REL Technical Brief

REL 2008 – No. 002

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

A second follow-up year for *Measuring how benchmark assessments affect student achievement*

Prepared by

Susan Henderson
Learning Innovations at WestEd

Anthony Petrosino
Learning Innovations at WestEd

Sarah Guckenburg
Learning Innovations at WestEd

Stephen Hamilton
Learning Innovations at WestEd

April 2008

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.

REL Technical Briefs is a new report series from Fast Response Projects that helps educators obtain evidence-based answers to their specific requests for information on pressing education issues. REL Technical Briefs offer highly targeted responses across a variety of subjects, from reviews of particular studies or groups of studies on No Child Left Behind Act implementation issues, to compilations or quick summaries of state or local education agency data, appraisals of particular instruments or tools, and very short updates of Issues & Answers reports. All REL Technical Briefs meet IES standards for scientifically valid research.

April 2008

REL Northeast and Islands received a request to update the Issues & Answers report, *Measuring how benchmark assessments affect student achievement* (http://ies.ed.gov/ncee/edlabs/projects/project.asp?id=43) with a second year of follow-up data to assess whether there were differences in grade 8 mathematics achievement between program and comparison schools.

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.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *REL Technical Brief—a second follow-up year for "Measuring how benchmark assessments affect student achievement"* (REL Technical Brief, REL Northeast and Islands 2007–No. 002). 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 REL Technical Brief is available on the regional educational laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

This technical brief examines whether, after two years of implementation, schools in Massachusetts using quarterly benchmark exams aligned with state standards in middle school mathematics showed greater gains in student achievement than those not doing so. A quasi-experimental design, using covariate matching and comparative interrupted time-series techniques, was used to assess school differences in changes in mathematics performance between program and comparison schools. Following up on an earlier report with just one year of post-implementation data, the study found no significant differences between schools using this practice and those not doing so after two years.

The brief summarizes findings from a follow-up study to the Issues & Answers report, "Measuring how benchmark assessments affect student achievement" (REL 2007–No. 039). The follow-up study adds another year of post-implementation data to examine the impact of benchmark assessments on grade 8 mathematics achievement, using the same data sources, methods, and reporting as the original study.

The study examines whether, after two years of implementation, schools in Massachusetts using quarterly benchmark exams aligned with state standards in middle school mathematics showed greater gains in student achievement than those not doing so. A quasi-experimental design, using covariate matching and comparative interrupted time-series techniques, was used to assess differences in changes in mathematics performance between program and comparison schools.

The follow-up study finds no significant differences between schools using this practice and those not doing so after two years. Limitations include the lack of data on what benchmark assessment practices comparison schools may be using, having only 22 treatment and 44 comparison schools, and having only two years of post-implementation data—perhaps still too few to observe an impact from the intervention.

Technical brief

Why this brief?

This technical brief summarizes findings from a follow-up study to the Issues & Answers report, "Measuring how benchmark assessments affect student achievement" (REL 2007–No. 039). That report examined the impact of benchmark assessments on one year of post-implementation data (for 2006 only). This brief examines the impact of benchmark assessments on grade 8 mathematics achievement using two years of post-implementation data (for 2006 and 2007).

Although the initial report did not find statistically significant differences between the program and comparison schools after one year, it received high visibility among regional policymakers because of the interest in the topic and the study's quasi-experimental methodology. The Massachusetts Department of Elementary and Secondary Education and members of the Regional Education Laboratory Northeast and Islands Governing Board asked the study leaders to repeat the statistical analysis conducted in the initial study, including a second year of post-implementation data, to determine whether benchmark assessments affected math achievement. The results can inform upcoming legislation on benchmark assessment policy in Massachusetts and Rhode Island and may be of interest to education decisionmakers throughout the region.

Analytical approach

As in the initial study, the research team used a quasi-experimental design to determine whether schools using quarterly benchmark exams made greater gains in mathematics achievement than schools that did not. The project compared 22 Massachusetts middle schools that received grants under a pilot program to implement a particular benchmark assessment with 44 statistically matched schools that did not receive the grants.

The follow-up study examined the same comparison schools, identified through a covariate matching procedure (Henderson, Petrosino, Guckenburg, & Hamilton 2007). The covariate matching procedure produced two groups of schools equated on prior math performance (using pretest data) and sociodemographic characteristics (using a composite sociodemographic index). Although there were two statistically significant differences between the two matched groups (proportions of African-American students and of Hawaiian and Pacific Islander students), the study controlled for these and other covariates in the analysis.

The team verified that each of the 22 treatment and 44 comparison schools remained intact middle schools and reported grade 8 math achievement scores by examining their school profiles on the Massachusetts Elementary and Secondary Education web site (http://profiles.doe.mass.edu). After verifying the schools, researchers obtained the 2007 grade 8 math scores from the Massachusetts Comprehensive Assessment System (MCAS) achievement test and updated the original project database.

To analyze the data, the researchers used the same descriptive and interrupted time-series methods as the initial report (Henderson, Petrosino, Guckenburg, & Hamilton 2007). But each analysis included a second year of post-intervention data to identify any departure from trend (see Bloom 2003; Henderson, Petrosino, Guckenburg, & Hamilton 2007).

The research team used a quasi-experimental design to determine whether schools using quarterly benchmark exams made greater gains in mathematics achievement than schools that did not

Results

The researchers repeated the descriptive analysis of the original report, adding a second year of post-implementation data from 2007. Scaled means on the grade 8 MCAS mathematics test for 2001–07 indicated that scores for both groups increased slightly during the past two

TABLE 1
Mean scaled grade 8 mathematics scores for program and comparison schools in the Massachusetts Comprehensive Assessment System, 2001–07

Year	Treatment 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
2007	229.42	227.80

Source: Authors' analysis based on data described in Henderson, Petrosino, Guckenburg, & Hamilton 2007.

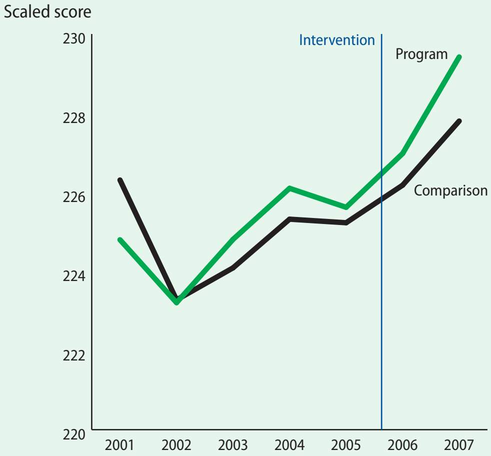
years but that the program schools did about 1.5 points better (table 1).

This finding is similarly evident in plotting the trend mean scaled grade 8 math scores. Although the scaled scores for both groups increased after the intervention in 2006, the increase for program schools seems to be greater (figure 1).

There was a statistically significant increase in the mean scaled grade 8 mathematics scores for program schools when taking the follow-up period (2006 and 2007) into account. The program schools had slightly higher mean scaled scores than expected without the program—3.2 points on the grade 8 MCAS mathematics test (see appendix A, table A2). But this small, statistically significant increase also occurred in the comparison schools, where mean scaled scores were slightly above the predicted trend (appendix A, table A3). The increase in comparison schools was 2.2 points on the grade 8 MCAS math test.

Although both groups improved, the question is whether the difference between the groups was significant. Difference-in-difference analysis showed that although the program effect grew to 0.98 of a mathematics test point from about 0.38 in the year-one analysis (see appendix A, table A4), the difference was not statistically significant. The most likely interpretation remains that the achievement of both groups is increasing slightly and that—although the program schools seem to be improving a little more—the difference could have been due to chance rather than to any program effect. So, though the trend line for the program schools, at least looking at scaled scores alone, was slightly higher than that of the comparison schools in 2007, the small increase for the program schools cannot confidently be attributed to the benchmark assessments and may be due to chance alone.

FIGURE 1
Mean scaled grade 8 mathematics scores for program and comparison schools in the Massachusetts Comprehensive Assessment System, 2001–07



Source: Authors' analysis based on data described in Henderson, Petrosino, Guckenburg, & Hamilton 2007.

Limitations

The follow-up analysis found no statistically significant difference between schools in their first two years implementing quarterly benchmark exams in middle school mathematics and those not doing so. There was a slightly larger gain on grade 8 MCAS scaled scores in math for the program schools in 2007. But the slight gain for scaled scores could not confidently be attributed to the benchmarking program. That conclusion might, however, be due to data limitations rather than to benchmark assessments being ineffective.

No data on benchmarking practices in comparison schools. As with the initial analysis, the follow-up included no data 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 benchmark assessment (Olson 2005), so it is possible that at least some comparison schools implemented their own versions of benchmarking. Given the prevalence of formative assessments under the No Child Left Behind Act of 2001, it is highly unlikely that a project with strictly controlled conditions could be implemented (that is, with schools using no benchmark assessments at all as the comparison group). So, the analyses may simply have compared a set of program schools receiving one type of benchmark intervention with comparison schools receiving other benchmarking types. If true, the findings would not be surprising.

Low statistical power. The study remained underpowered. That means that a small but important treatment effect for benchmarking could have gone undetected because there were only 22 program and 44 comparison schools. This highlights the need for future research to examine student-level data. Doing so would increase statistical power and enable researchers to explore the possible masking of nontrivial effects for subgroups. Student-level data would also enable researchers to examine performance by content strand, examining how districts performed on the standards they have prioritized through their benchmark assessments.

Too early to observe impact. With only two years of post-implementation data, it may still be too early to observe any impact from the intervention. Examining a third year of post-implementation data in the interrupted time-series design would enable researchers to better assess the impact.

With only two years of post-implementation data, it may still be too early to observe any impact from the intervention

References

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

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

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.

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

Appendix A. Details on the analysis

Analyses for two models, a baseline trend and linear trend model, were run for program and comparison schools separately to determine whether there was a post-interruption effect (that is, after the intervention in 2006 and 2007). The analyses measure a difference-in-difference effect (the effect between program and comparison schools). Covariates were then introduced to determine whether any estimates changed for time or for program impact when variables such as the percentage of African-American students enrolled at the schools were introduced. Table A1 lists variables used in the analysis.

First, using a "baseline mean model" as described by Bloom (2003), the researchers investigated whether there was a perceptible change after the implementation of the benchmarking program—between 2001–05 and the follow-up period, 2006–07. This was done for comparison schools and program schools separately. For program schools alone the increase in the math achievement scores in 2006 and 2007 was significant (variable "Ifollowup_1" in table A2). This increase represents a 3.20 test point improvement over what would have been expected without the program.

TABLE A1
Variables used in the analysis

Variable	Description
Afam	Percentage of students enrolled in the school who are African-American.
Asian	Percentage of students enrolled in the school who are Asian.
Hisp	Percentage of students enrolled in the school who are Hispanic.
Hqtper	Percentage of highly qualified teachers at the school.
Ifollowup_1	Effect for two follow-up years.
IfolXtre~1	Difference-in-difference estimate from whether schools were in the program.
Itreat_1	Effect from being in the program.
Intercept	Mean scores.
Iy20xtre~1	Interaction term between the year 2006 and whether a school was in the program
Lepper	Percentage of students in the school classified as limited English proficiency
Liper	Percentage of students in the school classified as low income
Totenrl	Number of students enrolled at the school.
White	Percentage of students enrolled in the school who are white.

TABLE A2
Baseline mean model, program schools only (22 schools, 137 observations)

Variable	Coefficient	Standard error	Probability
Intercept	225.01	0.873	0.000
Ifollowup_1	3.20	0.492	0.000

Source: Authors' analysis based on data described in Henderson, Petrosino, Guckenburg, & Hamilton 2007.

Table A3 underscores the importance of including a comparison group in the time-series analysis. For the comparison schools alone the higher improvement is also statistically significant, representing a gain of about 2.25 test points. Assessing the program based solely on the results in table A2 would have mistakenly attributed the gain in math scores to the benchmarking initiative.

Even so, a one point difference in scaled math scores remains between the program and comparison schools. Is the difference statistically significant? The follow-up difference-in-difference analysis in table A4 shows that the 0.94 difference (variable "Ifollowxtre~1") is not statistically significant. In the initial study the effect was 0.38. Given the modest improvement for program schools, further follow-up analysis may be warranted.

In the linear trend model the difference-in-difference estimates are very similar to the results in the baseline mean model (table A5). Again, the program impact is 0.92 scaled math points, but the difference is not significant and could have occurred by chance.

TABLE A3

Baseline mean model, comparison schools only (44 schools, 272 observations)

Variable	Coefficient	Standard error	Probability
Intercept	224.71	0.785	0.000
Ifollowup_1	2.25	0.465	0.000

Source: Authors' analysis based on data described in Henderson, Petrosino, Guckenburg, & Hamilton 2007.

TABLE A4

Baseline mean model, difference-in-difference estimate (66 schools, 409 observations)

Variable	Coefficient	Standard error	Probability
Intercept	224.71	0.734	0.000
Ifollowup_1	2.26	0.429	0.000
Itreat_1	0.292	0.292	0.82
IfolXtre~1	0.940	0.744	0.21

Source: Authors' analysis based on data described in Henderson, Petrosino, Guckenburg, & Hamilton 2007.

TABLE A5

Linear trend model, difference-in-difference estimate (66 schools, 409 observations)

Variable	Coefficient	Standard error	Probability
Intercept	223.98	0.800	0.000
Time	0.322	0.135	0.02
Ifollowup_1	1.21	0.610	0.05
Itreat_1	0.308	1.28	0.81
IfolXtre_~1	0.924	0.738	0.21

Source: Authors' analysis based on data described in Henderson, Petrosino, Guckenburg, & Hamilton 2007.

Table A6 shows that when covariates are introduced into the difference-in-difference analysis, results are similar to the baseline mean model (0.97 of a scaled test point gain, compared with 0.98 under the baseline mean model).

TABLE A6
Linear trend model, difference-in-difference estimate, with covariates (66 schools, 409 observations)

Variable	Coefficient	Standard error	Probability
Intercept	243.49	19.35	0.000
Time	0.366	0.138	0.01
Ifollowup_1	1.11	0.621	0.07
Itreat_1	–0.475	0.846	0.58
IfolXtre_~1	0.971	0.748	0.19
Afam	–0.114	0.206	0.58
Asian	–0.091	0.216	0.67
Hisp	–0.128	0.198	0.52
White	–0.175	0.200	0.38
Totenrl	0.002	0.002	0.27
Lepper	0.066	0.062	0.29
Liper	–0.234	0.032	0.000
Hqtper	0.087	0.035	0.01

Source: Authors' analysis based on data described in Henderson, Petrosino, Guckenburg, & Hamilton 2007.