Effects of a Data-Driven District Reform Model on State Assessments

Robert E. Slavin
Alan Cheung
Johns Hopkins University

GwenCarol Holmes
Alexandria City Public Schools

Nancy A. Madden
Success for All Foundation

Anne Chamberlain
Social Dynamics, LLC

Revision
July, 2011

This research was supported by a grant from the Institute for Education Science, U.S. Department of Education (No. R-305A040082). However, any opinions expressed are those of the authors and do not represent the positions or policies of the U.S. Department of Education.

Effects of a Data-Driven District-Level Reform Model

Abstract

A district-level reform model created by the Center for Data-Driven Reform in Education (CDDRE) provided consultation with district leaders on strategic use of data and selection of proven programs. 59 districts in seven states were randomly assigned to start CDDRE services either immediately or one year later. In addition, individual schools in each participating district were matched with control schools. Few important differences on state tests were found 1 and 2 years after CDDRE services began. The randomized design found positive effects on reading and math in fifth and eighth grade by Year 4. In the matched evaluation, positive, significant effects were seen on reading scores for fifth and eighth graders in Years 3 and 4. Effects were much larger for schools that selected proven programs than for those that did not.

For at least a quarter century, schools in the US have been in a constant state of reform. Commission reports, white papers, politicians, and the press periodically warn of dire consequences if America's schools are not substantially improved. In fact, on the 2009 National Assessment of Educational Progress (NCES, 2010) and on some international measures such as TIMSS (2007), PISA (2006), and PIRLS (2006), US schools have shown some gains in recent years, but the pace of change is slow. In particular, although the academic performance of middle class students is comparable to that of similar students in other countries, the most important problem in the US is the continuing low achievement of disadvantaged and minority students. For example, on the 2009 NAEP, 42% of White students scored proficient or better, while only 16% of African American, 17% of Hispanic, and 20% of American Indian students scored at this level. Among students who do not receive free lunches, 45% scored at proficient or better, while among those who receive free lunches, only 17% scored at proficient or better. Results in mathematics and at different grade levels showed similar gaps.

The continuing low performance of disadvantaged and minority students must be considered in light of substantial evidence showing positive effects of a wide range of educational innovations. Many interventions have been evaluated in rigorous experiments and found to improve student achievement, especially in reading and math, in comparison to traditional methods. Yet programs with strong evidence of effectiveness are rarely widely used, and those that are widely used rarely have much, if any, evidence of effectiveness. For example, there were five commercial reading texts that were emphasized in the federal Reading First program and were among the most widely used in the US during the period from 2000 to the present. The What Works Clearinghouse (2011a), in its beginning reading review, found supportive evidence for none of them. The same lack of evidence for these programs was reported in a review by Slavin, Lake, Chambers, Cheung, & Davis (2009). Reading programs that did have evidence of effectiveness from rigorous evaluations, such as various forms of tutoring, cooperative learning, and comprehensive school reform, are not used widely enough to have any meaningful impact on the national achievement gap. The same disconnect exists in math, where widely used textbook and CAI programs have little evidence of effectiveness (What Works Clearinghouse, 2011b, c; Slavin & Lake, 2008; Slavin, Lake, & Groff, 2009), while programs that do have extensive evidence of effectiveness are not widely used.

The limited application of proven programs is perhaps surprising in light of the extraordinary pressure schools have been under in recent years to improve student achievement. Under No Child Left Behind, schools have been subject to increasing sanctions leading up to closure or reconstitution if they do not meet standards on state accountability measures for a period of years. Because of the universal availability of data on student performance and the pressure to increase scores, it might be assumed that schools and districts would be intent on finding and adopting programs with strong evidence of effectiveness on the types of measures for which they are held accountable. Yet this is rarely the case.

Data-Driven Reform

The push to improve test scores has led to substantial interest in the use of data within schools and districts to drive decisions and motivate change. The focus of data-driven reform approaches is on obtaining timely, useful information, trying to understand the "root causes" behind the numbers, and designing interventions targeted to the specific areas most likely to be inhibiting success. The idea is both to focus resources and efforts most efficiently where they will make the biggest difference and to break the daunting task of turning around entire schools and districts into smaller, achievable tasks that can be accomplished in a reasonable time period, building a sense among front-line educators that they are capable of making a difference on enduring problems.

Data-driven reform involves collection, interpretation, and dissemination of data intended to inform and guide district and school reform efforts. Bernhardt (2003) identified four categories of data districts may analyze: student learning, demographics, school process, and teacher perceptions. These enable school leaders to identify specific problems faced by students and teachers, to break down the data to identify individual schools and demographic groups in need of particular help, and to suggest reasons for achievement gaps (Kennedy, 2003; Schmoker, 2003). Data-based decision making usually involves extensive professional development for school leaders to help them use data to set goals, prioritize resources, and make intervention plans (Conrad & Eller, 2003).

There is surprisingly little evidence on the effectiveness of data-driven reform strategies. That which does exist consists primarily of case studies of schools or districts that have made significant progress on state assessments. For example, the Council of the Great City Schools (2002) identified big-city districts that consistently "beat the odds" in raising student achievement, concluding that these districts were characterized by coherence, planfulness, and extensive use of data to inform district and school decisions. Case studies of other "positive outlier" districts and states have reached similar conclusions (CCSSO, 2002; Snipes, Doolittle, & Herlihy, 2002; Grissmer & Flanagan, 2001; Streifer, 2002; Symonds, 2003). However, such case studies only provide after-the-fact explanations of good results. We do not know, for example, whether schools and districts that did not make impressive gains may also have been trying to use the same data-driven strategies (see Herman et al., 2008).

Frequently, districts embarking on data-driven reform adopt benchmark assessments given several times a year to determine whether students are on track toward improvement on their state assessments. The idea is to find out early where problems may exist so that changes can be made before it is too late. There is evidence that more frequent assessment is more effective than annual assessment (e.g., Bangert-Drowns et al., 1991; Dempster, 1991; Schmoker, 1999), and in recent years, a few experimental and quasi-experimental evaluations of the use of such benchmark assessments have been reported. The findings are mixed. A Boston program, Formative Assessments of Student Thinking in Reading (FAST-R), provided teachers with data aligned with the Massachusetts MCAS reading assessments, which they gave to students every 3 to 12 weeks. Data coaches in each school helped teachers interpret and use the formative test data. A two-year evaluation of the program in 21 elementary schools found small, non-significant effects for third and fourth graders on MCAS and SAT-9 reading measures (Quint, Sepanik, & Smith, 2008). A one-year study of the use of benchmark assessments in 22 Massachusetts middle schools also showed no differences (Henderson, Petrosino, Guckenburg, & Hamilton, 2007). An analysis of first-year data from the present study by Carlson, Borman, & Robinson (in press) found significant but very small effects of the use of benchmark assessments on state mathematics assessments (ES = +0.06), but no significant effects on reading assessments (ES = +0.03).

A study by May & Robinson (2007) evaluated a benchmark assessment program used in high schools to prepare students for the Ohio Graduation Tests. The Personalized Assessment Reporting System (PARS) provided test reports for teachers, but also for students and their parents. Sixty districts were randomly assigned to use PARS or not to do so. There were no significant differences for 10th graders taking the Ohio Graduation Test for the first time, but there were positive effects for a subset of students who had initially failed the test. The second-chance students in PARS districts were more likely to take the test again and to score well on it.

Numerous studies have described "best practices" in the use of formative assessment data to help guide instruction. Examples include a study of "performance driven" school systems in California, Connecticut, and Texas by Datnow, Park & Wohlstetter (2007), studies of "data-informed districts" by Wayman, Jimerson, & Cho (2010), Wayman & Stringfield (2006), Wayman, Cho, & Shaw (2009), and Wayman, Cho, & Johnson (2007), and studies of evidence-based decision making in school district central offices (e.g., Bulkley, Christman, Goertz, & Lawrence, 2010; Honig, 2006; Honig & Coburn, 2008). All of these descriptive studies emphasize the need to make data important within systems, timely, and actionable, and to provide professional development and ongoing assistance to help teachers and administrators use the data intelligently, collaborative to decide on actions in response to findings, and follow through on solutions that flow from the data. Yet these studies do not establish a clear connection between effective use of data tools and student outcomes. Clearly, further research is needed to draw on the lessons of best practice and assess student outcomes over time.

Proven Programs

In all studies to date, the effects of implementing benchmark assessments with professional development to help educators interpret and respond appropriately to these assessments have been quite modest. It is perhaps too early to say that implementation of benchmark assessments is ineffective, but the expectation that providing periodic data on students' performance to teachers and administrators will greatly enhance achievement on accountability measures has not been convincingly demonstrated.

However, it may be that data-driven instruction will have more effect on achievement if assessment data are used to select proven programs which are known to be likely to improve outcomes in areas where weaknesses are observed. The theory of action implied in studies of benchmark assessments assumes that given frequent information on students' progress, teachers and administrators will adjust teaching strategies or school policies to respond to documented deficiencies. No one imagines that the assessment information in itself would lead to improved achievement; it is the educators' response to this information, the specific actions they take to remedy deficits, that are crucial. Yet these actions may or may not be implemented and may or may not be effective.

An alternative theory of action emphasizes the role of data-driven reform in encouraging educators to implement specific interventions known from research to be effective in improving student outcomes. By analogy, a physician's diagnostic procedures do not cure anything in themselves, but inform the selection from an armamentarium of proven treatments.

Cohen & Moffitt (2009, p. 226) make this point in their discussion of the accountability movement:

". . . the states' initiative (test-based accountability) is a version of standards-based reform, in which state policy makers and their allies seek to drive change in practice from the outside. Our analysis strongly suggests that this would be unlikely to work well, absent parallel efforts to build capacity from the inside."

Cohen & Moffitt (2009, pp. 226-227) go on to note that in order to build this capacity, the "most promising initial answers are comprehensive school reform designs . . .in which educational entrepreneurs carefully worked out designs for instruction."

The study reported in this article evaluated an approach like the one suggested by Cohen & Moffit (2009), in which district and school leaders were given data and assistance to identify key problems, as in all implementations of benchmark assessments, but were then helped and encouraged to select and implement proven programs likely to improve the identified outcomes. The longitudinal design allows for evaluation of the effects of adding quarterly benchmark assessments and then the effects of adding adoption of proven programs.

Center for Data-Driven Reform in Education

In 2004, the U.S. Department of Education funded a research center at Johns Hopkins University to create and evaluate a replicable approach to whole-district change based on the concepts of data-driven reform. The Center for Data-Driven Reform in Education (CDDRE) was intended to try to solve the problem of scale in educational reform by working with entire school districts. The idea was to help district and school leaders understand and supplement their data, as in the studies of benchmark assessments cited above. However, the emphasis of CDDRE was on going beyond formative assessments to help school leaders identify root causes underlying important problems, and then select and effectively implement programs directed toward solving those problems. The theory of action proposes that institutional change is facilitated by helping local decision makers not only understand their problems, but also making them aware of proven programs found to solve the problems identified in benchmark assessments or other data. A similar approach has been successfully used to improve outcomes such as reduced alcohol use and delinquent behaviors in a program called Communities That Care (Hawkins et al., 2009; Fagan, Brooke-Weiss, Cady, & Hawkins, 2009; Fagan, Hawkins, & Catalano, 2008). Another program, called PROSPER, which also helped communities select and implement proven programs, demonstrated positive effects on substance abuse (Spoth, Redmond, Shin, Greenberg, Clair, & Feinberg, 2007).

The CDDRE program offered to help schools adopt any program with strong evidence of effectiveness and partnered with several non-profit organizations that provide training and materials to support whole-school turnaround and have good evidence of effectiveness: Success for All (Slavin, Madden, Chambers, & Haxby, 2009), Direct Instruction (Adams & Engelmann, 1996), America's Choice (Supovitz, Poglinko, & Snyder, 2001), Modern Red Schoolhouse (2002), and Co-nect (Russell & Robinson, 2000). All of these were found to have "moderate" or better evidence of effectiveness by the Comprehensive School Reform Quality Center (CSRQ, 2006a, b).

Best Evidence Encyclopedia

In addition to information on proven whole-school reform models, CDDRE offered information to schools and districts on reading and math programs with strong evidence of effectiveness. Initially, it was expected that reviews of the evidence on such programs would soon be forthcoming from the What Works Clearinghouse, but the WWC reviews did not appear in time, so CDDRE created its own set of reviews, called the Best Evidence Encyclopedia (BEE; see www.bestevidence.org). These eventually covered elementary math (Slavin & Lake, 2008), secondary math (Slavin, Lake, & Groff, 2009), elementary reading (Slavin, Lake, Chambers, Cheung, & Davis, 2009), and secondary reading (Slavin, Cheung, Groff, & Lake, 2008).

The CDDRE Intervention

The services provided by CDDRE were designed to help district leaders understand and manage their own data, identify key areas of weakness and root causes for these deficits, recognize strengths and resources for reform, and then select and implement programs with strong evidence of effectiveness targeted to their identified areas of need. CDDRE consultants, all of whom had experience as superintendents, principals, or other leadership roles in education, provided approximately 30 days of on-site consultation to each district over a two-year period, depending on district size.

Data Review. CDDRE consultants cooperatively planned a series of meetings with district leaders and school teams (principal and key staff) to engage in a process of exploring all sources of data already collected by the district, including standardized test scores, attendance, disciplinary referrals, retentions, special education placements, and dropouts. CDDRE consultants and district leaders discussed the district's experiences with reform programs already in place, resources, state and federal mandates and constraints, and other factors relevant to the district's readiness for reform. Surveys of teachers collected information on their perceptions of school strengths and needs.

Benchmark Assessments. CDDRE created a set of state-specific benchmark assessments that assessed reading and mathematics achievement in grades 3-8 (in Pennsylvania, grades 3-11). These quarterly benchmark assessments, called 4Sight, were created from the same assessment blueprints as those used to construct the state assessments, and were written to mirror the state assessment’s content, coverage, difficulty, item types, proportions of open-ended items, and use of illustrations and other supports. The 4Sight benchmarks correlated with scores on the state test in the range of +0.80 to +0.85. 4Sight benchmarks were used 4-5 times per year to predict what students, student subgroups, classes, and schools would have scored on the state assessments. Special software enabled school leaders and teachers to examine the data by state standard, grade, class, student subgroup, and so on. The benchmark assessments provided district and school leaders with detailed, timely, actionable information on student achievement, giving them an opportunity to take action in time to affect yearly outcomes.

School Walk-Throughs. CDDRE consultants accompanied district leaders on visits to a cross-section of the district’s elementary, middle, and high schools. These structured walk-throughs provided insight for both the CDDRE consultants and the district administrators into the quality of instruction, classroom management, motivation, and organization of each school. They examined the implementation of various programs the schools were using, and focused on student engagement. In addition to informing CDDRE consultants, these walk-throughs were intended to help district leaders understand the real state of education in their own schools, to find out which of the many programs provided to their schools were actually in use, and to create a sense of urgency to take action.

Data-Based Solutions. Although many of the school leaders believed that the knowledge provided by benchmark assessments, data reviews, and walk-throughs were sufficient to cause reform to take place, the CDDRE model emphasized the idea that systematic reforms based on the data are essential if genuine progress is to be made. CDDRE consultants helped district and school leaders review potential solutions to the problems they identified. They emphasized programs and practices with strong evidence of effectiveness, those identified by the Best Evidence Encyclopedia or the What Works Clearinghouse. CDDRE consultants helped district and school leaders learn about research-proven solutions, and then advised them through a process of adopting and implementing them: obtaining teacher buy-in, ensuring high-quality professional development and follow-up, and doing formative assessments of program outcomes.

Focus of the Evaluation

The evaluation of the CDDRE process was intended to determine the value added to student achievement by the intervention throughout the districts involved. The intervention was delivered over a period of years, and had distinct components at different points in time that were expected to affect outcomes differentially. In the first year, all participating districts received extensive consulting on data-driven reform and almost all implemented benchmark assessments (unless they were already in use). Early-years outcomes therefore were exclusively evaluations of the data interpretation aspects of CDDRE. In later years, as many schools began to select and then implement proven programs, outcomes begin to reflect the effects of these programs. It was not the intention of the evaluation to examine impacts of particular programs, but rather to focus on the impact across the districts of the process that led to the selection and implementation of proven programs attuned to their needs. Since schools that implemented programs did so at different times in different subjects, the effects of the process would be expected to appear gradually over time.

Randomized Comparisons. The original design of the CDDRE intervention involved random assignment of pairs of similar districts within states to experimental or control conditions. A total of 59 districts in seven states (PA, AZ, MS, IN, OH, TN, AL) were recruited and randomly assigned in this way over a 3-year period. In order to facilitate recruitment, a delayed-treatment control group design was used, in which districts assigned to the control groups were eligible to receive the full treatment a year later.

In the first year, this delayed-treatment randomized design compared CDDRE schools to untreated schools, but after the first year, districts in the delayed treatment control group usually began to implement the CDDRE procedures. By the fourth and final year for the first cohort, experimental-control contrasts mostly compared fourth-year implementers to third-year implementers, since most control districts were using CDDRE procedures, a year behind their CDDRE counterparts. As a result, the intent to treat analysis did not show the full effects of the treatment as compared to ordinary practice, as there were few control groups that did not have some experience with the CDDRE process.

Matched Comparisons. Because of this delayed treatment problem, a second form of analysis was also used to compare CDDRE schools to matched schools outside of the experimental or control districts, but chosen to match the schools that implemented CDDRE. In this matched analysis, all of the selected districts that ever implemented CDDRE procedures were considered experimental groups, starting on whatever date they began to receive CDDRE services. Control schools that had never been involved with CDDRE were chosen from among all schools in non-participating districts in each state to match CDDRE schools in terms of prior state test scores, percent free lunch, ethnicity, urban/rural location, and school enrollment. The matched design allowed us to follow schools over time as they incorporated the CDDRE elements and to compare outcomes in CDDRE schools to those of schools as similar as possible to the experimental schools.

The randomized and matched analyses each had different advantages and limitations. The randomized analysis eliminated selection bias, in that all districts were assigned at random to immediate or delayed-treatment conditions. However, after the initial year, the comparison made in the randomized analysis was between districts and schools that received more or fewer years of intervention, rather than experimental versus business-as-usual control.

The matched analysis made a more policy-relevant comparison, between schools that received CDDRE services and those that did not. It focuses on the effect of the treatment on the treated, a typical follow-up analysis in a randomized design. It also nearly doubled the sample size, allowing for post-experimental comparisons of schools that did or did not implement reform models in reading or math. However, since CDDRE schools were in districts that chose to participate in the experiment, there may have been an element of self-selection bias in the comparison of these schools to those that did not have an opportunity to receive CDDRE services.

Because each of these designs had strengths and weaknesses, both are presented in this report as a triangulation of methodologies. That is, to the degree the alternative methods produce similar outcomes this strengthens causal claims. To the degree that they differ in outcomes, the differences can be examined for their substantive meanings.

For both forms of analysis, the research question was as follows:

- In comparison to control groups, what were the effects of CDDRE participation (controlling for pretests) on state tests of reading and mathematics at the elementary and middle school levels?

The matched design also permitted exploration of a second research question:

- Were effects of CDDRE participation (controlling for pretests) greater for schools that implemented proven reading and/or math programs than for schools that did not?

In addition to the overall impacts, both experimental designs enabled us to explore alternative theoretical models to explain outcomes. If positive achievement effects were seen in the early years, or if positive effects were found in schools that never adopted a proven program, this would support a conclusion that consultation and benchmark assessments have an independent effect on achievement. If positive effects were limited to the later years and to schools that did adopt one or more proven programs, this would support a conclusion that the program's effects are mediated primarily by adoption of proven programs.

Methods

Sample Selection

CDDRE districts were recruited by forming partnerships with state departments of education in the seven states listed earlier. The state departments then nominated districts with many low-achieving schools. The leadership of the nominated districts was approached by CDDRE staff and offered the opportunity to participate in the project, understanding that they would be randomly assigned to receive CDDRE services beginning either the following school year or a year later. The districts were recruited in three cohorts, beginning in spring of 2005 (n=20), 2006 (n=13), and 2007 (n=26). Within each district, district leaders could designate all schools or a subset of low-achieving schools to receive CDDRE services. Most of the 59 districts were in Pennsylvania (32), and there were 10 in Tennessee, 4 in Alabama, 4 in Arizona, 4 in Mississippi, 3 in Indiana, and 2 in Ohio. All were high-poverty Title I districts and schools, but they ranged from small rural districts to mid-sized urban ones.

Randomized Design

As districts were recruited for each cohort, they were matched with districts in the same state that were similar in demographic characteristics and prior achievement and then assigned at random (by coin flip) to the immediate intervention (experimental) or delayed treatment (control) groups. In four cases, no match was available, and districts were assigned individually by coin flip. The matching before random assignment was done just to reduce the possibility of inequalities within states and cohorts, and was not used in the design or analysis. There were a total of 391 elementary and 217 middle schools in the randomly assigned districts.

Table 1 shows demographic and pretest characteristics of all experimental and control schools in the randomized sample. As the table shows, the schools were very impoverished, with 64% of students qualifying for free- or reduced-price lunches. About 29% (Grade 5) and 31% (Grade 8) of the students were African American, 20% (at both levels) were Hispanic, and 48% (Grade 5) and 46% (Grade 8) were White. There were no significant differences between treatment and control schools on any of the baseline demographic characteristics. Enrollments, however, were significantly higher in the eighth grade treatment group (630 vs. 510, $p<.001$). In terms of pretest characteristics, no statistically significant differences were found between treatment and control schools for 5th grade reading and math and 8th grade reading. However, treatment schools scored significantly lower than control schools on 8th grade math ($p<.02$).

Since n’s of schools were lower in the cohorts that participated for 3 and 4 years, we also examined pretests for the final samples. These found no significant differences on pretest scores at any grade level or subject (tables available upon request).

==============

TABLE 1 HERE

==============

Matched Design

As noted earlier, the purpose of the matched analyses was to examine the impacts on schools of actually participating in CDDRE services. Because of the delayed treatment random assignment design, schools in most control districts in the randomized study began to receive CDDRE services just one year later than their corresponding districts in the experimental group. For example, control schools whose districts were randomly assigned in 2005, the first cohort, usually began implementation in 2006, at the same time as the experimental schools in the second cohort. For the matched analyses, schools were grouped in cohorts according to the year they began to receive services (2005, 2006, or 2007). This analysis examined effects of treatment on the treated (TOT); it excluded the one experimental district and six control districts that decided after random assignment not to take advantage of CDDRE services.

Control schools were identified from state records in districts not receiving CDDRE services. Matches were made based on prior achievement, percent free/reduced lunch, and student demographics. Characteristics of the CDDRE and control schools in the matched design are summarized in Table 2.

===================

TABLE 2 HERE

===================

As is apparent in Table 2, characteristics of the CDDRE and control schools in the matched design were very similar to those of the schools in the randomized analysis. Sixty-four percent of both fifth graders and eighth graders qualified for free- or reduced-price lunches. At both grade levels, 51% of students were White, 25% of fifth graders and 28% of eighth graders were African American, and 21% of fifth graders and 18% of eighth graders were Hispanic. Percent free lunch and pretest scores in reading and math were very similar in CDDRE and control schools. At both grade levels, percent White was the same in CDDRE and control schools, but CDDRE schools had somewhat higher percentages of Hispanic students and lower percentages of African American students. Pretests were equal in experimental and control schools except for fifth grade math, where control schools scored significantly higher than experimental schools ($p<.02$).

Measures

The measures for this study were the reading and math assessment scores for each state at the 5th and 8th grade levels: The Pennsylvania System of School Assessment (PSSA), the Tennessee Comprehensive Assessment Program (TCAP), the Alabama Reading and Mathematics Test (ARMT), the Arizona Instrument to Measure Standards (AIMS), the Mississippi Curriculum Test 2 (MCT-2), the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+), and the Ohio Achievement Test (OAT). Grades 5 and 8 were reported because these were the only grade levels tested at pretest in some states. Standard school-level scores on these measures were taken from raw scores provided by each state department of education for the year prior to CDDRE implementation through 2009. School-level means were used because individual-level scores were not available from some states.

Prior to analysis, all scores were transformed to z-scores within states and grade levels, to permit pooling across states and years. Note that this removes year-to-year variations likely to result from variations in test versions within states, which would affect control and experimental schools equally. A z-score of zero indicates that a school is scoring at the average for its set of matched experimental and control schools in a given year.

Analyses

All analyses used the school as the unit of analysis. A simple t-test was used to determine whether there was any significant difference in the achievement levels of treatment and control groups at pretest. The randomized evaluation used hierarchical linear modeling (HLM) with schools nested within districts, with pretests from the spring before implementation used as covariates. Comparisons were made each year using all schools (across cohorts) that were one to four years beyond random assignment. The matched evaluation used analyses of covariance (ANCOVAs) for each year using all schools that began implementation one to four years before posttest, controlling for pretests. Because schools joined the CDDRE project in successive waves or cohorts, the number of schools available for comparison at each post-test year diminishes over time. That is, while all schools included in the analysis have at least 2 years of post-test data, a smaller number (only the first cohort) accumulated 4 post-test years. Effect sizes were computed as the experimental-control difference in adjusted posttest scores divided by the unadjusted school-level standard deviation. School-level effect sizes were considered educationally important if they were at least +0.20 (equal to individual-level effect sizes in the range of +0.07 to +0.10, because school-level standards deviations are typically 2-3 times lower than individual ones).

Results

===================

TABLES 3-6 HERE

===================

Randomized Design

The outcomes for the randomized design are summarized in Tables 3-6. Table 3 shows that for fifth grade reading, treatment effects were not significant after one year (ES=+0.13, n.s.), two years (ES=+0.09, n.s.), or three years (ES=+0.24, p<.10), but were statistically significant after four years (ES=+0.50, p<.01). Eighth grade reading (Table 4), on the other hand, showed significant positive effects in Year 1 (ES=+0.26, p<.05) and Year 2 (ES=+0.23, p<.05), but not Year 3 (ES=+0.05, n.s.) or Year 4 (ES=+0.24, n.s.).

Mathematics effects for fifth grade (Table 5) were non-significant in Year 1 (ES=+0.25, n.s.), Year 2 (ES=+0.07, n.s.), and Year 3 (ES=+0.24, p<.10), but were significant in Year 4 (ES=+0.33, p<.05), mirroring the pattern seen in reading. Table 6 shows positive effects for eighth grade math in Year 1 (ES=+0.17, p<.01) but not Year 2 (ES=+0.08, n.s.) or Year 3 (ES=+0.01, n.s.). Math effects approached statistical significance in Year 4 (ES=+0.31, p<.10).

===================

FIGURES 1-6 HERE

===================

Matched Design

The findings for the matched design are summarized in Figures 1-6. Figure 1 shows overall outcomes for fifth grade reading. At the end of the first implementation year, CDDRE schools scored significantly lower than control schools (ES= -0.15, p<.05). CDDRE schools were nonsignificantly lower than controls in Year 2 (ES= -0.13, n.s.), but in Year 3, CDDRE schools scored significantly higher than controls (ES= +0.28, p<.01), and in Year 4, CDDRE schools scored non-significantly higher (ES = +0.20, n.s.).

Figure 2 breaks down the elementary reading findings according to schools that did or did not actually implement research-proven programs in reading. Overall, 30% of third-year elementary schools and 42% of fourth-year schools had adopted reading programs, most of which were cooperative learning programs provided by the Success for All Foundation (Slavin, Madden, Chambers, & Haxby, 2009). These programs all provided extensive professional development and supplementary materials to help students work in cooperative small groups to apply systematic phonics skills and use metacognitive strategies to comprehend text of increasing difficulty. Figure 2 shows that schools adopting reading programs scored substantially higher than control schools in the third year (ES = +0.49. p<.001) and in the fourth year (ES = +0.39, p<.02). Schools that did not implement a proven program scored lower than their controls in Years 1 and 2, and did not differ from controls in Years 3 and 4.

Figure 3 shows overall reading scores for eighth grade. At the end of the first implementation year, CDDRE schools did not differ from controls (ES= -0.12, n.s.), and were not significantly higher in the second year (ES= +0.16, p<.08). They scored significantly higher than controls in the third year (ES = +0.35, p<.02). In the fourth year, effect sizes remained positive but due to the smaller sample size, the differences were not significant (ES = +0.29, p<.10).

As in the elementary schools, middle schools that adopted proven reading programs had larger positive effects than those that did not. Overall, 28% of third-year and 33% of fourth year middle schools adopted proven programs. All schools that chose an intervention adopted the Success for All Foundation's Reading Edge program (Slavin, Chamberlain, Daniels, & Madden, 2009; Slavin, Daniels, & Madden, 2005). As shown in Figure 4, middle schools that adopted a reading program by the third year scored significantly higher than controls (ES = +0.37, p<.05). Fourth-year schools had a slightly higher effect size, but due to smaller n's, the differences were not significant (ES = +0.45, p<.07). Middle schools that did not adopt a reading program still had positive effects in Year 3 (ES = +0.33, p<.02), but not in Year 4 (ES = +0.13, n.s).

Fifth grade math scores, depicted in Figure 5, do not show any differences between CDDRE and control schools. Among elementary schools, only 8% of schools chose a proven math program by Years 3 or 4, so separate analyses for those that did adopt a program were not attempted. In eighth grade, overall effects were also not significant in Years 3 or 4 (see Figure 6), and only 8% of schools chose a math program by Years 3 or 4.

===================

FIGURES 7-8 HERE

===================

Outcomes by Cohort

In both the randomized and matched analyses, the most positive effects were seen for schools using CDDRE services for the longest time, especially those in the fourth year cohort in the randomized evaluation and the third and fourth year cohorts (in reading) in the matched analyses. An alternative explanation for such findings in a study combining multiple cohorts could be that the longest-implementing first cohort may have been superior to its control group all along. To test for this possibility, Figures 7 and 8 break down the matched findings for fifth and eighth grade reading, respectively, according to cohort. In both cases, the pattern of effects is similar to that for the combined analyses. Effects are small in Years 1 and 2 and then rise in Years 3 and 4 for the cohorts shown separately.

Discussion

The findings of the randomized and matched evaluations of CDDRE show similar but not identical patterns. In both analyses, effects were small in Years 1 and 2 at both grade levels and subjects. The randomized study found particularly strong outcomes for both subjects and grade levels in Year 4, while the matched evaluation showed particularly positive reading but not math effects in Years 3 and 4. The two analyses do not match perfectly, but they do both show positive effects of the CDDRE intervention in the later years. An earlier analysis of first-year data from all grades by Carlson, Borman, & Robinson (2010) similarly found quite small significant positive effects of CDDRE on math and no effects on reading.

The findings of the evaluation of CDDRE were somewhat different in the randomized and matched study and in reading and math, but there were some trends worthy of note. First, there were few important school-level (ES ≥ +0.20) first-year or second-year effects in either subject or grade level. Clearly, the provision of workshops and implementation of benchmarks was not sufficient to bring about significant changes in student performance. These findings are in accord with the outcomes of previous large-scale evaluations of benchmark assessment plans by Quint et al. (2008) and Henderson et al. (2007). They are also in accord with the program’s theory of action; first and second year interventions were analogous to taking a patient’s temperature, not providing a treatment.

By the third implementation year, school and district leaders were, in many locations, beginning to take action based on the data. Reading scores at both grade levels were significantly higher in CDDRE than in control schools in the matched comparisons. There were positive effects in both subjects and both grade levels by Year 4 in the randomized evaluation. However, the reading impacts depended substantially on whether or not schools actually adopted proven programs. In the third and fourth years of their CDDRE implementations, elementary and middle schools that implemented a proven program scored substantially higher than their control groups, while important positive reading effects were seen for schools that did not implement a proven program only in the third year cohort in middle schools. In mathematics, there were no significant overall effects in the matched comparisons, and few schools adopted math programs.

What the findings imply is that helping school leaders understand student data is not enough to produce gains in achievement. Schools must actually take action to change teaching and learning. The findings support a model of change in which initial consultation and implementation of benchmark assessments motivate school leaders to adopt proven programs, and it is these programs that lead to achievement gains, not the consultation or benchmarks in themselves.

An important finding of the CDDRE experiment was that it was possible to get many schools to adopt proven reading programs by engaging them in a process of self-examination, benchmark assessments, and exposure to information on proven alternatives. Very few schools had adopted any of the proven programs before their involvement with CDDRE, and by the fourth year, 42% of elementary schools and 33% of middle schools had implemented reading programs. However, the process was slow and uneven, as has been reported in similar programs in areas other than education (Fagan, Brooke-Weiss, Cady, & Hawkins, 2009). Fewer than half of all eligible schools eventually adopted reading programs, and only 8% adopted math programs. The same programs chosen by the CDDRE schools are typically adopted outside the CDDRE experiment in a period of months, especially if funding is available to help defray the costs (as was true when Obey-Porter comprehensive school reform funding was available in the late 1990's, for example; see Slavin, 2008). Something like the CDDRE process may be necessary for schools and districts otherwise unlikely to implement proven programs, but schools that are already aware of their needs may benefit from effective methods fairs or other less customized and faster methods of informing them about the effective alternatives available to them.

Where CDDRE appeared to make its largest differences, in reading at both grade levels among schools that implemented proven reading programs, the magnitude of the effects was surprisingly large, averaging effect sizes of +0.49 in Year 3 and +0.39 in Year 4 in fifth grade and +0.37 in Year 3 and +0.45 in Year 4 in eighth grade. Fourth-year effects in the randomized analyses were similarly large for fifth graders in both subjects. These effect sizes cannot be compared to individual-level effect sizes, because standard deviations (the denominator of the effect size formula) are 2-3 times larger among students than among schools. However, the ability to make this much difference on such a large scale is important. If outcomes of similar magnitude are seen in replications, these findings may point to a relatively inexpensive, readily scalable strategy for making a different in the performance of high-poverty schools.

References

Adams, G.L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR.* Seattle, WA: Educational Achievement Systems.

Bangert-Drowns, R.L., Kulik, C.C., Kulik, J.A, & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61* (2), 213-238.

Bernhardt, V.L. (2003). No schools left behind. *Educational Leadership, 60* (5), 26-30.

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. (in press). 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.*

Cohen, D. K., & Moffitt, S. L. (2009). *The ordeal of equality: Did federal regulation fix the schools?* Cambridge, MA: Harvard University Press.

Comprehensive School Reform Quality Center (2006a). *CSRQ Center report on elementary school comprehensive school reform models*. Washington, DC: American Institutes for Research.

Comprehensive School Reform Quality Center (2006b). *CSRQ Center report on secondary school comprehensive school reform models*. Washington, DC: American Institutes for Research.

Conrad, W.H., & Eller, B. (2003, April). *District data-informed decision making*. Paper presented at the annual meetings of the American Educational Research Association, Chicago.

Council of the Great City Schools (2002). *Beating the odds II.* Washington, DC: Authors.

Datnow, A., Park, V., Wohlstetter, P. (2007). Achieving with Data. *How high-performing school systems use data to improve instruction for elementary students.* Los Angeles, California: Center on Educational Governance, University of Southern California.

Dempster, F. N. (1991). Synthesis of research on reviews and tests. *Educational Leadership, 72*(8), 71–76.

Fagan, A. A., Brooke-Weiss, B. L., Cady, R., Hawkins, J. D. (2009). If at first you don't succeed…keep trying: Strategies to enhance coalition/school partnerships to implement school-based prevention programming. *Australian and New Zealand Journal of Criminology, 42(3),* 387-405.

Fagan, A. A., Hawkins, J. D., & Catalano, R. F. (2008). Using community epidemiologic data to improve social settings: The Communities That Care Prevention System. In M. Shinn & H. Yoshikawa (Eds.), *Toward Positive Youth Development: Transforming Schools and Community Programs* (pp. 292-312). New York: Oxford University Press.

Grissmer, D. & Flanagan, A. (2001). Searching for indirect evidence for the effects of statewide reforms. In *Brookings papers on education policy*, ed. D. Ravitch. Washington, DC: Brookings Institution.

Hawkins, J. D., Oesterle, S. Brown, E. C., Arthur, M. W., Abbott, R. D., Fagan, A. A., & Catalano, R. F. (2009). Results of a type 2 translational research trial to prevent adolescent drug use and delinquency: A test of Communities That Care. *Archives of Pediatrics and Adolescent Medicine*, 163 (9), 789-798.

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.

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., Redding, S., & Darwin, M. (2008). *Turning around chronically low-performing schools: A practice guide* (NCEE #2008-4020). Washington DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/practiceguides.

Honig, M.I. (2006). Street-level bureaucracy revisited: Frontline district central office administrators as boundary spanners in education policy implementation. *Educational Evaluation and Policy Analysis 28*(4), 357-383.

Honig, M.I., & Coburn, C.E. (2008). Evidence-based decision-making in school district central offices: Toward a research agenda. *Educational Policy, 22*(4), 578-608.

International Association for the Evaluation of Educational Achievement (IEA) (2003). *Trends in international mathematics and science study (TIMSS), 1995, 1999, and 2003.* Boston, MA: Author.

Kennedy, E. (2003). *Raising test scores for all students: An administrator's guide to improving standardized test performance*. Thousand Oaks, CA: Corwin Press.

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.

Modern Red Schoolhouse (2002). *School improvement: Research results for Modern Red Schoolhouse*. Nashville, TN: Author.

National Association of Educational Progress. (2009). *The nation's report card: Reading 2009.* Washington, DC: US Department of Education.

National Center for Education Statistics (2010). *The nation's report card.* Washington, DC: Author.

Quint, J., Sepanik, S., Smith, J., & MDRC (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Students Thinking in Reading (FAST-R) Program in Boston Elementary Schools*. MDRC (ERIC Document Reproduction Service No. ED503919) Retrieved January 14, 2011, from ERIC database.

Russell, M., & Robinson, R. (2000). *Co-nect retrospective outcomes study*. Boston, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Schmoker, M. (1999). *Results: The key to continuous school improvement (2nd ed.).* Alexandria, VA: ASCD.

Schmoker, M. (2003). First things first: Demystifying data analysis. *Educational Leadership, 60 (5),* 22-25.

Slavin, R. E. (2008). Comprehensive school reform. In C. Ames, D. Berliner, J. Brophy, L. Corno, & M. McCaslin (Eds.), *21st century education: A reference handbook*. Thousand Oaks, CA: Sage.

Slavin, R. E., Chamberlain, A., Daniels, C., & Madden, N. A. (2009). The Reading Edge: A randomized evaluation of a middle school cooperative reading programme. *Effective Education, 1*(1), 13-26.

Slavin, R.E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best evidence synthesis. *Reading Research Quarterly*, 43(3), 290-322.

Slavin, R.E., Daniels, C., & Madden, N.A. (2005). The Success for All Middle School: Adding content to middle grades reform. *Middle School Journal*, 36 (5), 4-8.

Slavin, R.E., & Lake, C. (2008). Effective programs in elementary math: A best evidence synthesis. *Review of Educational Research*, 78(3), 427-515.

Slavin, R.E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best evidence synthesis. *Review of Educational Research*, 79 (2), 839-911.

Slavin, R.E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best evidence synthesis. *Review of Educational Research, 79* (4), 1391-1465.

Slavin, R.E., Madden, N.A., Chambers, B. & Haxby, B. (2009). *Two million children: Success for All*. Thousand Oaks, CA: Corwin.

Snipes, J., Doolittle, F., & Herlihy, C. (2002). *Foundations for success: Case studies of how urban school systems improve student achievement*. Washington, DC: Council of the Great City Schools.

Spoth, R. L., Redmond, C., Shin, C., Greenberg, M. Clair, S., & Feinberg, M. (2007). Substance abuse outcomes at eighteen months past baseline from the PROSPER community-university partnership trial. *American Journal of Preventive Medicine,32* (5), 395-402.

Streifer, P. (2002). *Using data to make better educational decisions.* Lanham, MD: Scarecrow Press.

Supovitz, J.A., Poglinco, S.M., & Snyder, B.A. (2001). *Moving mountains: successes and challenges of the America's Choice comprehensive school reform design*. Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

What Works Clearinghouse (2011a). *Beginning reading topic report*. At http://ies.ed.gov/ncee/wwc/reports/beginning_reading/topic/

What Works Clearinghouse (2011b). *Elementary mathematics topic report*. At http://ies.ed.gov/ncee/wwc/reports/elementary_math/topic/

What Works Clearinghouse (2011c). *Middle school mathematics topic report*. At http://ies.ed.gov/ncee/wwc/reports/middle_math/topic/

Wayman, J. C. & Stringfield, S. (2006). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education 112*(4), 549-571.

Wayman, J. C., Cho, V., & Johnston, M. T. (2007). *The data-informed district: A district-wide evaluation of data use in the Natrona County School District.* Austin: The University of Texas.

Wayman, J. C., Cho, V., & Shaw, S. M. (2009). *First-year results from an efficacy study of the Acuity data system.* Austin: The University of Texas.

Wayman, J. C., Jimerson, J. B., & Cho, V. (2010, October). *District policies for the effective use of student data.* Paper presented at the 2010 Annual Convention of the University Council for Educational Administration, New Orleans LA.

l

Table 1
Comparison of Baseline Demographic Characteristics: Randomized Design

5th Grade

Variable	Condition	N	Mean	SD	p-value
Reading Pretest (z)	Treatment	165	-0.07	1.04	0.25
	Control	190	0.06	0.96	
Math Pretest (z)	Treatment	165	0.01	1.06	0.97
	Control	189	-0.01	0.93	
Free Lunch (%)	Treatment	187	63	25.64	0.41
	Control	204	65	24.54	
Enrollment	Treatment	187	470	240.78	0.25
	Control	204	445	179.91	
Female (%)	Treatment	187	48	2.98	0.88
	Control	204	48	4.48	
African American (%)	Treatment	187	29	31.82	0.83
	Control	204	29	31.00	
Hispanic (%)	Treatment	187	18	25.13	0.38
	Control	204	21	31.29	
White (%)	Treatment	187	50	35.51	0.42
	Control	204	47	37.46	

8th Grade

Variable	Condition	N	Mean	SD	p-value
Reading Pretest (z)	Treatment	92	-0.10	0.98	0.18
	Control	120	0.07	0.99	
Math Pretest (z)	Treatment	80	-0.25	1.00	0.02
	Control	120	0.17	0.92	

Free Lunch (%)	Treatment	95	63	26.02	0.54
	Control	122	65	25.23	
Enrollment	Treatment	95	630	319.19	0.00
	Control	122	510	222.86	
Female (%)	Treatment	95	49	2.61	0.49
	Control	122	49	4.26	
African American (%)	Treatment	95	32	34.39	0.59
	Control	122	30	31.35	
Hispanic (%)	Treatment	95	18	25.49	0.31
	Control	122	22	33.03	
White (%)	Treatment	95	48	38.01	0.55
	Control	122	45	37.54	

Table 2
Comparison of Baseline Demographic Characteristics: Matched Design

5th Grade

Variable	Condition	N	Mean	SD	p-value
Reading Pretest (z)	Treatment	272	-0.06	0.99	0.12
	Control	284	0.06	1.00	
Math Pretest (z)	Treatment	263	-0.10	1.07	0.02
	Control	295	0.09	0.92	
Free Lunch (%)	Treatment	272	62	23.37	0.28
	Control	284	65	22.56	
Enrollment	Treatment	272	466	216	0.30
	Control	284	485	223	
Female (%)	Treatment	272	48	3.26	0.24
	Control	284	48	2.70	
African American (%)	Treatment	272	23	28.72	0.11
	Control	284	27	33.22	
Hispanic (%)	Treatment	272	23	32.20	0.03
	Control	284	18	28.16	
White (%)	Treatment	272	51	38.34	0.87
	Control	284	50	37.95	

8th Grade

Variable	Condition	N	Mean	SD	p-value
Reading Pretest (z)	Treatment	152	-0.02	0.95	0.74
	Control	149	0.02	1.00	
Math Pretest	Treatment	158	-0.03	1.06	0.59
	Control	167	0.03	0.92	

Free Lunch (%)	Treatment	152	64	22.83	0.70
	Control	149	63	24.55	
Enrollment (%)	Treatment	152	601	287	0.27
	Control	149	562	310	
Female (%)	Treatment	152	48	3.11	0.12
	Control	149	48	5.01	
African American (%)	Treatment	152	23	28.86	0.00
	Control	149	33	36.84	
Hispanic (%)	Treatment	152	22	31.18	0.02
	Control	149	14	30.26	
White (%)	Treatment	152	52	39.18	0.66
	Control	149	50	38.83	

Table 3

Multilevel Models Predicting District-Level 5th Grade Reading Outcomes

N=59 (389 schools)

Level 1 model: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2 model: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Mean Pretest})_j + \gamma_{02}(\text{CDDRE})_j + u_{0j}$

Multilevel Models Predicting Reading Outcomes

	5th Grade Reading Outcomes											
	Year 1 (N=59)			Year 2 (N=59)			Year 3 (N=33)			Year 4 (N=19)		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	-0.04	0.07	-0.57	-0.03	0.04	-0.71	-0.06	0.07	-0.91	0.04	0.07	0.66
Mean Pretest	0.67**	0.12	5.40	0.80**	0.10	8.32	0.75**	0.11	6.68	0.80**	0.13	6.23
Treatment	**+0.13**	0.14	-1.99	**+0.09**	0.09	0.38	**+0.24**[a]	0.12	2.00	**+0.50****	0.10	5.02
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
District mean achievement	0.11	114.9**	56	0.01	58.92	56	0.04	40.97[a]	56	0.01	14.17	16
Within-district variation	0.71			0.72			0.68			0.72		

Note: a p <0.10 * $p < .05$; ** $p < .01$.

Table 4

Multilevel Models Predicting District-Level 8th Grade Reading Outcomes

N=59 (217 schools)

Level 1 model: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2 model: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Mean Pretest})_j + \gamma_{02}(\text{CDDRE})_j + u_{0j}$

Multilevel Models Predicting Reading Outcomes

	8th Grade Reading Outcomes											
	Year 1 (N=59)			Year 2 (N=59)			Year 3 (N=33)			Year 4 (N=18)		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	-0.02	0.06	-0.43	-0.05	0.05	-0.96	-0.15	0.06	-2.50	-0.15	0.11	-1.33
Mean Pretest	0.76**	0.09	8.72	0.64**	0.08	7.88	0.66**	0.08	8.43	0.79**	0.11	7.32
Treatment	**+0.26***	0.11	2.34	**+0.23***	0.10	2.32	**+0.05**	0.14	0.37	**+0.24**	0.21	1.16
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
District mean achievement	0.03	45.19	56	0.01	42.52	56	0.01	22.37	30	0.03	16.24	15
Within-district variation	0.68			0.76			0.68			0.63		

Note: a<0.10 * $p < .05$; ** $p < .01$.

Table 5

Multilevel Models Predicting District-Level 5th Grade Math Outcomes

N=56 (374 schools)

Level 1 model: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2 model: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Mean Pretest})_j + \gamma_{02}(\text{CDDRE})_j + u_{0j}$

Multilevel Models Predicting Math Outcomes

	5th Grade Math Outcomes											
	Year 1 (N=56)			Year 2 (N=56)			Year 3 (N=31)			Year 4 (N=18)		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	-0.02	0.08	-0.29	-0.01	0.05	-0.23	-0.05	0.07	-0.61	0.07	0.11	0.68
Mean Pretest	0.64**	0.14	4.71	0.71**	0.12	5.99	0.74**	0.13	5.43	0.58**	0.15	3.79
Treatment	**+0.25**	0.16	1.55	**+0.07**	0.10	0.70	**+0.24**^{a}	0.10	1.82	**+0.33**^{a}	0.18	1.78
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
District mean achievement	0.22	190.9**	53	0.04	69.88^{a}	53	0.06	45.48*	28	0.08	25.38*	15
Within-district variation	0.62			0.74			0.66			0.76		

Note: a p<0.10 * *p* < .05; ** *p* < .01.

Table 6

Multilevel Models Predicting District-Level 8th Grade Math Outcomes

N=59 (217 schools)

Level 1 model: $Y_{ij} = \beta_{0j} + r_{ij}$

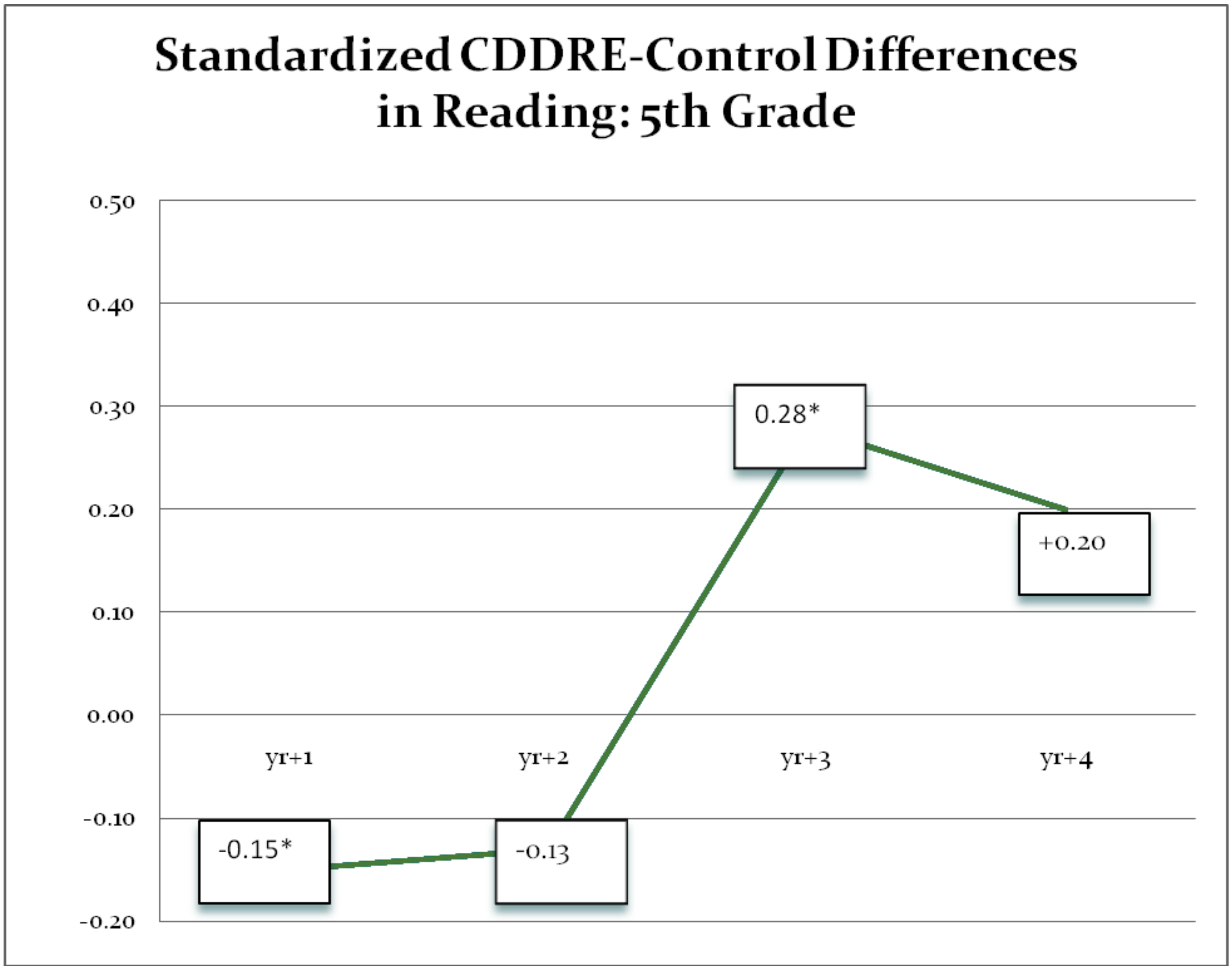
Level 2 model: $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Mean Pretest})_j + \gamma_{02}(\text{CDDRE})_j + u_{0j}$

Multilevel Models Predicting Math Outcomes

	8th Grade Math Outcomes											
	Year 1 (N=59)			Year 2 (N=59)			Year 3 (N=34)			Year 4 (N=18)		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	0.03	0.04	0.75	0.03	0.04	0.64	-0.03	0.07	-0.48	-0.05	0.07	-0.70
Mean Pretest	0.79**	0.06	13.89	0.76**	0.06	13.20	0.75**	0.08	8.78	0.81**	0.15	5.57
Treatment	**+0.17****	0.08	2.09	**+0.08**	0.08	1.11	**+0.01**	0.14	0.08	**+0.31**[a]	0.15	2.01
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
District mean achievement	0.01	36.97	56	0.01	33.83	56	0.01	29.38	31	0.02	13.36	15
Within-district variation	0.64			0.64			0.61			0.77		

Note: a p<0.10 * *p* < .05; ** *p* < .01.

Figure 1

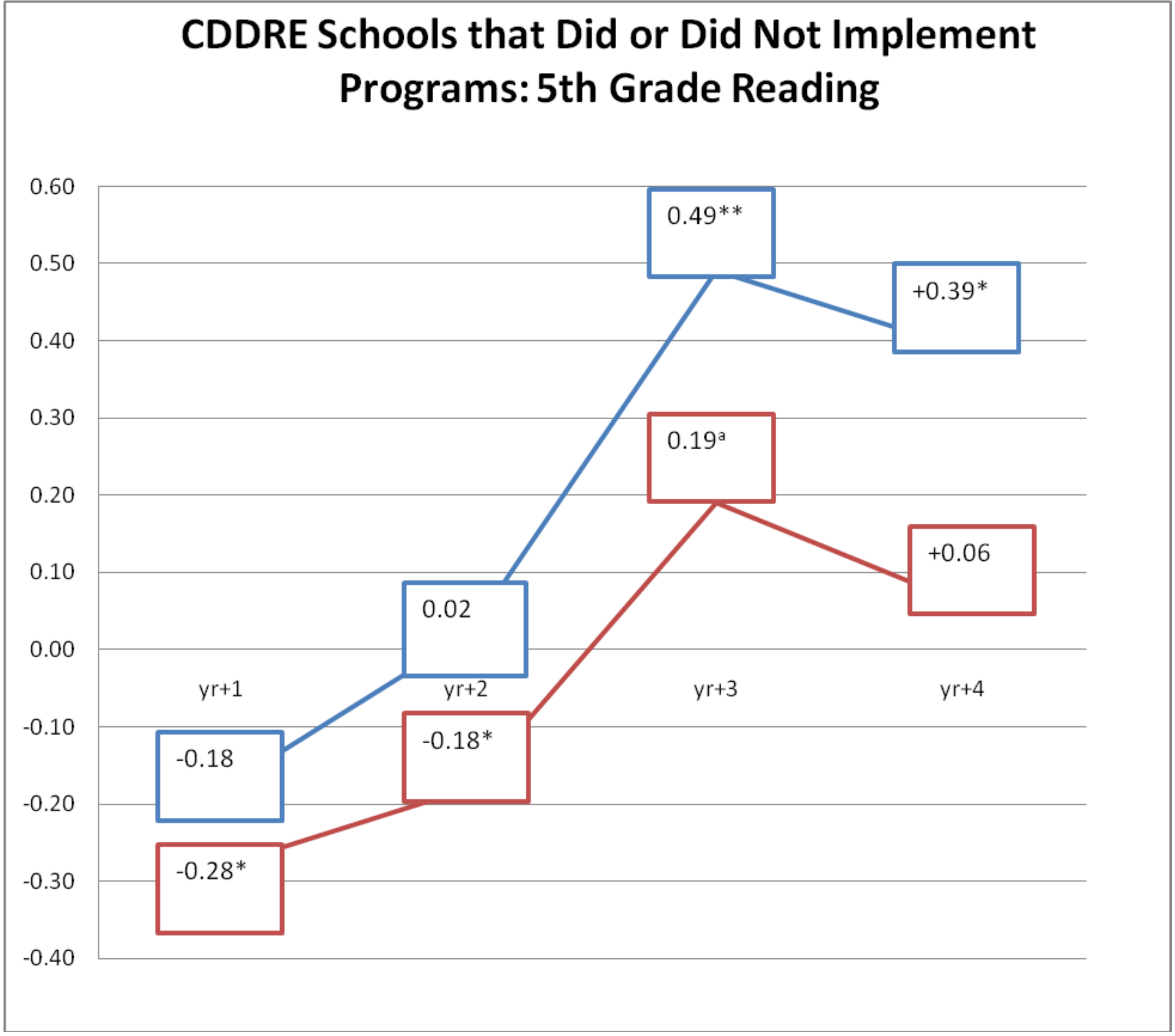


Adjusted posttest scores

	Baseline	yr+1	yr+2	yr+3	yr+4
Treatment	-0.06 (0.99)	-0.08(1.04)	-0.07 (1.05)	0.14 (0.96)	0.09 (0.96)
Control	+0.06 (1.00)	0.06 (0.93)	0.05 (0.91)	-0.14 (1.01)	-0.11 (1.00)
ES	**-0.12**	**-0.15**	**-0.13**	**0.28**	**0.20**
p-value	0.12	0.05	0.11	0.01	0.13
N (schools)	T=272, C=284	T=272, C=284	T=216, C=221	T=135, C=142	T=69, C=71

a p< .10
* p< .05
** p< .01

Figure 2

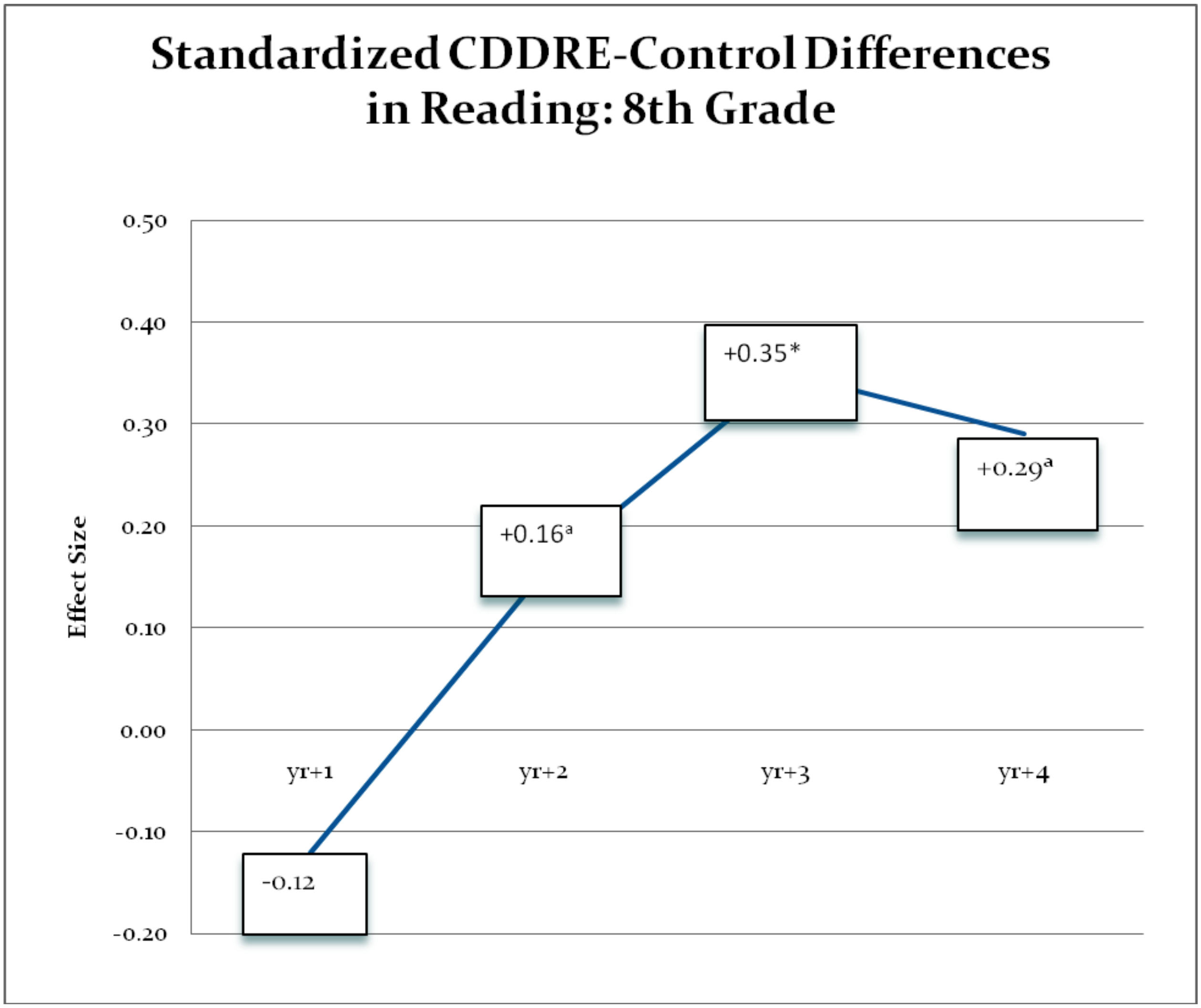


Effect Sizes

	Baseline	yr+1	yr+2	yr+3	yr+4
Treatment	**-0.09**	**-0.18**	**0.02**	**0.49**	**0.39**
p-value	0.20	0.20	0.84	0.00	0.02
N (schools)	T=39, C=142	T=39, C=142	T=39, C=142	T=40, C=142	T=29, C=71
Control	**-0.12**	**-0.28**	**-0.18**	**0.19**	**0.06**
p-value	0.38	0.01	0.05	0.07	0.02
N (schools)	T=95, C=142	T=95, C=142	T=95, C=142	T=95, C=142	T=40, C=71

a p< .10
* p< .05
** p< .01

Figure 3

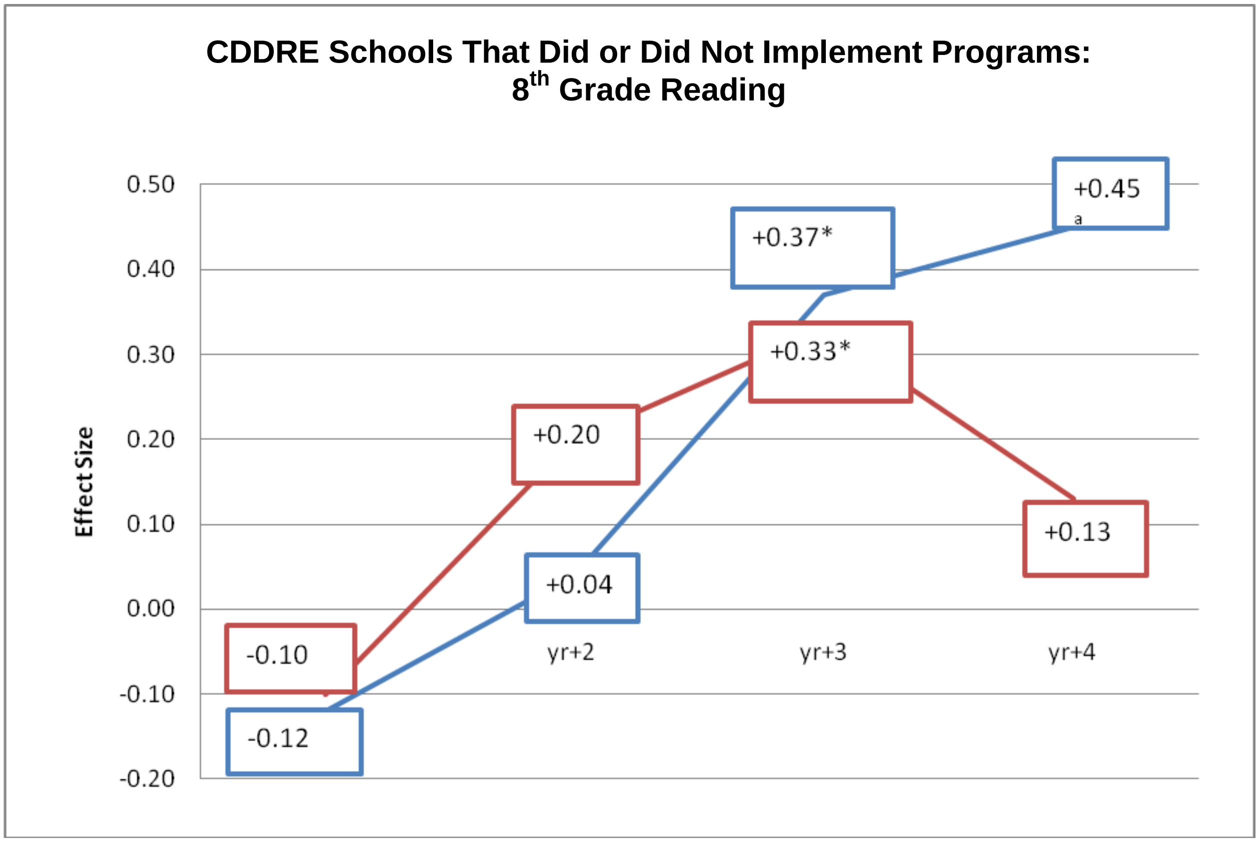


Adjusted posttest scores

	Baseline	**yr+1**	**yr+2**	**yr+3**	**yr+4**
Treatment	-0.02 (0.95)	-0.06 (1.04)	0.08 (0.88)	0.15 (1.00)	0.14 (1.01)
Control	0.02 (1.00)	0.06 (0.93)	-0.08(1.07)	-0.18 (0.93)	-0.13(0.93)
ES	**-0.04**	**-0.12**	**0.16**	**0.35**	**0.29**
p-value	0.74	0.12	0.08	0.02	0.10
N (schools)	T=152, C=149	T=152, C=149	T=101, C=99	T=58, C=50	T=33, C=34

a p< .10
* p< .05
** p< .01

Figure 4



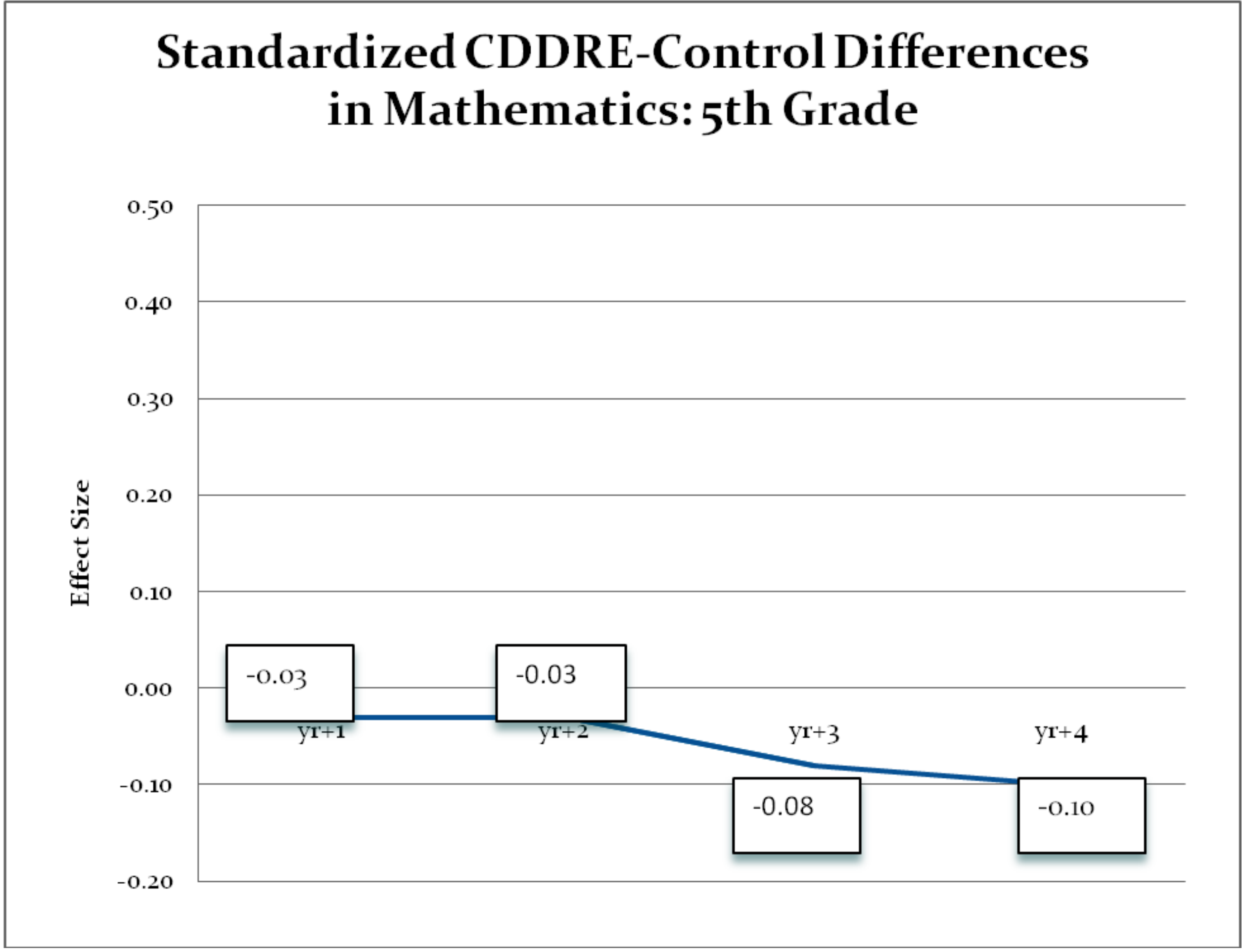
CDDRE Schools That Did or Did Not Implement Programs: 8th Grade Reading

Effect Sizes

	Baseline	yr+1	yr+2	yr+3	yr+4
Treatment	**-0.10**	**-0.12**	**0.04**	**0.37**	**0.45**
p-value	0.43	0.48	0.84	0.05	0.07
N (schools)	T=15 , C=50	T=15 , C=50	T=15 , C=50	T=16 , C=50	T=11, C=34
Control	**0.13**	**-0.10**	**0.20**	**0.33**	**0.13**
p-value	0.55	0.43	0.14	0.02	0.52
N (schools)	T=42, C=50	T=42, C=50	T=42, C=50	T=42, C=50	T=22, C=34

a p< .10
* p< .05
** p< .01

Figure 5

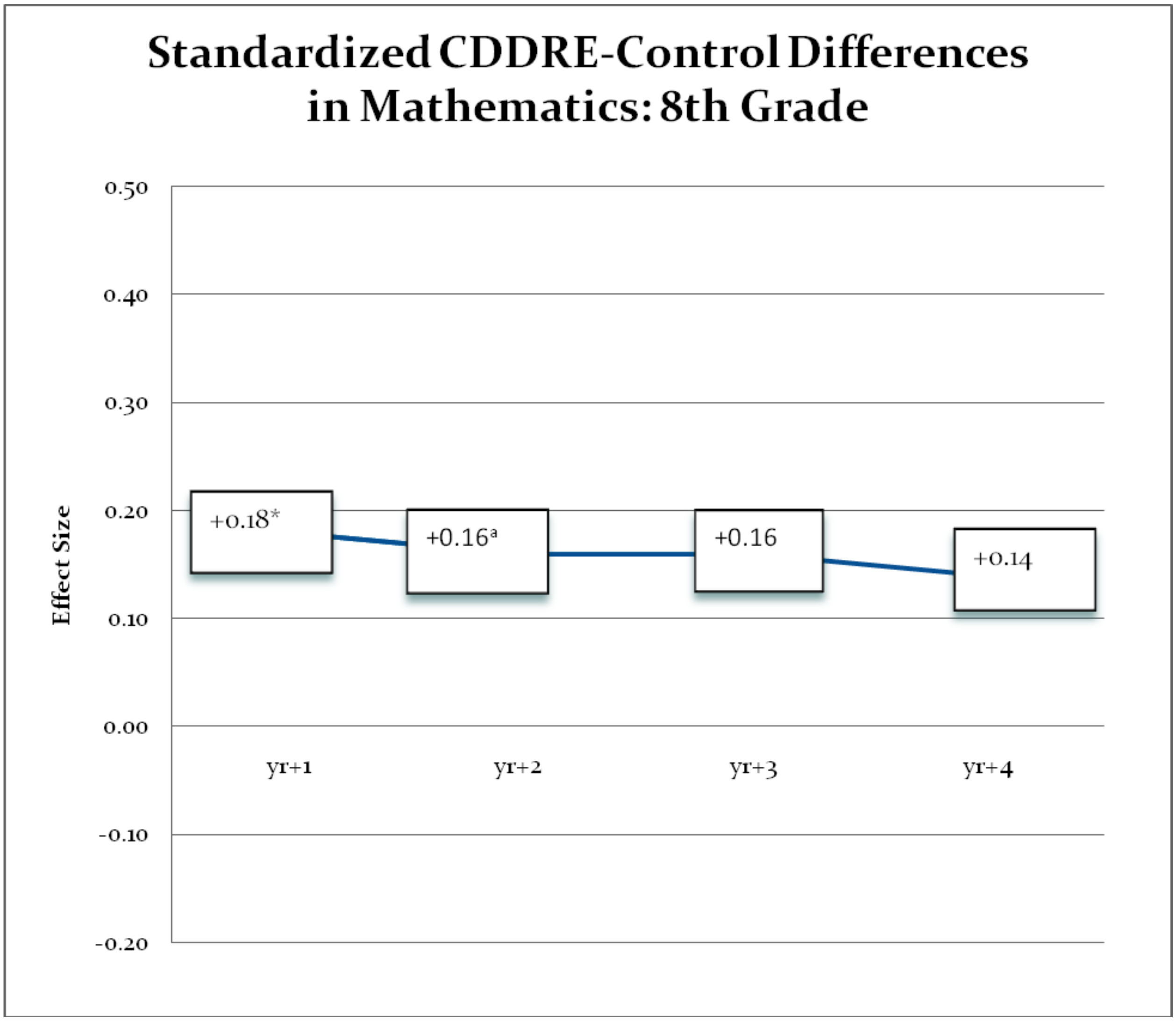


Adjusted posttest scores

	Baseline	**yr+1**	**yr+2**	**yr+3**	**yr+4**
Treatment	-0.10 (1.07)	-0.02 (1.05)	-0.02 (1.05)	-0.05 (1.05)	-0.05(1.07)
Control	0.09 (0.92)	0.01 (0.94)	0.01 (0.93)	0.03 (0.94)	0.04 (0.91)
ES	**-0.22**	**-0.03**	**-0.03**	**-0.08**	**-0.10**
p-value	0.02	0.32	0.36	0.39	0.15
N (schools)	T=263, C=295	T=263, C=295	T=220, C=230	T=126, C=134	T=80, C=86

a p< .10
* p< .05
** p< .01

Figure 6



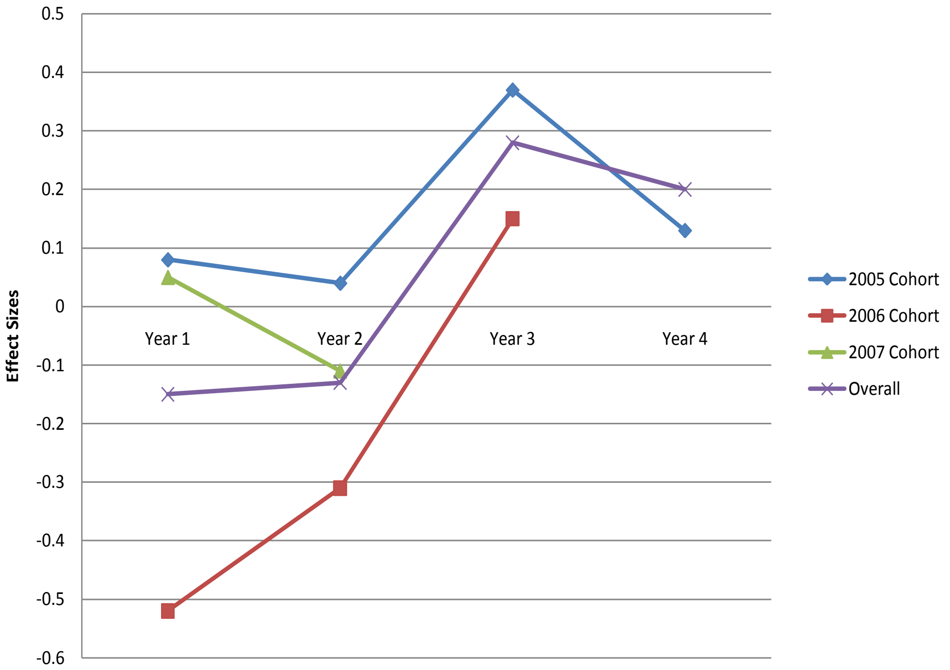
Adjusted posttest scores

	Baseline	**yr+1**	**yr+2**	**yr+3**	**yr+4**
Treatment	-0.03 (1.06)	0.09 (0.96)	0.08 (1.00)	0.07 (1.10)	0.07 (1.11)
Control	0.03 (0.92)	0.09 (1.00)	-0.08 (0.97)	-0.09 (0.90)	-0.05 (0.83)
ES	**-0.06**	**0.18**	**0.16**	**0.16**	**0.14**
p-value	0.59	0.02	0.06	0.13	0.45
N (schools)	T=158, C=167	T=158, C=167	T=134, C=136	T=62, C=78	T=33, C=41

a p< .10
* p< .05
** p< .01

Figure 7

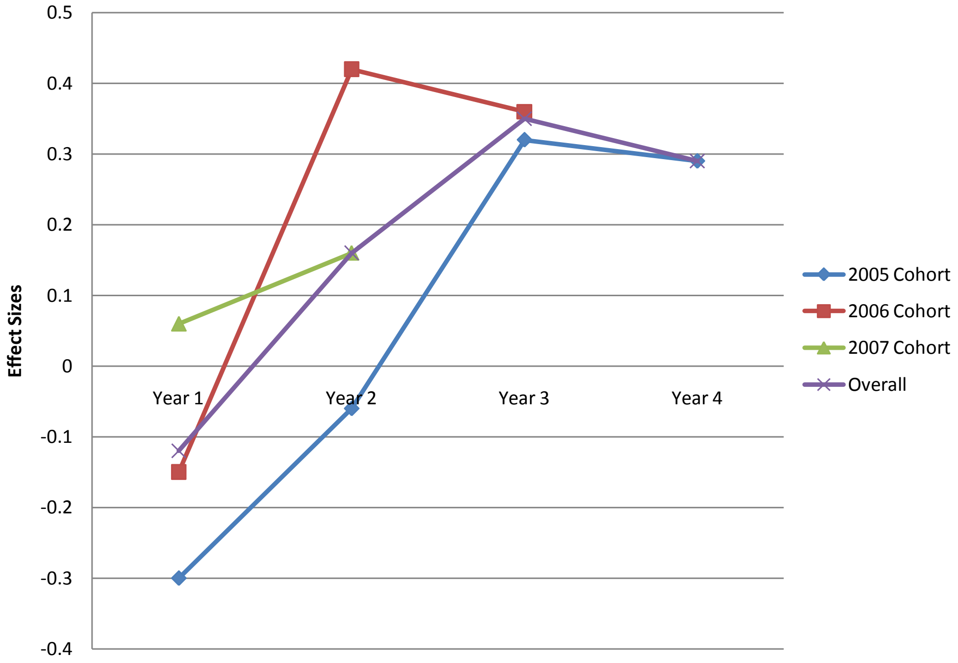
CDDRE 5th Grade Reading by Cohort



Adjusted Posttest Scores

	Baseline	Year 1	Year 2	Year 3	Year 4
2005 Cohort	0.10	0.08	0.04	0.37	0.13
2006 Cohort	-0.06	-0.52	-0.31	0.15	
2007 Cohort	-0.38	0.05	-0.11		
Overall	**-0.12**	**-0.15**	**-0.13**	**0.28**	**0.20**

Figure 8
CDDRE 8th Grade Reading by Cohort



Adjusted Posttest Scores

	Baseline	Year 1	Year 2	Year 3	Year 4
2005 Cohort	0.32	-0.3	-0.06	0.32	0.29
2006 Cohort	-0.45	-0.15	0.42	0.36	
2007 Cohort	-0.15	0.06	0.16		
Overall	**-0.04**	**-0.12**	**0.16**	**0.35**	**0.29**