
White	−0.145	0.222	0.510
Totenrl	0.001	0.001	0.380
Lepper	0.053	0.068	0.430
Liper	−0.234	0.035	0.000
Hqtper	0.077	0.038	0.040

Source: Authors' analysis based on data described in text.

Finally, given that Massachusetts Comprehensive Assessment System (MCAS) mathematics scaled scores are transformations from raw scores,[23] researchers examined the raw scores that represent the actual numeric score that the students received on the MCAS. The results were nearly identical, for both the baseline mean and linear trend models, to the analyses reported above. These analyses are available upon request.

APPENDIX E
MASSACHUSETTS CURRICULUM FRAMEWORKS FOR GRADE 8 MATHEMATICS (MAY 2004)

Number sense and operations strand

Topic 1: Numbers

Grades 7–8:

8.N.1. Compare, order, estimate, and translate among integers, fractions and mixed numbers (rational numbers), decimals, and percents.

8.N.2. Define, compare, order, and apply frequently used irrational numbers, such as $\sqrt{2}$ and π.

8.N.3. Use ratios and proportions in the solution of problems, in particular, problems involving unit rates, scale factors, and rate of change.

8.N.4. Represent numbers in scientific notation, and use them in calculations and problem situations.

8.N.5. Apply number theory concepts, including prime factorization and relatively prime numbers, to the solution of problems.

Grade (All):

3.N.6. Select, use, and explain various meanings and models of multiplication (through 10 × 10). Relate multiplication problems to corresponding division problems, for example, draw a model to represent 5×6 and $30 \div 6$.

Topic 2: Operations

Grades 7–8:

8.N.6. Demonstrate an understanding of absolute value, for example, $|-3| = |3| = 3$.

8.N.7. Apply the rules of powers and roots to the solution of problems. Extend the Order of Operations to include positive integer exponents and square roots.

8.N.8. Demonstrate an understanding of the properties of arithmetic operations on rational numbers. Use the associative, commutative, and distributive properties; properties of the identity and inverse elements ($-7 + 7 = 0$; $\frac{3}{4} \times \frac{4}{3} = 1$); and the notion of closure of a subset of the rational numbers under an operation (the set of odd integers is closed under multiplication but not under addition).

Topic 3: Computation

Grades 7–8:

8.N.9. Use the inverse relationships of addition and subtraction, multiplication and division, and squaring and finding square roots to simplify computations and solve problems, such as multiplying by ½ or 0.5 is the same as dividing by 2.

8.N.10. Estimate and compute with fractions (including simplification of fractions), integers, decimals, and percents (including those greater than 100 and less than 1).

8.N.11. Determine when an estimate rather than an exact answer is appropriate and apply in problem situations.

8.N.12. Select and use appropriate operations (addition, subtraction, multiplication, division, and positive integer exponents) to solve problems with rational numbers (including negatives).

Patterns, relations, and algebra strand

Topic 4: Patterns, relations, and functions

Grades 7–8:

8.P.1. Extend, represent, analyze, and generalize a variety of patterns with tables, graphs, words, and, when possible, symbolic expressions. Include arithmetic and geometric progressions, such as compounding.

Topic 5: Symbols

Grades 7–8:

8.P.2. Evaluate simple algebraic expressions for given variable values, such as $3a^2 - b$ for $a = 3$ and $b = 7$.

8.P.3. Demonstrate an understanding of the identity $(-x)(-y) = xy$. Use this identity to simplify algebraic expressions, such as $(-2)(-x+2) = 2x - 4$.

Topic 6: Models

Grades 7–8:

8.P.4. Create and use symbolic expressions and relate them to verbal, tabular, and graphical representations.

8.P.5. Identify the slope of a line as a measure of its steepness and as a constant rate of change from its table of values, equation, or graph. Apply the concept of slope to the solution of problems.

Topic 7: Change

Grades 7–8:

8.P.6. Identify the roles of variables within an equation, for example, $y = mx + b$, expressing y as a function of x with parameters m and b.

8.P.7. Set up and solve linear equations and inequalities with one or two variables, using algebraic methods, models, and graphs.

8.P.8. Explain and analyze—both quantitatively and qualitatively, using pictures, graphs, charts, or equations—how a change in one variable results in a change in another variable in functional relationships, for example, $C = \pi d$, $A = \pi r^2$ (A as a function of r), $A_{\text{rectangle}} = lw$ ($A_{\text{rectangle}}$ as a function of l and w).

8.P.9. Use linear equations to model and analyze problems involving proportional relationships. Use technology as appropriate.

8.P.10. Use tables and graphs to represent and compare linear growth patterns. In particular, compare rates of change and x- and y-intercepts of different linear patterns.

Geometry strand

Topic 8: Properties of shapes

Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships.

Grades 7–8:

8.G.1. Analyze, apply, and explain the relationship between the number of sides and the sums of the interior and exterior angle measures of polygons.

8.G.2. Classify figures in terms of congruence and similarity, and apply these relationships to the solution of problems.

Topic 9: Locations and spatial relationships

Specify locations and describe spatial relationships using coordinate geometry and other representational systems.

Grades 7–8:

8.G.3. Demonstrate an understanding of the relationships of angles formed by intersecting lines, including parallel lines cut by a transversal.

8.G.4. Demonstrate an understanding of the Pythagorean theorem. Apply the theorem to the solution of problems.

Topic 10: Transformations and symmetry

Apply transformations and use symmetry to analyze mathematical situations.

Grades 7–8:

8.G.5. Use a straight-edge, compass, or other tools to formulate and test conjectures, and to draw geometric figures.

8.G.6. Predict the results of transformations on unmarked or coordinate planes and draw the transformed figure, for example, predict how tessellations transform under translations, reflections, and rotations.

Topic 11: Visualization and models

Use visualization, spatial reasoning, and geometric modeling to solve problems.

Grades 7–8:

8.G.7. Identify three-dimensional figures (prisms, pyramids) by their physical appearance, distinguishing attributes, and spatial relationships such as parallel faces.

8.G.8. Recognize and draw two-dimensional representations of three-dimensional objects (nets, projections, and perspective drawings).

Measurement strand

Topic 12: Measurable attributes and measurement systems

Grades 7–8:

8.M.2. Given the formulas, convert from one system of measurement to another. Use technology as appropriate.

Topic 13: Techniques and tools

Apply appropriate techniques, tools, and formulas to determine measurements.

Grades 7–8:

8.M.3. Demonstrate an understanding of the concepts and apply formulas and procedures for determining measures, including those of area and perimeter/circumference of parallelograms, trapezoids, and circles. Given the formulas, determine the surface area and volume of rectangular prisms, cylinders, and spheres. Use technology as appropriate.

8.M.4. Use ratio and proportion (including scale factors) in the solution of problems, including problems involving similar plane figures and indirect measurement.

8.M.5. Use models, graphs, and formulas to solve simple problems involving rates (velocity and density).

Data analysis, statistics, and probability strand

Topic 14: Data collection

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them.

Grades 7–8:

8.D.1. Describe the characteristics and limitations of a data sample. Identify different ways of selecting a sample, for example, convenience sampling, responses to a survey, random sampling.

Topic 15: Statistical methods

Select and use appropriate statistical methods to analyze data.

Grades 7–8:

8.D.2. Select, create, interpret, and utilize various tabular and graphical representations of data, such as circle graphs, Venn diagrams, scatterplots, stem-and-leaf plots, box-and-whisker plots, histograms, tables, and charts. Differentiate between continuous and discrete data and ways to represent them.

Topic 16: Inferences and predictions

Develop and evaluate inferences and predictions that are based on data.

Grades 7–8:

8.D.3. Find, describe, and interpret appropriate measures of central tendency (mean, median, and mode) and spread (range) that represent a set of data. Use these notions to compare different sets of data.

Topic 17: Probability

Understand and apply basic concepts of probability.

Grades 7–8:

8.D.4. Use tree diagrams, tables, organized lists, basic combinatorics ("fundamental counting principle"), and area models to compute probabilities for simple compound events, for example, multiple coin tosses or rolls of dice.

Algebra I course

Topic AI.N: Number sense and operations

Understand numbers, ways of representing numbers, relationships among numbers, and number systems.

Understand meanings of operations and how they relate to one another.

Compute fluently and make reasonable estimates.

Grades 9–12:

AI.N.1. Identify and use the properties of operations on real numbers, including the associative, commutative, and distributive properties; the existence of the identity and inverse elements for addition and multiplication; the existence of nth roots of positive real numbers for any positive integer n; the inverse relationship between taking the nth root of and the nth power of a positive real number; and the density of the set of rational numbers in the set of real numbers. (10.N.1)

AI.N.2. Simplify numerical expressions, including those involving positive integer exponents or the absolute value, for example, $3(2^4 - 1) = 45$, $4|3 - 5| + 6 = 14$; apply such simplifications in the solution of problems. (10.N.2)

AI.N.3. Find the approximate value for solutions to problems involving square roots and cube roots without the use of a calculator, for example, $\sqrt{}(3^2 - 1) \approx 2.8$ (10.N.3)

AI.N.4. Use estimation to judge the reasonableness of results of computations and of solutions to problems involving real numbers. (10.N.4)

Topic AI.P: Patterns, relations, and algebra

Understand patterns, relations, and functions.

Represent and analyze mathematical situations and structures using algebraic symbols.

Use mathematics models to represent and understand quantitative relationships.

Grades 9–12:

AI.P.3. Demonstrate an understanding of relations and functions. Identify the domain, range, dependent, and independent variables of functions.

AI.P.4. Translate between different representations of functions and relations: graphs, equations, point sets, and tabular.

AI.P.5. Demonstrate an understanding of the relationship between various representations of a line. Determine a line's slope and x- and y-intercepts from its graph or from a linear equation that represents the line. Find a linear equation describing a line from a graph or a geometric description of the line, for example, by using the "point-slope" or "slope y-intercept" formulas. Explain the significance of a positive, negative, zero, or undefined slope. (10.P.2)

AI.P.6. Find linear equations that represent lines either perpendicular or parallel to a given line and through a point, for example, by using the "point-slope" form of the equation. (10.G.8)

AI.P.7. Add, subtract, and multiply polynomials. Divide polynomials by monomials. (10.P.3)

AI.P.8. Demonstrate facility in symbolic manipulation of polynomial and rational expressions by rearranging and collecting terms, factoring ($a^2 - b^2 = (a + b)(a - b)$, $x^2 + 10x + 21 = (x + 3)(x + 7)$, $5x^4 + 10x^3 - 5x^2 = 5x^2 (x^2 + 2x - 1)$), identifying and canceling common factors in rational expressions, and applying the properties of positive integer exponents. (10.P.4)

AI.P.9. Find solutions to quadratic equations (with real roots) by factoring, completing the square, or using the quadratic formula. Demonstrate an understanding of the equivalence of the methods. (10.P.5)

AI.P.10. Solve equations and inequalities including those involving absolute value of linear expressions ($|x - 2| > 5$) and apply to the solution of problems. (10.P.6)

AI.P.11. Solve everyday problems that can be modeled using linear, reciprocal, quadratic, or exponential functions. Apply appropriate tabular, graphical, or symbolic methods to the solution. Include compound interest, and direct and inverse variation problems. Use technology when appropriate. (10.P.7)

AI.P.12. Solve everyday problems that can be modeled using systems of linear equations or inequalities. Apply algebraic and graphical methods to the solution. Use technology when appropriate. Include mixture, rate, and work problems. (10.P.8)

Topic AI.D: Data analysis, statistics, and probability

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them.

Select and use appropriate statistical methods to analyze data.

Develop and evaluate inferences and predictions that are based on data.

Understand and apply basic concepts of probability.

Grades 9–12:

AI.D.1. Select, create, and interpret an appropriate graphical representation (scatterplot, table, stem-and-leaf plots, circle graph, line graph, and line plot) for a set of data and use appropriate statistics (mean, median, range, and mode) to communicate information about the data. Use these notions to compare different sets of data. (10.D.1)

AI.D.2. Approximate a line of best fit (trend line) given a set of data (scatterplot). Use technology when appropriate. (10.D.2)

AI.D.3. Describe and explain how the relative sizes of a sample and the population affect the validity of predictions from a set of data. (10.D.3)

Algebra II course

Topic AII.N: Number sense and operations

Understand numbers, ways of representing numbers, relationships among numbers, and number systems.

Understand meanings of operations and how they relate to one another.

Compute fluently and make reasonable estimates.

Grades 9–12:

AII.N.1. Define complex numbers (such as $a + bi$) and operations on them, in particular, addition, subtraction, multiplication, and division.

Relate the system of complex numbers to the systems of real and rational numbers. (12.N.1)

AII.N.2. Simplify numerical expressions with powers and roots, including fractional and negative exponents.

Topic AII.P: Patterns, relations, and algebra

Understand patterns, relations, and functions.

Represent and analyze mathematical situations and structures using algebraic symbols.

Use mathematical models to represent and understand quantitative relationships.

Analyze change in various contexts.

Grades 9–12:

AII.P.1. Describe, complete, extend, analyze, generalize, and create a wide variety of patterns, including iterative and recursive patterns such as Pascal's Triangle. (12.P.1)

AII.P.2. Identify arithmetic and geometric sequences and finite arithmetic and geometric series. Use the properties of such sequences and series to solve problems, including finding the formula for the general term and the sum, recursively and explicitly. (12.P.2)

AII.P.3. Demonstrate an understanding of the binomial theorem and use it in the solution of problems. (12.P.3)

AII.P.4. Demonstrate an understanding of the exponential and logarithmic functions.

AII.P.5. Perform operations on functions, including composition. Find inverses of functions. (12.P.5)

AII.P.6. Given algebraic, numeric and/or graphical representations, recognize functions as polynomial, rational, logarithmic, or exponential. (12.P.6)

AII.P.7. Find solutions to quadratic equations (with real coefficients and real or complex roots) and apply to the solutions of problems. (12.P.7)

AII.P.8. Solve a variety of equations and inequalities using algebraic, graphical, and numerical methods, including the quadratic formula; use technology where appropriate. Include polynomial, exponential, and logarithmic functions; expressions involving the absolute values; and simple rational expressions. (12.P.8)

AII.P.9. Use matrices to solve systems of linear equations. Apply to the solution of everyday problems. (12.P.9)

AII.P.10. Use symbolic, numeric, and graphical methods to solve systems of equations and/or inequalities involving algebraic, exponential, and logarithmic expressions. Also use technology where appropriate. Describe the relationships among the methods. (12.P.10)

AII.P.11. Solve everyday problems that can be modeled using polynomial, rational, exponential, logarithmic, and step functions, absolute values and square roots. Apply appropriate graphical, tabular, or symbolic methods to the solution. Include growth and decay; logistic growth; joint ($I = Prt$, $y = k(w_1 + w_2)$), and combined ($F = G(m_1m_2)/d^2$) variation. (12.P.11)

AII.P.12. Identify maximum and minimum values of functions in simple situations. Apply to the solution of problems. (12.P.12)

AII.P.13. Describe the translations and scale changes of a given function $f(x)$ resulting from substitutions for the various parameters a, b, c, and d in $y = af(b(x + c/b)) + d$. In particular, describe the effect of such changes on polynomial, rational, exponential, and logarithmic functions. (12.P.13)

Topic AII.G: Geometry

Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships.

Specify locations and describe spatial relationships using coordinate geometry and other representational systems.

Apply transformations and use symmetry to analyze mathematical situations.

Use visualization, spatial reasoning, and geometric modeling to solve problems.

Grades 9–12:

AII.G.1. Define the sine, cosine, and tangent of an acute angle. Apply to the solution of problems. (12.G.1)

AII.G.2. Derive and apply basic trigonometric identities ($\sin^2\theta + \cos^2\theta = 1$, $\tan^2\theta + 1 = \sec^2\theta$) and the laws of sines and cosines. (12.G.2)

AII.G.3. Relate geometric and algebraic representations of lines, simple curves, and conic sections. (12.G.4)

Topic AII.D: Data analysis, statistics, and probability

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them.

Select and use appropriate statistical methods to analyze data.

Develop and evaluate inferences and predictions that are based on data.

Understand and apply basic concepts of probability.

Grades 9–12:

AII.D.1. Select an appropriate graphical representation for a set of data and use appropriate statistics (quartile or percentile distribution) to communicate information about the data. (12.D.2)

AII.D.2. Use combinatorics (such as "fundamental counting principle," permutations, and combinations) to solve problems, in particular, to compute probabilities of compound events. Use technology as appropriate. (12.D.6)

Geometry course

Topic G.G: Geometry

Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships.

Specify locations and describe spatial relationships using coordinate geometry and other representational systems.

Apply transformations and use symmetry to analyze mathematical situations.

Use visualization, spatial reasoning, and geometric modeling to solve problems.

Grades 9–12:

G.G.1. Recognize special types of polygons (such as isosceles triangles, parallelograms, and rhombuses). Apply properties of sides, diagonals, and angles in special polygons; identify their parts and special segments (such as altitudes, midsegments); determine interior angles for regular polygons. Draw and label sets of points such as line segments, rays, and circles. Detect symmetries of geometric figures.

G.G.2. Write simple proofs of theorems in geometric situations, such as theorems about congruent and similar figures, parallel or perpendicular lines. Distinguish between postulates and theorems. Use inductive and deductive reasoning, as well as proof by contradiction. Given a conditional statement, write its inverse, converse, and contrapositive.

G.G.3. Apply formulas for a rectangular coordinate system to prove theorems.

G.G.4. Draw congruent and similar figures using a compass, straightedge, protractor, or computer software. Make conjectures about methods of construction. Justify the conjectures by logical arguments. (10.G.2)

G.G.10. Apply the triangle inequality and other inequalities associated with triangles (such as the longest side is opposite the greatest angle) to prove theorems and solve problems.

G.G.11. Demonstrate an understanding of the relationship between various representations of a line. Determine a line's slope and x- and y-intercepts from its graph or from a linear equation that represents the line. Find a linear equation describing a line from a graph or a geometric description of the line, for example, by using the "point-slope" or "slope y-intercept" formulas. Explain the significance of a positive, negative, zero, or undefined slope. (10.P.2)

G.G.12. Using rectangular coordinates, calculate midpoints of segments, slopes of lines and segments, and distances between two points, and apply the results to the solutions of problems. (10.G.7)

G.G.13. Find linear equations that represent lines either perpendicular or parallel to a given line and through a point, for example, by using the "point-slope" form of the equation. (10.G.8)

G.G.14. Demonstrate an understanding of the relationship between geometric and algebraic representations of circles.

G.G.15. Draw the results, and interpret transformations on figures in the coordinate plane, for example, translations, reflections, rotations, scale factors, and the results of successive transformations. Apply transformations to the solution of problems. (10.G.9)

G.G.16. Demonstrate the ability to visualize solid objects and recognize their projections and cross sections. (10.G.10)

G.G.17. Use vertex-edge graphs to model and solve problems. (10.G.11)

G.G.18. Use the notion of vectors to solve problems. Describe addition of vectors and multiplication of a vector by a scalar, both symbolically and pictorially. Use vector methods to obtain geometric results. (12.G.3)

Topic G.M: Measurement

Understand measurable attributes of objects and the units, systems, and processes of measurement.

Apply appropriate techniques, tools, and formulas to determine measurements.

Grades 9–12:

G.M.1. Calculate perimeter, circumference, and area of common geometric figures such as parallelograms, trapezoids, circles, and triangles. (10.M.1)

G.M.2. Given the formula, find the lateral area, surface area, and volume of prisms, pyramids, spheres, cylinders, and cones, (find the volume of a sphere with a specified surface area). (10.M.2)

G.M.3. Relate changes in the measurement of one attribute of an object to changes in other attributes, for example, how changing the radius or height of a cylinder affects its surface area or volume. (10.M.3)

G.M.4. Describe the effects of approximate error in measurement and rounding on measurements and on computed values from measurements. (10.M.4)

G.M.5. Use dimensional analysis for unit conversion and to confirm that expressions and equations make sense. (12.M.2)

Precalculus course

Topic PC.N: Number sense and operations

Understand numbers, ways of representing numbers, relationships among numbers, and number systems.

Understand meanings of operations and how they relate to each other.

Compute fluently and make reasonable estimates.

Grades 9–12:

PC.N.1. Plot complex numbers using both rectangular and polar coordinates systems. Represent complex numbers using polar coordinates, that is, $a + bi = r(\cos\theta + i\sin\theta)$. Apply DeMoivre's theorem to multiply, take roots, and raise complex numbers to a power.

Topic PC.P: Patterns, relations, and algebra

Understand patterns, relations, and functions.

Represent and analyze mathematical situations and structures using algebraic symbols.

Use mathematical models to represent and understand quantitative relationships.

Analyze change in various contexts.

Grades 9–12:

PC.P.1. Use mathematical induction to prove theorems and verify summation formulas, for example, verify

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

PC.P.2. Relate the number of roots of a polynomial to its degree. Solve quadratic equations with complex coefficients.

PC.P.3. Demonstrate an understanding of the trigonometric functions (sine, cosine, tangent, cosecant, secant, and cotangent). Relate the functions to their geometric definitions.

PC.P.4. Explain the identity $\sin^2\theta + \cos^2\theta = 1$. Relate the identity to the Pythagorean theorem.

PC.P.5. Demonstrate an understanding of the formulas for the sine and cosine of the sum or the difference of two angles. Relate the formulas to DeMoivre's theorem and use them to prove other trigonometric identities. Apply to the solution of problems.

PC.P.6. Understand, predict, and interpret the effects of the parameters a, ω, b, and c on the graph of $y = \text{as in } (\omega(x - b)) + c$; similarly for the cosine and tangent. Use to model periodic processes. (12.P.13)

PC.P.7. Translate between geometric, algebraic, and parametric representations of curves. Apply to the solution of problems.

PC.P.8. Identify and discuss features of conic sections: axes, foci, asymptotes, and tangents. Convert between different algebraic representations of conic sections.

PC.P.9. Relate the slope of a tangent line at a specific point on a curve to the instantaneous rate of change. Explain the significance of a horizontal tangent line. Apply these concepts to the solution of problems.

Topic PC.G: Geometry

Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships.

Specify locations and describe spatial relationships using coordinate geometry and other representational systems.

Apply transformations and use symmetry to analyze mathematical situations.

Use visualization, spatial reasoning, and geometric modeling to solve problems.

Grades 9–12:

PC.G.1. Demonstrate an understanding of the laws of sines and cosines. Use the laws to solve for the unknown sides or angles in triangles. Determine the area of a triangle given the length of two adjacent sides and the measure of the included angle. (12.G.2)

PC.G.2. Use the notion of vectors to solve problems. Describe addition of vectors, multiplication of a vector by a scalar, and the dot product of two vectors, both symbolically and geometrically. Use vector methods to obtain geometric results. (12.G.3)

PC.G.3. Apply properties of angles, parallel lines, arcs, radii, chords, tangents, and secants to solve problems. (12.G.5)

Topic PC.M: Measurement

Understand measurable attributes of objects and the units, systems, and processes of measurement.

Apply appropriate techniques, tools, and formulas to determine measurements.

Grades 9–12:

PC.M.1. Describe the relationship between degree and radian measures, and use radian measure in the solution of problems, in particular problems involving angular velocity and acceleration. (12.M.1)

PC.M.2. Use dimensional analysis for unit conversion and to confirm that expressions and equations make sense. (12.M.2)

Topic PC.D: Data analysis, statistics, and probability

Formulate questions that can be addressed with data collect, organize, and display relevant data to answer them.

Select and use appropriate statistical methods to analyze data.

Develop and evaluate inferences and predictions that are based on data.

Understand and apply basic concepts of probability.

Grades 9–12:

PC.D.1. Design surveys and apply random sampling techniques to avoid bias in the data collection. (12.D.1)

PC.D.2. Apply regression results and curve fitting to make predictions from data. (12.D.3)

PC.D.3. Apply uniform, normal, and binomial distributions to the solutions of problems. (12.D.4)

PC.D.4. Describe a set of frequency distribution data by spread (variance and standard deviation), skewness, symmetry, number of modes, or other characteristics. Use these concepts in everyday applications. (12.D.5)

PC.D.5. Compare the results of simulations (e.g. random number tables, random functions, and area models) with predicted probabilities. (12.D.7)

Scaled score ranges of the Massachusetts Comprehensive Assessment System by performance level and year

TABLE E1
Scaled score ranges of the Massachusetts Comprehensive Assessment System by performance level and year

	2001–02		2003–05
Warning	200–203	Warning	200–202
	204–207		204–206
	208–211		208–210
	212–215		212–214
	216–219		216–218
Needs improvement	220–223	Needs improvement	220–222
	224–227		224–226
	228–231		228–230
	232–235		232–234
	236–239		236–238
Proficient	240–243	Proficient	240–242
	244–247		244–246
	248–251		248–250
	252–255		252–254
	256–259		256–258
Advanced	260–263	Advanced	260–262
	264–267		264–266
	268–271		268–270
	272–275		272–274
	276–280		276–280

Source: 2001 data are from http://www.doe.mass.edu/mcas/2001/interpretive_guides/fullguide.pdf; 2002 data are from http://www.doe.mass.edu/mcas/2002/interpretive_guides/fullguide.pdf; 2003 data are from http://www.doe.mass.edu/mcas/2003/interpretive_guides/full.pdf; 2004 data are from http://www.doe.mass.edu/mcas/2004/interpretive_guides/full.pdf; 2005 data are from http://www.doe.mass.edu/mcas/2005/interpretive_guides/full.pdf.

NOTES

Many thanks to Thomas Hanson (WestEd), for his methodological expertise, patience, and Stata syntax writing ability, and to Laura O'Dwyer (Boston College), Craig Hoyle (EDC), Thomas Cook (Northwestern University), William Shadish (UC Merced), Howard Bloom (MDRC), David Wilson (George Mason University), and Natalie Lacireno-Paquet (Learning Innovations) for their assistance.

1. An effect size of 0.40 means that the experimental group is performing, on average, about 0.40 of a standard deviation better than the control group (Valentine and Cooper, 2003). An effect size of 0.40 represents a roughly 20 percent improvement over the control group.

2. Bloom (2003) might argue that 2006 should be interpreted as an "implementation" year rather than a post-test year.

3. Scaled scores are constructed by converting students' raw scores (say, the number of questions correct) on a test to yield comparable results across students, test versions, or time. Raw scores were also examined and produced similar results to the scaled scores (see appendix C).

4. To report adequate yearly progress determinations the MCAS has four performance levels: warning (scoring 200–19), needs improvement (220–39), proficient (240–59), and advanced (260–80).

5. Statistical power refers to the ability of the statistical test to detect a true treatment effect, if one exists. Although there are other design features that can influence the statistical power of a test, researchers are generally most concerned with sample size, because it is the component they have the most control over and can normally plan for.

6. Using Stata's program for computing statistical power in repeated measures designs (such as time series)—and assuming a type I error rate (α) of 0.05 (two-sided), correlations between the annual test score measures of 0.70, and statistical power of 0.80—there was sufficient statistical power for the originally designed study (with 25 program schools and 50 comparison schools) to detect a post-intervention difference between program and comparison schools of 0.41 standard deviations. The loss of three program schools and six comparison schools (with power at 0.80) increased the minimum detectible effect size to 0.44 (the actual pretest–post-test correlation was 0.74). An effect of such magnitude would be generally considered moderate (Cohen, 1988; Lipsey & Wilson, 1993), although it is relatively large by education intervention standards (Bloom, Richburg-Hayes, & Black, 2005).

7. The 2001–03 achievement data were not publicly available and had to be requested from the Massachusetts Department of Education.

8. The initial study plan called for including both sixth grade and eighth grade. But that set would have excluded too many of the program schools and a large pool of comparison schools.

9. Student level data for 2001–03 had to be aggregated at the school level. The 2001–03 achievement data were provided by the Massachusetts Department of Education at the student level but were not linked to the student level demographic data for the same years.

10. Twenty-five schools were originally identified as treatment schools, but three were found to be newly configured or established schools, meaning they had no data on student achievement for previous years and could not be included in the time series analysis. They were therefore excluded from the study.

11. The authors are very grateful to Thomas Hanson, WestEd, for his assistance throughout

on the covariate matching procedure, and to Craig Hoyle (EDC) and William Shadish (University of California, Merced) for advice.

12. Percentages are preferred over total numbers. For example, 100 Asian students in a Boston school might be 10 percent of the school's total population; in a different school, it could represent 50 percent or more of the total enrolled students.

13. To create SCI, the following formula was used:

$$\text{Academic Performance} = \alpha + \beta 1{*}\text{Enrollment} + \beta 2{*}\text{Income} + \beta 3{*}\text{English Language Learners} + \beta k{*}\text{Race/Ethnic} + \varepsilon$$

α represents the intercept, or the value of Y if X is 0. In this instance, it would be the predicted score on the 2005 Mathematics CPI if enrollment, income, English language learners and race/ethnicity were all zero.

Academic Performance represented the average school performance on the 2005 Composite Performance Index for eighth-grade mathematics. The 2005 CPI is actually an average of two years of the school's eighth-grade mathematics scores on the Massachusetts Comprehensive Assessment System (MCAS) test.

Enrollment is the school's total enrollment.

Income is the percentage of students classified as low income.

Race/Ethnic is the percentage of students of different racial or ethnic groups (with percentage of White students the omitted reference group).

$\beta 1$, $\beta 2$, $\beta 3$, and βk represent the relationships of school enrollment, percentage of low-income students, percentage of English language learners, and percentage of different race/ethnicity groups to the 2005 Mathematics CPI scores, respectively.

ε represents the error term. Error in this instance refers to the difference between the predicted and actual Mathematics CPI 2005 scores. In this study, it represents the adjusted mathematics score (2005 CPI Mathematics Score minus the SCI or predicted score).

14. This section on Mahalanobis Distance was informed by the relevant section in StatSoft, Inc. (2001).

15. The measure was created by the noted statistician, Prasanta Chandra Mahalanobis, in 1930. See http://en.wikipedia.org/wiki/Prasanta_Chandra_Mahalanobis.

16. Many thanks to Thomas Hanson of WestEd, who drafted the Stata syntax for creating the Mahalanobis Distance measures. The syntax was:

```
*Compute Mahalanobis Distance
matrix drop _all
mkmat scix true, matrix(xvar)
matrix accum cov = scix true, noc dev
matrix cov = cov/(r(N)-1)
matrix factorx= (xvar) * (inv(cov)) * (xvar')
matrix factor= (vecdiag(factorx))'
svmat factor, names(factor)

sort dstlcl factor
```

17. It was too difficult to replicate in SPSS the Mahalanobis Distance method used in Stata. Unfortunately, the original Stata syntax is very idiosyncratic and would require extensive programming time to convert it into similar syntax in SPSS.

18. David Kantor provided excellent guidance to the team in the use of the Mahapick program.

19. To prevent individual schools from being identified, tables detailing the samples of schools that each iteration produced have been omitted.

20. There is also a statistically significant difference between the program and comparison schools on the percentage of students enrolled at the school classified as "Hawaiian/Pacific Islander." This is not troubling because it represents an extremely small share of the student population.

21. The Stata syntax used was:

```
renames mthscaledgrd62001 mthscaledgrd62002
mthscaledgrd62003
mthscaledgrd62004 mthscaledgrd62005
mthscaledgrd62006 mt82001 mt82002sc
mt82003sc mt82004sc mt82005sc mt82006sc
\ math6s2001 math6s2002 math6s2003
math6s2004 math6s2005 math6s2006
math8s2001 math8s2002 math8s2003
math8s2004 math8s2005 math8s2006

gen byte treat=1 if stygrouppetrosino == 2
replace treat=0 if stygrouppetrosino == 3

keep school treat grsix greight totenrl-mrnh
lepper liper hqtper math6s2001 math6s2002
math6s2003 math6s2004 math6s2005
math6s2006 math8s2001 math8s2002
math8s2003 math8s2004 math8s2005
math8s2006

order school treat grsix greight totenrl-mrnh
lepper liper hqtper math6s2001 math6s2002
math6s2003 math6s2004 math6s2005
math6s2006 math8s2001 math8s2002
math8s2003 math8s2004 math8s2005
math8s2006

reshape long math6s math8s, i(school) j(year)

gen byte y2006=0
replace y2006=1 if year == 2006
gen time=year-2001

egen schlid=group(school)

order schlid school year time y2006
compress

save benchmark1, replace

log using benchmark1, text replace
```

22. The following is the Stata syntax for the analyses. The relevant analyses are the "difference-in-difference" estimates (bold), as they represent the pretest–post-test estimate for the program group schools compared with the comparison schools.

```
* Baseline Mean Model - no covariates.
xtreg math8s y2006 if treat == 0, i(schlid)
xtreg math8s y2006 if treat == 1, i(schlid)

* Difference-in-Difference Baseline Mean
Model - no covariates
xi: xtreg math8s i.y2006*i.treat, i(schlid)

* Linear Baseline Trend Model - no covariates.
xtreg math8s time y2006 if treat == 0,
i(schlid)
xtreg math8s time y2006 if treat == 1,
i(schlid)

* Difference-in-Difference Linear Baseline
Trend Model - no covariates
xi: xtreg math8s time i.y2006*i.treat, i(schlid)

* Baseline Mean Model - covariates.
xtreg math8s y2006 afam asian hisp white totenrl
lepper liper hqtper if treat == 0, i(schlid)
xtreg math8s y2006 afam asian hisp white totenrl
lepper liper hqtper if treat == 1, i(schlid)

* Difference-in-Difference Baseline Mean
Model - covariates
xi: xtreg math8s i.y2006*i.treat afam asian
hisp white totenrl lepper liper hqtper, i(schlid)

* Linear Baseline Trend Model - covariates.
xtreg math8s time y2006 afam asian hisp
white totenrl lepper liper hqtper if treat == 0,
i(schlid)
xtreg math8s time y2006 afam asian hisp
white totenrl lepper liper hqtper if treat == 1,
i(schlid)
```

```
* Difference-in-Difference Linear Baseline Trend Model - no covariates
xi: xtreg math8s time i.y2006*i.treat afam asian hisp white totenrl lepper liper hqtper, i(schlid)
```

23. MCAS scaled scores are transformations from raw scores to criterion-referenced cut-points established during standard setting exercises that took place in 1998 for grade 8 mathematics. Each year the MADOE finds the raw score that is most equal to the difficulty level set to 220, 240, and 260 and then builds linear equations to determine two different point intervals between 200 and 220, 220 and 240, 240 and 260, and 260 and 280.

REFERENCES

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: interaction of research and policy. *Educational Measurement: Issues and Practice,* 25(4), 36–46.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice,* 5(1).

Black, P., & Wiliam, D. (1998b). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan,* 80(2). Retrieved June 21, 2007, from http://www.pdkintl.org/kappan/kbla9810.htm

Bloom, B. (1984). The search for methods of instruction as effective as one-to-one tutoring. *Educational Leadership,* 41(8), 4–17.

Boston, C. (2002). *The concept of formative assessment.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms. *Evaluation Review,* 27(1), 3–49.

Bloom, H. S., Richburg-Hayes, L., and Black, A. (2005). "Using covariates to improve precision: empirical guidance for studies that randomize schools to measure the impacts of educational interventions." Working Paper. New York: MDRC.

Cohen, J. (1988). *Statistical power for the behavioral sciences.* Mahway, NJ: Lawrence Erlbaum.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: design and analysis for field settings.* Chicago: Rand McNally.

Cotton, K. (1996). *School size, school climate, and student performance.* School Improvement Research Series, Series X, Close-Up 20. Portland, OR: Northwest Regional Educational Laboratory. Retrieved from http://www.nwrel.org/scpd/sirs/10/c020.html

Hannaway, J. (2005). Poverty and student achievement: a hopeful review. In J. Flood and P. Anders (Eds.), *Literacy development of students in urban schools* (pp. 3–21). Newark, DE: International Reading Association.

Herman, J. L., & Baker, E. L. (2005). Making benchmark testing work for accountability and improvement: quality matters. *Educational Leadership,* 63(3), 48–55.

Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap.* Washington, DC: Brookings Institution.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist,* 48, 1181–1209.

Massachusetts Department of Education. (2007). *About the MCAS.* Retrieved June 23, 2007, from http://www.doe.mass.edu/mcas/about1.html

Olson, L. (2005, November). Benchmark assessments offer regular checkups on student achievement. *Education Week,* 25 (13), 13–14.

Rubin, D. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics* 36(2), 293–298.

Shadish,W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized and casual inference.* Boston, MA: Houghton Mifflin.

StatSoft, Inc. (2001). *Electronic statistics textbook.* Tulsa, OK: StatSoft. Retrieved June 11, 2007, from http://www.statsoft.com/textbook/stathome.html

Taylor, R. (1994). *Research methods in criminal justice.* New York: McGraw-Hill.

U. S. Department of Education. (2007). *Improving math performance.* Washington, DC: Author. Retrieved June 11, 2007, from http://www.ed.gov/programs/nclbbrs/math.pdf

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: issues in the interpretation of effect sizes.* Washington, DC: What Works Clearinghouse.