# The Impact of Indiana's System of Benchmark Assessments on Mathematics Achievement

by

Spyros Konstantopoulos

Michigan State University

Shazia Miller

Arie van der Ploeg

AIR

Cheng-Hsien Li

Anne Traynor

Michigan State University

## Abstract

Benchmark assessments have become increasingly common in education research the last few years. We use high quality data from a large-scale school-level cluster randomized experiment to examine the effects of *mClass and Acuity* benchmark assessments on mathematics achievement in 59 schools in the State of Indiana. Results indicate that the treatment effects are positive, but not consistently significant. The treatment effects are smaller in lower grades (i.e., kindergarten to second grade), but larger in upper grades (i.e., third to eighth grade). Significant treatment effects are detected in grades 3 to 8, especially in fifth and sixth grade mathematics.

The last 10 years with the passage of the No Child Left Behind (NCLB) Act, states were required to introduce school accountability systems that measure annual gains in student achievement. The NCLB mandates resulted in educational reforms through an abundance of school interventions that aimed to improve student achievement (Dunn & Malvenon, 2009). Among the various assessment solutions proposed as methods of improving student performance are diagnostic and formative assessments. Although there may not be an explicit agreed upon universal definition for such types of assessments, the definition typically involves a process that is designed to help teachers to use assessment-based evidence to facilitate ongoing learning and instruction (Dunn & Malvenon, 2009). Regardless, these assessments are nowadays considered by many educational leaders, in states and districts throughout the United States, as an effective lever to increase student achievement (Carlson, Borman, & Robinson, 2011). Assessment based models of educational reform are common and take advantage of the data obtained from student assessments in order to change instruction and improve student achievement. Such types of assessments are data-driven and aim to improve ongoing classroom instruction and provide feedback on student performance in order to promote learning for all students. The underlying principle is that via up to date diagnostic information about students' learning teachers can offer ongoing constructive feedback and individualized instruction that meets the students' needs and improves student learning further.

Typically data-driven assessments are implemented in entire schools and, thus, are considered whole school interventions. Currently, there is a need to evaluate rigorously school interventions at a large scale and examine the consistency of the intervention at the state level. The state of Indiana was the first to implement statewide technology-supported interim formative assessments to be taken by all K–8 students multiple times each school year. Indiana expects

teachers to use the constantly updated diagnostic information about student learning to improve ongoing instruction for individual students or for the whole classroom and ultimately increase student achievement. The interim assessments are hypothesized to help teachers identify areas for instructional need by providing immediate, detailed insight on student strengths and weaknesses. In 2008 the Indiana Department of Education (IDOE) began the roll-out of its system of interim assessments. The theory undergirding this decision was based on the premise that education is about decreasing gaps between students' current and intended knowledge. Assessment may be viewed as the measurement of these gaps, providing critical feedback and more exact documentation of the gaps between current and desired status, with repeated measurements documenting changing gaps.

In the present study we provide evidence of the effectiveness of interim assessments on mathematics achievement in Indiana schools. We designed and conducted a cluster randomized experimental to test the hypothesis that teachers who have access to objective data to monitor student progress to guide their choices about day-to-day instruction will produce students who perform higher on state assessments. We used data from nearly 60 schools in the State of Indiana that were randomly assigned to a treatment and control group. Approximately one half of the schools implemented the intervention and evaluated whether or not interim assessments lead to improved instructional practices and student outcomes.

We designed and conducted a rigorous experimental study and collected high quality data to determine the effectiveness of the intervention on student achievement. We examined whether interim assessments implemented by schools in Indiana produced significant effects on student performance on the state's annual Indiana Statewide Testing for Educational Progress–Plus (ISTEP+) measures. Because the data were produced from a well-designed randomized

experiment our estimates have high internal validity and justify causal inferences about the intervention effects (Shadish, Cook, & Campbell, 2002). In addition, because our sample included multiple schools (i.e., 59 schools) from various school districts in the state of Indiana our estimates should have higher external validity than those obtained from convenient and localized samples. To account for the nesting of students within schools we employed multilevel models to examine the effectiveness of the intervention. The results of this study are of significance to education researchers and policymakers, given their current level of interest in assessment levers to accelerate school improvement.

## Literature Review

Increased attention on ways to assess students in order to differentiate and individualize instruction has recently emerged in contrast to more traditional summative testing. Instead of summing up class, school, or district results against a defined set of standards and as part of an accountability system, formative assessment or assessment for learning focuses on active feedback loops that assist teachers and students during the learning process (Heritage, 2010; Perie, Marion, & Gong, 2007; Sadler, 1989; Stiggins, 2002). For example, Perie et al. (2007) argue that formative assessment is administered by the teacher for the explicit purpose of diagnosing student learning, identifying student gaps in knowledge and understanding, and redirecting teaching and learning. Over the last decade, much of the discussion of formative assessment has been dichotomous to that of summative assessment.

Recently, however, and because of the pressure on school and district administrators to produce results on state summative assessments, vendors of data-driven products have

introduced to states and districts assessments that have been referred to as benchmark assessments (Perie et al., 2007). These assessment products have the potential to improve student performance and help schools accomplish adequate yearly progress (Perie et al., 2007). The term interim assessment has also emerged lately as a benchmark assessment that is typically administered on a smaller scale (e.g., school- or district-wide) than those of summative assessments (Perie et al., 2007). These assessments are administered multiple times per academic year. Although teachers can individualize and differentiate instruction with interim assessment data, the tools are not created specifically for or limited to those uses.

In this paper we make use of the term interim assessment, but we also use the terms data-driven and benchmark assessments, which are broader terms. The effects of data-driven decision making such as interim assessments on student achievement have only recently been documented in the literature. Some researchers have argued that although interim assessment is currently defined more clearly, it is unclear that the expectations that district and school staff have for these tests have been sufficiently realized (Bulkley, Christman, Goertz, & Lawrence, 2010). In particular, data provided by district leaders, school principals, and school teachers in schools that have implemented benchmark assessments suggest that the benchmark assessment value is based on the hope of establishing a link between district policy and teacher instruction. In other studies, focus group and survey data of school teachers have indicated that teachers believe that interim assessments can be useful in redesigning lessons, modifying instruction, and preparing students for standardized testing (Clune & White, 2008).

Some districts and school leaders are making efforts to develop interim assessment tools as a response to state sanctions to align curricula with government standards and state tests in order to regularly measure student progress. This is clearly an attempt to make changes in

instructional strategies. However, in order to achieve meaningful instructional changes principals and teachers need to act accordingly and be equipped with the necessary skills and knowledge, because interim assessments can only identify areas that need improvement, but the school agents are the ones who can use that information to attain targeted improvements (Blanc et al., 2010). Other work has provided support to this notion. Specifically, analysis of teacher interview data has suggested that teachers use interim assessments to gain information about their students' learning in mathematics, but they do not use interim assessments to assess students' conceptual understanding (Nabors-Olah, Lawrence, & Riggan, 2010). In a similar vein, teachers typically do not have the training to fully understand the technical aspects of the assessment tools, which may limit their ability to use the assessment tools most effectively. It is also conceivable that principals and teachers have difficulty distinguishing between formative and interim assessments and their potential uses (Li, Marion, Perie, & Gong, 2010).

## Empirical Evidence

A recent meta-analysis on the effects of data-driven assessments revealed that providing teachers with specific information about students' progress and areas of strengths and weaknesses improved student mathematics achievement (Gersten, et al., 2009). However, other recent primary studies on the effects of benchmark assessments on student achievement have produced mixed results. Through a large-scale cluster randomized experiment May and Robinson (2007) evaluated Ohio's Personalized Assessment Reporting System (PARS) for the Ohio Graduation Tests (OGT). The PARS program produced test score reports for teachers, administrators, students, and parents with the objective of monitoring student progress through

easily accessible data and changing teacher instruction to improve student learning. The authors compared 10th grade student achievement between 51 treatment and 49 control schools during the pilot year and found that the impact of the first year of PARS on student achievement was not significant. However, PARS effects were evident for students who retook the OGT assessments.

Another large-scale study employed a quasi-experimental design to examine whether middle schools participating in a Massachusetts pilot program utilizing quarterly benchmarks demonstrated greater mathematics gains than matched schools that did not participate in the program (Henderson, Petrosino, Guckenburg, & Hamilton, 2007). Pre-implementation test scores and other variables were used to match 44 comparison schools to the 22 treatment schools. The results revealed no statistically significant differences between treated and comparison schools. The authors noted that data limitations were a possible cause for the lack of significant findings.

In a different evaluation study, Quint, Sepanik, & Smith (2008) investigated the impact of Formative Assessment Student Thinking in Reading (FAST-R) on third and fourth grade classrooms in 21 Boston elementary schools during the 2005-2006 and 2006-2007 school years. The authors created a comparison group of 31 elementary schools and employed interrupted time series to determine achievement differences between treatment and control schools. The effects of FAST-R on the Massachusetts Comprehensive Assessment System (MCAS) test scores were generally positive, but not statistically significant. The results for reading achievement on the Stanford Achievement Test produced mixed results that were also statistically insignificant.

The impact of the Center for Data-Driven Reform in Education (CDDRE) on student achievement was also examined in a recent study (Carlson, Borman, & Robinson, 2011). The CDDRE intervention is a data-driven decision-making process that emphasizes instructional change based on benchmark assessment results among other things. The authors analyzed data

from a multistate district-level cluster randomized experiment to investigate the potential benefits of CDDRE. The sample included 509 schools across 56 districts in seven states. The results of the first year of the experiment indicated significant positive effects on mathematics scores, but the positive effects on reading scores did not reach statistical significance.

A follow-up study investigated the impact of CDDRE over a four-year period (Slavin, Cheung, Holmes, Madden, & Chamberlain, 2011). A total of 391 elementary schools and 217 middle schools were included in the analysis. Multilevel models were used to analyze the impact of CDDRE on grade 5 and 8 student achievement in mathematics and reading. In addition, the authors created matched comparison groups to examine the effects of treatment on the treated using ANCOVA. Both the experiment and the matched design analyses yielded effects that were generally small in both grades The results of the experiment analysis showed significant positive effects on both mathematics and reading and grade levels in the fourth year, whereas the results of the matched design analysis indicated strong positive results for reading only.

Most recently, the impact of the Measures of Academic Progress (MAP) benchmark assessment on reading achievement was examined (Cordray, Pion, Brandt, & Molefe, 2011). MAP is a product of Northwest Evaluation Association (NWEA) that is a widely used and commercially available system designed to incorporate benchmark assessment and training in differentiated instruction. Thirty-two elementary schools from five districts in the State of Illinois were randomly assigned to treatment and control conditions at grades 4 and 5 (with one grade per school being assigned to treatment and the other assigned to control). Over 170 teachers and nearly 4,000 students were included in the analyses. The study found that MAP was implemented with moderate fidelity, but MAP teachers were not more likely to differentiate instruction than their non-MAP colleagues. The researchers found no statistically significant

differences for either grade in reading achievement on the Illinois State Achievement Test or the MAP composite score.

## Methods

### Data

The randomized experiment was conducted in Indiana during the 2009-2010 academic year and included K-8 schools that had volunteered to be part of the intervention the Spring of 2009. The design was a two-level cluster randomized design (see Boruch, Weisburd, & Berk, 2010 for a discussion on these designs). Students were nested within schools, and schools were nested within treatment and control groups. Random assignment took place at the school level, that is, schools were randomly assigned to treatment and control conditions. Since the intervention was designed as a whole-school implementation, the conceptual match between design and practice was satisfactory.

Participating schools volunteered to participate in late Spring of 2009. From that list of schools, our initial objective was to assign 25 schools to a treatment and 25 schools to a control group. However, to facilitate participation we decided to use an unbalanced design instead with a larger number of schools in the treatment group (i.e., 30 treatment and 20 control schools). Specifically, our final sample included 59 schools, 35 of which were randomly assigned to the treatment condition, while the remaining 24 schools were randomly assigned to the control condition. Of the 35 treatment schools 31 participated in the experiment and of the 24 control schools 18 participated in the experiment for the whole year. The total number of participating schools altogether was 49. The schools in the treatment condition received *mCLASS* and/or

*Acuity*, while the control schools did not receive any treatment and should not have had any assessment program in place or have received a similar treatment the previous year. Because of random assignment of schools to conditions the results are likely to be causal. Overall, some 20,000 students participated in the study during the 2009-2010 school year.

The study was a large-scale randomized experiment where Indiana schools that volunteered to participate were randomly assigned to a treatment and a control condition. In 2006, the Indiana Legislature charged the Indiana State Board of Education (ISBE) and the IDOE to develop a long-term plan for a new assessment system that would be less expensive and less time consuming to measure individual student growth from year to year, and provide diagnostic information to teachers to use to improve ongoing instruction and ultimately student learning. This new system would be fully online. Among the proposals considered was one for new technology-enabled classroom-based diagnostic assessments for all K-8 grade students that would provide immediate feedback to the teacher and the student.

The plan required that the assessments be voluntary. In schools that chose to use them, IDOE would cover costs. The plan also tasked IDOE to ensure alignment of test content to Indiana standards and grade-level expectations. IDOE identified two commercial products through a standard public agency request for information, request for proposals, and negotiated bidding process. The first program was Wireless Generation's *mCLASS:Math* as the K–2 solution, and the second program was CTB/McGraw-Hill's *Acuity* product for Grades 3–8. From IDOE's perspective, this is a single intervention, a system of periodic diagnostic assessments that is *consistent*, because students throughout Indiana take the same assessments; *periodic*, because students are tested at the same three time points during the school year statewide; and *diagnostic*, because the assessments identify and report individual learning needs.

Indiana began the roll-out of the assessment program in summer 2008 by training teachers from more than 500 schools teaching some 220,000 K–8 grade students. These teachers and students used the diagnostic assessments during the 2008–09 school year. Additional schools volunteered to participate in the assessments each of the next several years. IDOE staff expects that essentially all elementary schools and students statewide will be active participants by 2013–14.

## The Intervention

With the *mCLASS,* the screening and diagnostic probes are conducted face-to-face, with students and teachers working together. The student performs language tasks while the teacher records characteristics of the work on a personal digital assistant (PDA). Teachers are guided through the assessments by the PDA and, through the PDA interface, they can immediately view results and compare them to prior performance. Detailed individual and group reports as well as ad hoc queries are available to the classroom teacher and other authorized personnel. In addition, at any point, teachers are able to monitor individual student progress in the classroom using short one-on-one, one-minute probes and then see those results linked to previous results graphically on the PDA screen.

*Acuity* provides online assessments in reading, mathematics, science, and social studies for grades 3-8. These assessments are 30- to 35-item multiple-choice online tests that can be completed within a class period, usually in group settings. They are closely aligned to Indiana standards and designed to be predictive of ISTEP$^{+}$ results. *Acuity* also permits teachers to construct practice or progress monitoring assessments from extensive banks of aligned items for

more frequent progress monitoring. Instructional resources—packaged student exercises to practice weak skills or explore others— are also available and may be assigned to specific students directly from *Acuity*'s diagnostic displays. Teacher access to most reports and queries is immediate.

## Analysis

We used student and school data for grades K-8. We conducted analysis both on the initial number of schools that were randomly assigned to conditions (Intention to Treat or ITT) and on the participating schools (Treatment on Treated or TOT). The outcome was mathematics scores of ISTEP+ and the main independent variable was treatment (coded as one for treatment schools and zero otherwise).

To capture the dependency in the data (i.e., students nested within schools) we used two-level models with students at the first level and schools at the second level. Schools were treated as random effects and the between-school variance of these effects indicated differences in mathematics achievement across schools. We conducted several analyses using data across all grades (i.e., k through 8), using k to 2 grade data and 3 to 8 grade data separately. We also conducted within grade analyses for each grade separately.

In each case we regressed mathematics scores on the treatment variable that was coded as a binary indicator, and other student and school covariates. The regression model for student *i* in school *j* is

$$y_{ij} = \beta_{00} + \beta_{10} Treatment_j + \mathbf{X}_{ij}\mathbf{B}_{20} + \mathbf{Z}_j\mathbf{B}_{30} + \mathbf{G}_{ij}\mathbf{B}_{40} + \upsilon_j + \varepsilon_{ij} \qquad (1)$$

where *y* is the outcome (i.e., mathematics scores), $\beta_{00}$ is the constant term, $\beta_{10}$ is the estimate of the treatment effect, *Treatment* is a binary indicator for the treatment, **X** is a row vector of student level predictors such as age, race, gender, SES, special education status, limited English proficiency status, $\mathbf{B}_{20}$ is a column vector of regression estimates of student predictors, **Z** is a row vector of school level predictors such as percent female, minority, disadvantaged, or limited English proficiency students in the schools, $\mathbf{B}_{30}$ is a column vector of regression estimates of school predictors, **G** represents grade fixed effects, $\mathbf{B}_{40}$ is a column vector of grade fixed effects estimates, $\upsilon$ is a school level residual, and $\varepsilon$ is a student level residual. The variance of $\upsilon$ captures the nesting of students within schools. We replicated the analysis described above for grades k to 2 and grades 3 to 8 separately to determine *mClass* and *Acuity* effects respectively. We also conducted within grade analyses for grades k through 6 that had sufficient data to estimate the treatment effect. In the within-grade analysis the grade dummies were omitted. For grades 4, 5, and 6 we also ran models that included prior student achievement as a covariate. This analysis was not possible to conduct in other grades given that prior scores were not available or the data were very sparse.

In order to put mathematics scores across grades into one common scale we used linear equating methods (e.g., creating z scores) (see Glick & White, 2003; Konstantopoulos, 2006). The standardization creates comparable indexes of achievement across grades under the assumption that the tests are linearly equitable (see Holland & Rubin, 1982; Kolen & Brennan, 1995). We standardized mathematics scores two different ways: across or within grades. In particular, in one case, we used the overall standard deviation of the outcome for all data across all grades to standardize mathematics scores. In another case, we standardized mathematics

scores within each grade using the grade-specific standard deviation of the outcome and then we pooled all scores across grades together for the analysis. The standard deviations within each grade were very similar and thus the within-grade standardization seemed reasonable. Both methods created data that produced very similar treatment estimates. Therefore, we report in the following section results that were produced from the within-grade standardization of the outcome. The grade dummies in this case do not have any impact on the treatment estimates.

# Results

## Analysis Across Grades

The preliminary analysis involved tests that checked whether random assignment of schools to conditions was successful for several observed variables at the school level. The random assignment indicated intention to treat. We used t-tests of independent samples to determine significant differences between the two conditions for several school-level observed variables such as proportion of female, minority, disadvantaged, special education, limited English proficiency students as well as prior school achievement. The results indicated that for these variables random assignment was successful (see Table 1). We replicated these analyses using data from participating schools and the results were by the large similar, that is, no significant differences between treatment and control conditions were found with respect to the variables of interest (see Table 2). Tables 1 and 2 report the mean differences between treatment and control schools, the standard error and theses mean differences, and the p-values of the t-tests used to examine the statistical significance of the mean differences. The mean differences

were typically smaller than their standard errors and never larger than two times their standard errors. As a result, all p-values were larger than .05, the standard level of statistical significance.

---

Insert Tables 1 & 2 Here

---

In the primary analysis the regression estimates were mean differences in standard deviation units between treatment and control groups. Positive estimates indicated a positive treatment effect. The results of the ITT analyses are reported in Table 3. The ITT analysis provided estimates for schools that were assigned randomly to treatment and control groups by design. All treatment effects were positive which suggests an overall positive treatment effect. The estimates obtained from the grade k-8 analysis were on average around one-tenth of a standard deviation. The estimates however were not statistical significant at the .05 level. The estimates from the grade k-2 analysis were also positive, insignificant and nearly one-half as large. Finally, the estimates from the grade 3 to 8 analysis were also positive, larger in magnitude and reached statistical significance at the .05 level when student and school covariates were included in the model. The significant treatment effect was slightly smaller than one-fifth of a standard deviation, which is not a trivial effect.

---

Insert Table 3 Here

---

The results of the TOT analyses are reported in Table 4. The TOT analysis provided estimates for schools that participated in the study and, therefore, selection bias is a possibility. The results were for the most part similar to those reported in Table 3. All treatment effects were positive as in Table 3, but the magnitude of the effects was much larger for the grade k-8 and 3-8 analyses. The treatment effects from the grade 3-8 analysis were positive and significant and nearly one-fourth of a standard deviation. In sum, the estimates presented in Tables 3 and 4 point to positive treatment effects that are not consistently significant. It appears that the treatment is more pronounced in grades 3 to 8, which would suggest an *Acuity* effect.

---

Insert Table 4 Here

---

Analysis Within Grades

We also conducted analyses within each grade (i.e., kindergarten to sixth grade) to determine grade-specific treatment effects. The results from the ITT analysis are summarized in Table 5. The overwhelming majority of the treatment effect estimates was positive. However, the estimates were not significant at the .05 level in most grades. In kindergarten, first, and second grades the estimates of the treatment effect were small and close to zero. Third and fourth grade estimates were larger but still insignificant. The treatment effect was positive and significant however in the fifth and sixth grades, especially when covariates were included in the model. The estimates were consistently larger than one-fourth of a standard deviation, which is a

considerable effect. The results from the TOT analysis are summarized in Table 6. Overall the results were similar to those reported in Table 5. However, the estimates were larger in grades 3 to 6 and the treatment effect was statistically significant in the third grade when covariates where included in the model, as well as in the fifth and sixth grades. The estimates in fifth and sixth grades were nearly one-third of a standard deviation. Generally, the within grade analysis yielded some interesting findings, which suggested consistent treatment effects in the fifth and sixth grades.

---

Insert Tables 5 & 6 Here

---

## Discussion

We examined the effects of a data-driven intervention on mathematics achievement using data from a cluster randomized experiment. We collected data of high quality that should facilitate causal inferences about the treatment effect. In addition, the findings have some generality since the analysis included schools from different parts of the State of Indiana. The findings overall suggested that the treatment effect was positive but not significant across all grades (i.e., K to 8). The treatment effect was smaller in lower grades (i.e., kindergarten to second grade), but larger in upper grades (i.e., third to eighth grade). Significant treatment effects were produced in the grade 3-8 analysis especially when covariates were included in the model. The effects of the TOT analysis were more pronounced and typically significant in the grade 3-8

analysis. These results are consistent in terms of the sign of the effect (i.e., positive), but inconsistent in terms of statistical significance. It seems for example, that *mClass* did not affect mathematics achievement much, whilst *Acuity* seems to have affected mathematics achievement positively and considerably in upper grades (e.g., fifth and sixth grade).

The within grade analysis revealed that in fifth grade mathematics the effect of the treatment was positive, significant, and not trivial. The treatment effect was consistently as large as one-fourth of a standard deviation and indicated an important annual gain in mathematics achievement (see Hill, Bloom, Black, & Lipsey, 2008). The estimates of the TOT analysis were even larger in magnitude and nearly one-third of a standard deviation. All other estimates were insignificant. Hence, it is unclear whether the intervention had any systematic effects on student achievement except for fifth and sixth grade mathematics.

The estimates produced from the ITT and the TOT analyses were overall similar qualitatively. The TOT estimates however, were larger and significant in more grades than the ITT estimates. Still, it is difficult to know whether selection bias is evident in the TOT estimates, since the two types of estimates are not that different. In addition, the tests we used to examine the degree to which random assignment was successful did not indicate any significant differences between treatment and control schools for either the ITT or the TOT analyses. The tests however, were based on schools means and, therefore, it is possible that differences could not be detected because of lower statistical power.

We also ran sensitivity analysis that omitted seventh and eighth grade data, but the estimates were very similar to the ones reported here. This was expected since the data from these grades were sparse. For grades 4, 5, and 6 we also ran models that included prior student

achievement as a covariate and the results were similar to those reported here. Other analyses were conducted including and excluding students who joined participating schools while the treatment was being implemented or students who left the experiment while in progress. The results of these analyses were very similar to those reported here indicating that student movement in and out of the participating schools did not affect the treatment estimates most likely. Overall it appears that the estimates are robust and show treatment effects in grades 3 to 8, especially in grades 5 and 6.

References

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E., Lawrence, N. R. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*, 205-225.

Boruch, R., Weisburd, D., & Berk, R. (2010). Place Randomized Trials. In A. Piquero & David Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp 481-502). New York: Springer.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's Benchmark Assessment System. *Peabody Journal of Education, 85*(2), 186-204.

Carlson, D., Borman, G.D., & Robinson, M. (2011). A multi-state district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33,* 378-398.

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence public schools.* WCER Working Paper No. 2008-10. Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Cordray, D., Pion, G., Brandt, C., and Molefe, A. (2011). *The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement.* Manuscript submitted for publication.

Dunn, K. E., & Malvenon, S. W. (2009). A critical review of research on formative assessment:

The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment Research and Evaluation, 14(7),* 1-11.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., Flojo, J. (2009). Mathematics Instruction for Students With Learning Disabilities: A Meta-Analysis of Instructional Components. *Review of Educational Research, 79,* 1202-1242.

Glick, J. E., & White, M. J. (2003). The academic trajectories of immigrant youths: Analysis within and across cohorts. *Demography, 40,* 759–783.

Henderson, S., Petrosino, A. Guckenburg, S., & Hamilton, S. (2007a). *Measuring how benchmark assessments affect student achievement* (Issues and 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.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom.* Thousand Oaks, CA: Corwin Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic Press.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer.

Konstantopoulos, S. (2006). Trends of School Effects on Student Achievement: Evidence from NLS:72, HSB: 82, and NELS:92. *Teachers College Record, 108*, 2550-2581.

Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education, 85*, 163-185.

May, H., & Robinson, M. A. (2007). *A randomized evaluation of Ohio's Personalized Assessment Reporting System (PARS).* Philadelphia: University of Pennsylvania Consortium for Policy Research in Education.

Nabors-Olah, L., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark data: How to analyze results. *Peabody Journal of Education, 85*, 226-245.

Quint, J., Sepanik, S., & Smith, J., (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Student Thinking in Reading (FAST-R) Program in Boston Elementary Schools.* New York: MDRC.

Perie, M., Marion, S., Cong, B. (2007). *A Framework for Considering Interim Assessments*. Dover, NH: National Center for the Improvement of Educational Assessment.

Sadler, D. R. (1989). Formative assessment and the design of instructional strategies. *Instructional Science, 18,* 119-144.

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

Slavin, R. E., Cheung, A., Holmes, G. C., Madden, N. A., Chamberlain, A. (2011). *Effects of a data-driven district reform model.* Baltimore, MD: Johns Hopkins University's Center for Research and Reform in Education.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-767.

Table 1. T-tests that check random assignment on observed variables of interest: All Schools

| Variable | $M_d$ | $SE_d$ | p-value |
|---|---|---|---|
| Grades K to 8: 58 Schools | | | |
| Proportion of Female Students | -0.006 | 0.010 | 0.563 |
| Proportion of Minority Students | 0.016 | 0.079 | 0.845 |
| Proportion of Disadvantaged Students | 0.042 | 0.054 | 0.430 |
| Proportion of Special Education Students | 0.011 | 0.014 | 0.440 |
| Proportion of Limited English Proficiency Students | 0.026 | 0.014 | 0.078 |
| Grades K to 2: 56 Schools | | | |
| Proportion of Female Students | -0.008 | 0.010 | 0.436 |
| Proportion of Minority Students | 0.013 | 0.082 | 0.874 |
| Proportion of Disadvantaged Students | 0.042 | 0.056 | 0.459 |
| Proportion of Special Education Students | 0.014 | 0.014 | 0.341 |
| Proportion of Limited English Proficiency Students | 0.026 | 0.015 | 0.077 |
| Grades 3 to 8: 57 Schools | | | |
| Proportion of Female Students | -0.006 | 0.010 | 0.580 |
| Proportion of Minority Students | 0.009 | 0.081 | 0.916 |
| Proportion of Disadvantaged Students | 0.048 | 0.055 | 0.384 |
| Proportion of Special Education Students | 0.011 | 0.014 | 0.440 |
| Proportion of Limited English Proficiency Students | 0.025 | 0.014 | 0.087 |
| Spring 2009 Math Scores | 6.290 | 6.542 | 0.342 |
| Spring 2009 ELA Scores | 1.340 | 5.410 | 0.806 |

Note: $M_d$ = Difference between treatment and control group school means; $SE_d$ = standard error of the mean difference.

Table 2. T-tests that check random assignment on observed variables of interest: Participating Schools

| Variable | $M_d$ | $S_d$ | *p* -value |
|---|---|---|---|
| Grades K to 8: 49 Schools | | | |
| Proportion of Female Students | -0.001 | 0.010 | 0.921 |
| Proportion of Minority Students | -0.014 | 0.094 | 0.886 |
| Proportion of Disadvantaged Students | 0.025 | 0.061 | 0.686 |
| Proportion of Special Education Students | 0.007 | 0.014 | 0.616 |
| Proportion of Limited English Proficiency Students | 0.022 | 0.016 | 0.175 |
| Grades K to 2: 44 Schools | | | |
| Proportion of Female Students | 0.014 | 0.013 | 0.329 |
| Proportion of Minority Students | -0.013 | 0.092 | 0.884 |
| Proportion of Disadvantaged Students | 0.025 | 0.065 | 0.697 |
| Proportion of Special Education Students | 0.013 | 0.017 | 0.470 |
| Proportion of Limited English Proficiency Students | 0.034 | 0.021 | 0.111 |
| Grades 3 to 8: 49 Schools | | | |
| Proportion of Female Students | -0.001 | 0.010 | 0.921 |
| Proportion of Minority Students | -0.014 | 0.094 | 0.886 |
| Proportion of Disadvantaged Students | 0.025 | 0.061 | 0.686 |
| Proportion of Special Education Students | 0.007 | 0.014 | 0.616 |
| Proportion of Limited English Proficiency Students | 0.022 | 0.016 | 0.175 |
| Spring 2009 Math Scores | 7.190 | 7.700 | 0.375 |
| Spring 2009 ELA Scores | 1.750 | 6.410 | 0.787 |

Note: $M_d$ = Difference between treatment and control group school means; $SE_d$ = standard error of the mean difference.

Table 3. Regression Estimates of Treatment Effects in Mathematics: Intention to Treat Analysis

| | Model I | | Model II | |
|---|---|---|---|---|
| Variable | Estimate | SE | Estimate | SE |
| Grades K to 8 | | | | |
| Treatment Effect | 0.074 | 0.083 | 0.128 | 0.070 |
| Number of Schools | 57 | | 56 | |
| Number of Students | 20428 | | 18800 | |
| Grades K to 2 | | | | |
| Treatment Effect | 0.045 | 0.119 | 0.084 | 0.099 |
| Number of Schools | 44 | | 44 | |
| Number of Students | 7644 | | 6948 | |
| Grades 3 to 8 | | | | |
| Treatment Effect | 0.141 | 0.088 | 0.186* | 0.077 |
| Number of Schools | 57 | | 56 | |
| Number of Students | 12784 | | 11852 | |

Note: Model I Includes Treatment; Model II Adds Student and School Characteristics
*$p \leqslant .05$

Table 4. Regression Estimates of Treatment Effects in Mathematics: Treatment on the Treated Analysis

| | Model I | | Model II | |
|---|---|---|---|---|
| Variable | Estimate | SE | Estimate | SE |
| Grades K to 8 | | | | |
| Treatment Effect | 0.148 | 0.088 | 0.189* | 0.070 |
| Number of Schools | 49 | | 49 | |
| Number of Students | 19102 | | 17931 | |
| Grades K to 2 | | | | |
| Treatment Effect | 0.045 | 0.119 | 0.084 | 0.099 |
| Number of Schools | 44 | | 44 | |
| Number of Students | 7644 | | 6948 | |
| Grades 3 to 8 | | | | |
| Treatment Effect | 0.229* | 0.092 | 0.257* | 0.076 |
| Number of Schools | 49 | | 49 | |
| Number of Students | 11458 | | 10983 | |

Note: Model I Includes Treatment; Model II Adds Student and School Characteristics

*$p \leqslant .05$

Table 5. Within Grade Regression Estimates of Treatment Effects in Mathematics: Intention to Treat Analysis

| | Model I | | Model II | |
|---|---|---|---|---|
| Variable | Estimate | SE | Estimate | SE |
| Grade k | | | | |
| Treatment Effect | -0.013 | 0.145 | 0.029 | 0.145 |
| Number of Schools | 38 | | 37 | |
| Number of Students | 2441 | | 2229 | |
| Grade 1 | | | | |
| Treatment Effect | 0.037 | 0.123 | 0.097 | 0.092 |
| Number of Schools | 38 | | 38 | |
| Number of Students | 2510 | | 2290 | |
| Grade 2 | | | | |
| Treatment Effect | 0.002 | 0.128 | 0.070 | 0.099 |
| Number of Schools | 44 | | 44 | |
| Number of Students | 2693 | | 2429 | |
| Grade 3 | | | | |
| Treatment Effect | 0.060 | 0.095 | 0.134 | 0.090 |
| Number of Schools | 57 | | 56 | |
| Number of Students | 3608 | | 3345 | |
| Grade 4 | | | | |
| Treatment Effect | 0.086 | 0.111 | 0.108 | 0.093 |
| Number of Schools | 57 | | 56 | |
| Number of Students | 3583 | | 3341 | |
| Grade 5 | | | | |
| Treatment Effect | 0.256* | 0.106 | 0.271* | 0.101 |
| Number of Schools | 56 | | 55 | |
| Number of Students | 3401 | | 3170 | |
| Grade 6 | | | | |
| Treatment Effect | 0.254 | 0.162 | 0.363* | 0.142 |
| Number of Schools | 26 | | 25 | |
| Number of Students | 1545 | | 1377 | |

Note: Model I Includes Treatment; Model II Adds Student and School Characteristics

*$p \leqslant .05$

Table 6. Within Grade Regression Estimates of Treatment Effects in Mathematics: Treatment on the Treated Analysis

| | Model I | | Model II | |
|---|---|---|---|---|
| Variable | Estimate | SE | Estimate | SE |
| Grade k | | | | |
| Treatment Effect | -0.013 | 0.145 | 0.029 | 0.145 |
| Number of Schools | 38 | | 37 | |
| Number of Students | 2441 | | 2229 | |
| Grade 1 | | | | |
| Treatment Effect | 0.037 | 0.123 | 0.097 | 0.092 |
| Number of Schools | 38 | | 38 | |
| Number of Students | 2510 | | 2290 | |
| Grade 2 | | | | |
| Treatment Effect | 0.002 | 0.128 | 0.070 | 0.099 |
| Number of Schools | 44 | | 44 | |
| Number of Students | 2693 | | 2429 | |
| Grade 3 | | | | |
| Treatment Effect | 0.121 | 0.098 | 0.190* | 0.09 |
| Number of Schools | 49 | | 49 | |
| Number of Students | 3224 | | 3079 | |
| Grade 4 | | | | |
| Treatment Effect | 0.161 | 0.121 | 0.160 | 0.099 |
| Number of Schools | 49 | | 49 | |
| Number of Students | 3200 | | 3074 | |
| Grade 5 | | | | |
| Treatment Effect | 0.338* | 0.115 | 0.334* | 0.107 |
| Number of Schools | 48 | | 48 | |
| Number of Students | 3018 | | 2903 | |
| Grade 6 | | | | |
| Treatment Effect | 0.298 | 0.172 | 0.391* | 0.144 |
| Number of Schools | 24 | | 24 | |
| Number of Students | 1369 | | 1308 | |

Note: Model I Includes Treatment; Model II Adds Student and School Characteristics

*p $\leqslant$ .05