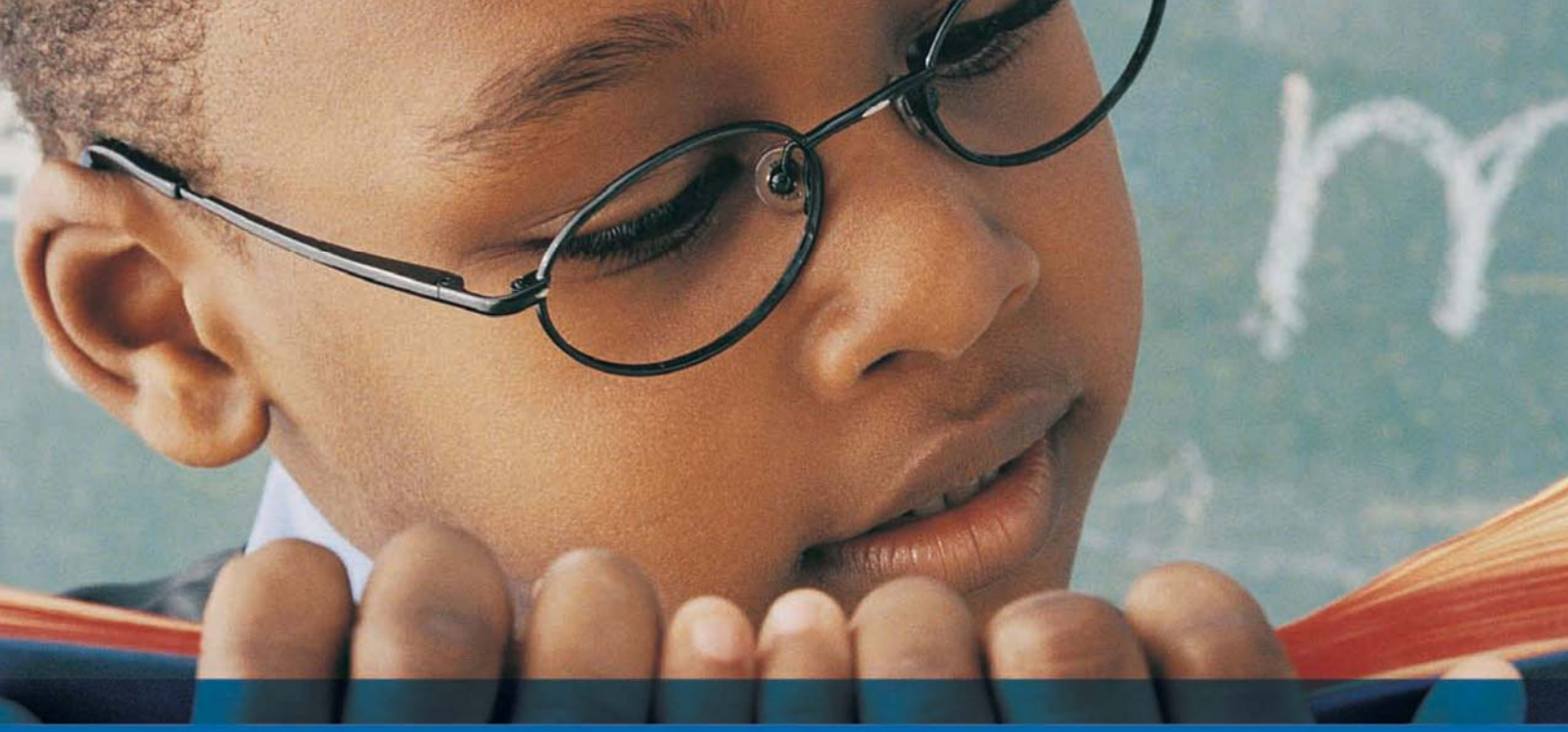


Evaluating the Implementation of Comprehensive School Reform and Its Impact on Growth in Student Achievement

Yu Zhang, Jamie Shkolnik, and Olatokunbo Fashola

American Institutes for Research®

1000 Thomas Jefferson Street, NW | Washington, D.C. 20007-3835 | 202.403.5000 | www.air.org

Evaluating the Implementation of Comprehensive School Reform and Its Impact on Growth in Student Achievement

A paper prepared for the annual meeting of the
American Educational Research Association, Montreal, Canada, April 11–15, 2005

Yu Zhang, American Institutes for Research®
Jamie Shkolnik, American Institutes for Research®
Olatokunbo Fashola, American Institutes for Research®

The authors thank Mike Garet, Daniel Aladjem, Kerstin LeFloch, Anja Kurki, James Taylor, Kazuaki Uekawa, Suzannah Herrmann, and Stephanie Miller for their very helpful comments on the previous version of this paper. The authors also express gratitude to Andrea Boyle and Alex Tombras who assisted in the data preparation.

The contents of this document were developed under a grant from the U.S. Department of Education (PR/Award Number R306S000012). The contents, however, do not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the federal government.

Contents

List of Figures

List of Tables

Evaluating the Implementation of Comprehensive School Reform and Its Impact on Growth in Student Achievement

Abstract

This study examines the relationship between the implementation of comprehensive school reform (CSR) and growth in student achievement. Survey data about CSR implementation and school-level achievement were collected in multiple years from a sample of CSR schools and compared with a sample of matched-pair schools. The sampled CSR schools adopted several promising CSR models. Findings indicate that implementation level of some components is a growth function of implementation length, with a large variation. On average, growth in mathematics and reading achievement among students at CSR schools is not consistently greater than the growth among similar students at comparison schools. However, after controlling for implementation level and limiting the implementation length to 3–5 years, overall CSR schools made larger growth in mathematics and reading achievement relative to the matched comparison schools when CSR schools had a high level of implementation or made progress in implementation level in some components over time. The relationship between CSR implementation level and growth in reading varies by CSR model.

Evaluating the Implementation of Comprehensive School Reform and Its Impact on Growth in Student Achievement

Introduction

This paper examines the relationship between comprehensive school reform (CSR) and student achievement gains. The earliest CSR design, the School Development Program, was introduced to U.S. schools in 1968, but the majority of CSR models were designed in the late 1980s (Herman et al., 2000). The CSR movement was introduced in the mid to late 1990s mostly because many of the existing reforms were unable to provide sustaining improvements in academic achievement. A few studies on national achievement trends and comparisons raised concerns about public education in America. The studies showed that students continued to perform lower academically than earlier U.S. cohorts and students from other industrialized countries (National Commission on Excellence in Education, 1983; Peak, 1996).

Although many efforts have been taken to reform schools through various methods—such as increasing funding, raising standards, improving professional development, and reorganizing schools—most have been discontinued after only a few years of implementation because they lacked effectiveness and failed to close the gap between wealthy and low income students (Cuban, 1984, 1990; Smith and O'Day, 1991; Darling-Hammond, 1994; Porter, 1994; Wang, Haertel, & Walberg, 1997). Instead focusing on failing schools, later studies approached the achievement gap by trying to determine the key characteristics that lead to high achievement in schools (Teddlie & Reynolds, 2000; Orfield, 1999; Murphy & Hallinger, 1993).

Incorporating characteristics of successful schools into the foundations of CSR designs can help to comprehensively transform low-achieving schools into model-like schools. A comprehensive approach becomes a CSR model when its components, materials, and methods of implementation can be assembled into an operational kit that is both understandable and affordable for schools. Some developers have evaluated their own studies and found positive effects on student achievement. For instance, Comer (1988) evaluated the School Development Program, and Slavin and colleagues (1996, 2000) and Slavin (2001) reported on the effectiveness of Success for All (SFA).

The growing need for evidence of the effectiveness of CSR programs has become not only an incentive for schools in need of improvement, but it has also attracted more attention and support from policymakers at different levels. In 1997, Congress created the Comprehensive School Reform Demonstration Program. The program is now known as the Comprehensive School Reform Program. Through it, grantee schools receive a minimum of $50,000 per year for 3 years to implement CSR programs. Today, the federal government's support of schoolwide projects, fueled by available funds, makes it possible for additional schools to adopt CSR models. Since 1998, nearly 6,000 schools have received funds and implemented more than 700 different CSR models (Southwest Educational Development Laboratory, n.d.).

To ensure that CSR Program funds are being used in compliance with federal guidelines, the No Child Left Behind (NCLB) Act of 2001 (SEC 1606, Part F) specifies criteria for a CSR program. One of these

criteria stipulates that a CSR Program-funded school has to demonstrate a model that "(11)(A) has been found, through scientifically based research to significantly improve the academic achievement of students participating in such program as compared to students in schools who have not participated in such program; or (B) has been found to have strong evidence that such program will significantly improve the academic achievement of participating children" (U.S. Department of Education, 2002b). These criteria establish clearly that the primary goal of implementing CSR is to improve student achievement. However, neither piece of legislation (i.e., the CSR Program or the NCLB Act) specifies which CSR designs have a strong effect on student achievement, or identifies who is responsible for evaluating the effects of federally funded CSR designs.

Although research on CSR models has been conducted for more than 10 years, measuring implementation of CSR is still an ongoing issue in existing literature and current research. Studies of the effects of CSR programs on student achievement have been inconsistent (Borman, Hewes, Overman, & Brown, 2003; Berends, Kirby, Naftel, & McKelvey, 2001; Slavin & Fashola, 1998; Herman et al., 2000, 1999; Cook et al., 1999; Cook, Hunt, & Murphy, 2000). These mixed findings indicate that at least three large issues confound the understanding of studies on CSR effects:

- **Time.** The pace and level at which CSR is implemented varies from school to school.
- **Design.** The various emphases and structures of CSR designs result in different levels of effect.
- **Rigor.** CSR programs are designed, conducted, and evaluated at various levels of rigor.

One way to verify the causal effect of CSR on student achievement is to attempt to replicate the positive findings, regardless of who designed the program. A reliable inference stems from the rigor of study design and analytical methods. An unbiased estimate of CSR's effect must be independent of persons conducting the evaluation and the design being studied. This, however, is not the case for the studies of CSR. Some earlier studies have been subject to bias because of less rigorous study designs or less objective conclusions reached by the evaluators (Borman et al., 2003).

This paper attempts to explain a relationship between CSR implementation and student achievement by using a larger sample of schools that are implementing various types of CSR models using a more rigorous research design than earlier studies.

The Theoretical Relationship Between CSR Implementation and Student Achievement

To some extent, many CSR models are research based. Some well-known CSR models are based on well-known educational theories or research experiences in teaching and learning practice. For instance, James Comer developed the School Development Program based on his experience and beliefs in the connections among social climate, interrelationships, and students' academic performance (Comer, 1984). Comer's programs center on schools with high concentrations of low-achieving or minority students and in which social networks between parents and schools are often lacking and a supportive climate for academic performance is less likely to be observed. Comer's program aims to build social support in schools and close the social and cultural distances between students and teachers. Comer's theory is supported by theories on the effect of social and cultural capital on student achievement (Coleman, 1988; Bourdieu, 1986). Empirically, evidence suggests that in schools with high concentrations of students from

single-parent families, students perform better when their parents have strong social relationships with the school and other parents (Epstein & Hollifield, 1996; Pong, 1998).

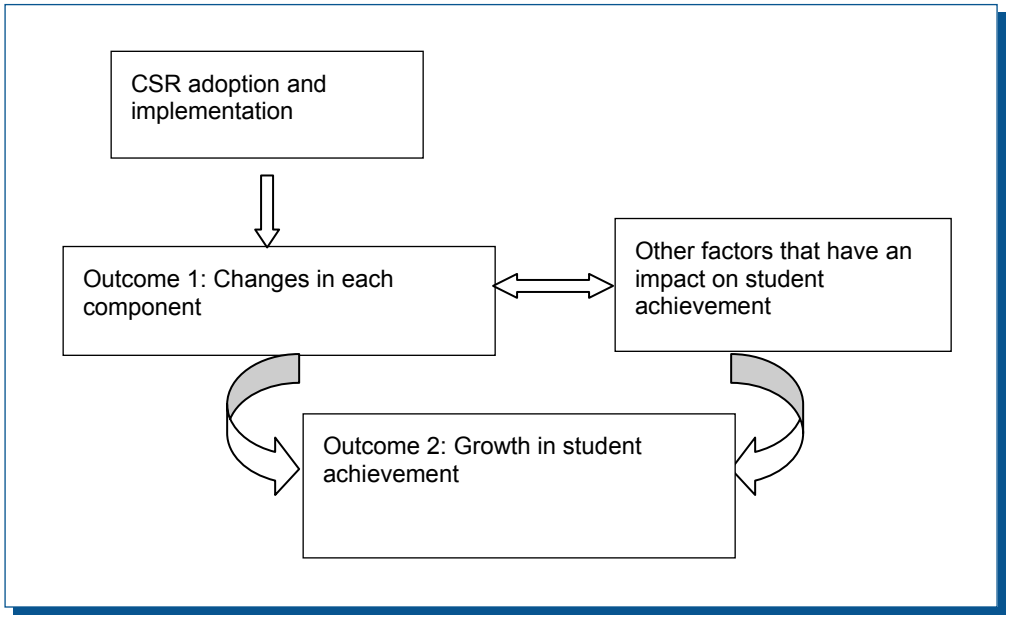
Henry Levin developed the Accelerated Schools Project (ASP)—a CSR model that emphasizes expanding the faith and potential for success of all children, which is consistent with John Dewey's effective education (Dewey, 1966). Levin believes that combining an accelerated pace of learning with a rich and challenging curriculum is the way to help low-achieving students catch up academically. The philosophy behind the ASP is that the way to help low-performing students to catch up academically is to combine an accelerated rate of learning with a rich and challenging curriculum (Levin, 1989).

Robert Slavin and Nancy Madden and their colleagues at Johns Hopkins University developed SFA. Their work began as a response to school problems in Baltimore city. They generalized their approaches to address the common problems faced by failing urban schools. SFA addresses reading challenges of at-risk students and is supported by theories of early intervention and prevention. The program's developers believe that immediate and intensive intervention for reading-related learning problems at an early stage is more effective than waiting until students lag behind, at which time they may need special assistance (Slavin et al., 1996; Wasik & Slavin, 1993).

Unquestionably, an effective way to prevent failure in academic performance is to provide students with an environment that encourages learning and cognitive development. CSR developers attempt to create a new environment through rebuilding whole schools. Based on various perspectives and beliefs about the best way to help low-achieving students, it is natural for developers from different fields to design models based on various theories and philosophies and their own theoretical frameworks. CSR designs vary in emphases, but they often share common goals and characteristics. Part of this commonality among CSR designs is reflected in 11 core components of CSR (U.S. Department of Education, 2002a). Because these components cover most major aspects of schooling, implementing and improving these core components is expected to eventually lead to reforming these schools comprehensively. The path of changing core components of school organization to improve student achievement is supported by some theoretical assumptions and empirical studies on school effectiveness, structure, and organization (Teddlie & Reynolds, 2000; Orfield, 1999; Murphy & Hallinger, 1993). The fragmented nature of early reform strategies may be a contributing factor to why such strategies may not have been able make balanced changes to the whole school and may be the reason that some of these strategies failed. In comparison, a CSR model implies changes for the whole school, maintaining consistency and coherence of school activities within the balanced changes. The expectation is that when a CSR model is implemented successfully, the school becomes completely new and model-like, school effectiveness and improvement follow, and student achievement will eventually improve.

Broadly speaking, successful CSR adoption and implementation should lead to two outcomes. First, schools are reorganized to fit the CSR design components. Second, the changes resulting from implementation make schools functional for educational productivity, and thus growth in student achievement proceeds gradually (see Figure 1). Logistically, student achievement improves slowly during the first 2 years of implementation and accelerates in year 3 and after.

Figure 1. Theoretical Framework of CSR Implementation and Student Achievement



The Empirical Relationship Between CSR Implementation and Student Achievement

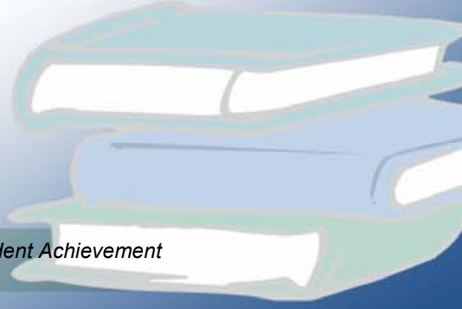
Many evaluations have found that implementing CSR changes schools in some aspects (Desimone, 2002). However, the extent to which these initial changes will continue to affect student achievement is still uncertain. In an earlier review of CSR, Herman and colleagues (1999) reviewed studies on CSR designs and found that only three CSR models demonstrated strong effects on student learning: Direct Instruction and SFA (for grades K–6), and High Schools That Work (for grades 9–12). Some CSR models showed marginal effects: Community for Learning, Different Ways of Knowing, Expeditionary Learning, Outward Bound, and School Development Program. Other models revealed weaker effects or sometimes no effects.

Some more recent studies have also reported evidence of CSR's effects. Bloom, Rock, Ham, Melton, and O'Brien (2001) found that students attending some schools implementing ASP scored significantly higher in 3rd-grade mathematics and reading at the end of 5th year of CSR implementation, compared to their achievement levels in the year prior to adoption of the model. Borman and his colleagues (2002, in press) demonstrated that the implementation of SFA had significant effects on students' reading scores. Cook and his evaluation team (2000) found that student achievement improved in Chicago schools that implemented the School Development Program. Findings on effectiveness of CSR models, however, are not always consistent even within the model designs. For example, Cook and his team (1999) also conducted a study of the School Development Program in Prince George's County, Maryland, but they failed to find a significant effect similar to that found in Chicago. Similarly, the effect of Co-nect was reported in Herman and her and colleagues' study (1999), but a recent study by Ross and Lowther (2003) showed mixed results in student achievement.

Borman and his colleagues (2003) used a meta-analysis to summarize findings from existing studies on the overall effects of CSR. They concluded that the effects of CSR on student achievement are positive overall and promising. However, they also cautioned against hastily reaching false conclusions about these findings given the possibility that (a) earlier findings may be less reliable, (b) evaluations performed by the CSR developers may yield estimates of effects higher than evaluations performed by others, and (c) studies using experimental or quasi-experimental treatment–control comparisons may yield effect estimates lower than studies based on analyses of CSR pre- and post-gain scores. Because studies that failed to reject the null hypothesis are less likely to be published, a meta-analysis based solely on published studies cannot fully correct for the bias of overestimated effect (Light & Pillemer, 1984). Other researchers have also expressed similar concerns in that because most of the earlier evaluations of CSR effects were conducted by models' developers or their affiliates, their conclusions may not be as objective as those conducted by independent parties (Desimone, 2000; Slavin, 2002). Moreover, the study methods used in earlier evaluation studies are not rigorous enough to generate reliable results (Cook, 2002). A few studies using experimental designs have found CSR positive effects on, for example, SFA (Borman et al., in press) and School Development Program (Cook et al., 1999; 2000). The effects of the other numerous CSR designs have not been studied with rigorous designs and objective evaluators, especially studies with samples composed of different model designations and implementations in various contexts.

In accordance with the theories of CSR, it is reasonable to expect growth in student achievement in many evaluation studies, regardless of the size of the impact. Empirical studies, however, have already revealed that CSR implementation does not guarantee improvement in student achievement. Although studies with rigorous designs and relevant samples would improve the power and precision of detecting CSR effects, researchers must realize that variations in designs and emphases in implementation will continually challenge studies on the relationship between CSR implementation and student achievement. Because CSR implementation is a complex process and its success depends somewhat on available resources and additional support from outside the school setting (e.g., funding), some schools experience difficulties and phase in the implementation of CSR more slowly than others (Glennan, 1998; Berends et al., 2001). Failure to fully implement a CSR model makes it difficult to attribute growth in student achievement to implementation of the model. A closer view of the challenges of CSR implementation and the variation across designs is essential to the study of the association between CSR implementation and student achievement.

The nature of CSR sometimes makes it difficult to successfully implement the program and ultimately accomplish the goal of implementing student academic success. As such, schools often encounter difficulties or challenges that affect changes. Some challenges in implementation are related to the very nature of the schools in which CSR is most frequently implemented—that is, low-income and low-performing schools with high concentrations of low-performing students. In 2000, 85% of CSR Program schools were Title I schools (Southwest Educational Development Laboratory, n.d.). Approximately 70% of schools were located in large or mid-size cities or towns, and the average poverty rate of schools was 70% (U.S. Department of Education, 2000). Because of the characteristics of Title 1 schools, they are usually under pressure to improve, and CSR is a path that many of these schools take to improvement. School faculty turnover is also a frequently reported challenge. Staff turnover is usually high in schools that serve high populations of at-risk students. The turnover rate is usually even higher in schools in which model designs require substantive changes in the faculty's behaviors (Smith et al., 1997), and senior teachers are less likely to apply the CSR's instructional practice when the reform is mandated



(Ross et al., 1997; Slaton, Atwood, Shake, & Hales, 1997). Thus, a lot of the buy-in and initial professional development and teacher assistance are lost, and the second year becomes like the first year. Ross and colleagues, Stringfield, Williams, and Wright (2003), for instance, reported that the proportion of new teachers is higher in CSR schools than in matching non-CSR schools and that new teachers are also more likely to leave during the initial year of implementation (Newmann, Smith, Allensworth, & Bryk, 2002). One can definitely conclude that successful CSR implementation is impeded when instructional staff do not change their current practices, and this resistance by teachers makes it difficult to ensure coherence in instruction and CSR activities for improvement. The incoherence in instruction ultimately impedes student achievement (Newmann et al., 2002). Ultimately, the relationship between CSR implementation and growth in student achievement depends on increasing levels of implementation over time. The relationship between implementation and achievement is weaker when implementation remains flat or declines over time.

The nature of a CSR model's design may also challenge its implementation in some settings. It is very possible that the focus of CSR models may be inconsistent with the priorities or standards (e.g., curriculum, instruction, accountability, and reform approaches) set by the district or state that is attempting to implement the model. CSR implementation is facilitated if the model emphases do not conflict with priorities in school and district policies (Bodilly, 1996; Smith et al., 1997). Although most CSR designs prescribe implementation guidelines, various challenges in the practice of implementation create a priority-conflicting environment and can make the connection between CSR implementation and student performance uncertain at best. Some schools implement some but not all core components or receive different levels and types of support (Berends, Bodilly, & Kirby, 2002; Berends, 2000; Bodilly, 1998; Muncey & McQuillan, 1996). When schools experience consistent difficulties in implementing models, full implementation may be either ultimately delayed or remain incomplete. For instance, in a study conducted by RAND, researchers found that approximately half of the schools implemented the core components after 2 years of model adoption (Glennan, 1998). In another case, Bloom and colleagues (2001) found that many schools using the ASP model implemented components of curriculum and instruction in the 3rd or 4th years of implementation.

Research Questions

When a school fails to implement CSR successfully, regardless of the direction of student achievement, school level changes cannot be attributed to the CSR model, because it was not implemented correctly. To build such an inference, it is important to determine whether each step of implementation and improvement in student achievement, albeit gradual, can be observed. In this study, the relationship between student achievement and the movement from adoption to full implementation of a CSR model is examined through three sequential questions that serve as the main premise for this study:

1. Is the level of implementation a function of the number of years of implementation?
2. Is growth in student achievement a function of the number of years of implementation?
3. Is a high level of or positive change in implementation over time associated with growth in student achievement?

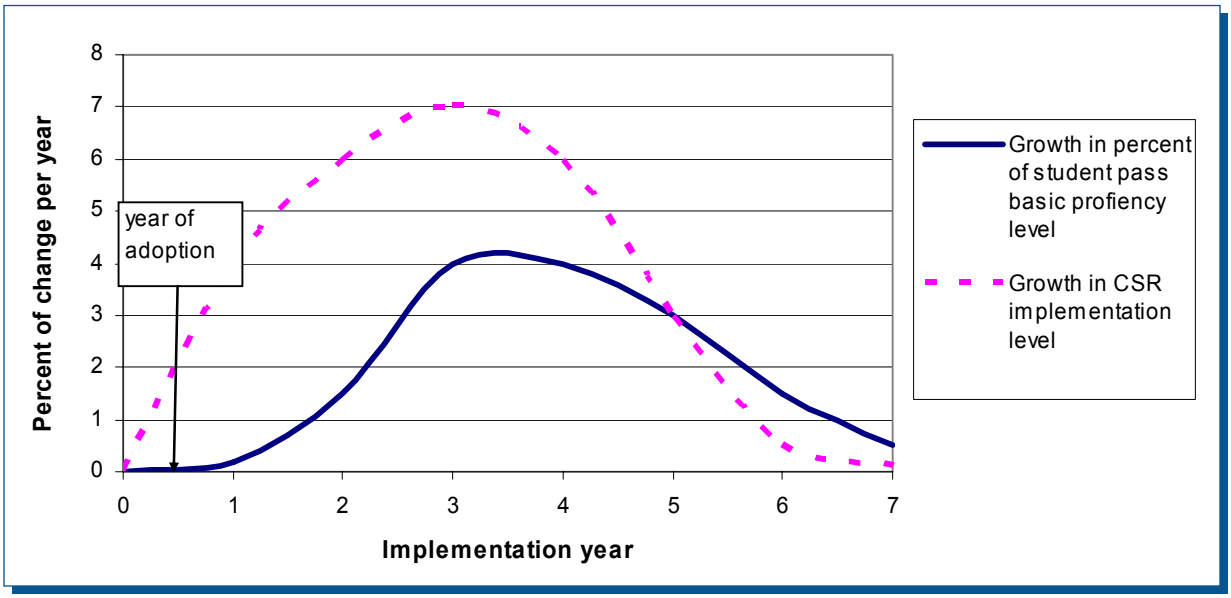
Question 1 asks whether CSR schools with longer implementation periods have a higher level of implementation. The level of implementation is expected to improve as the number of years of

implementation increases. The level of implementation is also expected to increase rapidly during the early implementation years and then plateau during the later years. For question 2, changes in student achievement are expected to lag slightly behind the changes in the level of implementation. CSR schools are also expected to experience greater changes in student achievement than comparable schools not implementing CSR. Similarly, with CSR treatment, improvements in student achievement in CSR schools are not expected to be lower than comparable non-CSR schools. These hypotheses create the possibility of a relationship between CSR implementation and improvement in student achievement. Question 3 examines the positive direction of this link.

The relationship between the first two questions is sequential and reflects the possibility that a later event (student achievement) is conditional upon a prior event (the degree of success of CSR implementation). According to this logic, in longitudinal observations over implementation years, one should expect (a) increase in implementation level, (b) improvement in student achievement, (c) greater improvements in student achievement in schools with greater increases in implementation, (d) uneven improvements in implementation and student achievement in different implementation years, and (e) improvement variations in CSR design. These five anticipated results are presented as tools to help the reader understand the hypothetical relationship between CSR and growth in student achievement.

Figure 2 shows the expected association between implementation and student achievement. Changes in both implementation level and student achievement are sequential and may be a function of implementation year, with unequal ratios of increases of student achievement from year to year (Garet, Zhang, & O'Day, 2003). Both increases in implementation and student achievement may increase to a certain point and then stop or even decrease. If the pattern displayed in Figure 2 represents an accurate relationship between CSR implementation and growth in student achievement, then expecting a similar rate of growth regardless of the implementation level or length of implementation is actually difficult and unrealistic. For example, in schools in which CSR has been implemented for more than 5 years, the level of implementation gain would drop rather than improve during the subsequent implementation years. Also, in schools in which CSR has been implemented for 1 or 2 years, the growth rate in student achievement may not be of the same magnitude as the growth rate of the level of implementation. In other words, during the initial phase-in of CSR implementation, initial observations of large gains in student achievement may lag behind initial observations of large gains in implementation (within the CSR school). Thus, the relationship is complex, but ultimately the length, change, and level of implementation are the main factors when modeling the relationship.

Figure 2. Expected Relationship Between Student Achievement and CSR Implementation, by Implementation Year



CSR models can be distinguished based on the extent to which they emphasize core components; that is, the extent to which a specific component is central to the CSR magnitude of emphasis on each component within each design (Fast, Aladjem, Shive, Herman, & Carter, 2001). Although CSR models share core components, they differ in model structure. Herman et al. (2000) indicate that CSR models are either structure or philosophy based. A structure-based model prescribes instruction with specific content and strategy. A philosophy-based model provides a general approach and requires teachers to implement the philosophy without the help of instruction- or curriculum-related professional development. Structure-based models that focus on instruction or curriculum are usually more likely to be implemented consistently than philosophy-based models.

CSR models can also be distinguished based on their primary emphases. A CSR model may have one or more focus (e.g., curriculum, instruction, or governance). For example, the Direct Instruction and SFA models are classified as structure-based models that emphasize curriculum and instruction. ASP and ATLAS (Authentic Learning and Assessment for All Students) Communities are philosophy-based models, but the former emphasizes curriculum and the latter focuses on school governance (Herman et al., 2000).

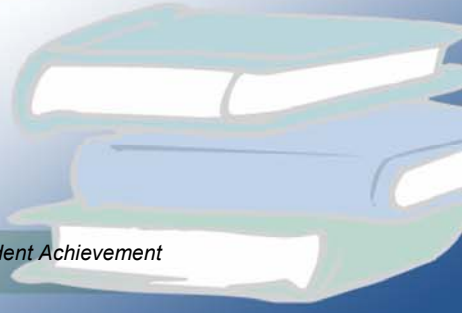
The specific CSR model being implemented by the school is another factor that can contribute to variation between implementation and growth. Accounting for this factor, this study included several CSR designs in its evaluation. The possible variation in the types of CSR designs in this study makes it possible to ask a fourth research question: How are CSR implementation levels and growth in student achievement associated across CSR models?

Data

To address the four research questions, several years of data are needed on levels of CSR implementation and school level student achievement. This study uses a quasi-experimental design to examine information from 3 years of CSR implementation in a sample of 649 schools from 21 school districts in 16 states. CSR schools were selected from lists provided by CSR model developers. The comparison schools were selected and matched with CSR schools from the same district serving the same grade levels. The matching was based on two school demographic characteristics—school minority composition and percentage of students who receive free or reduced-price lunch, which combined are often used as indicators of student achievement. The estimated effects of these two characteristics were weighted based on an analysis of existing data. CSR and comparison schools were paired based on closeness to specific matched comparison schools, as measured by Euclidean distance.

In 2001, the sampled CSR schools were implementing 8 CSR models. Three waves of data were collected from principals and districts in 2002, 2003, and 2004. Two waves of data were collected from teachers in 2002 and 2004. Measures on implementation level were developed from the survey data. The survey instruments were developed to address most core components of CSR, such as school leadership, professional development, and pedagogy. Principals responded to the same set of survey items for 3 years, and the variations in their responses across years were expected to reflect changes in implementation level. Because the developers of different CSR models may provide different responses to a survey item regarding the level of implementation, model developers were asked to describe in the survey what a full implementation of their program would look like. Their responses were used to assess the level of implementation fidelity. When a school's responses to a survey were similar to a model developer's responses, a higher implementation level rating was assigned to the school. Kurki, Aladjem and Carter (2005) present detailed information about how the measures of implementation were created.

School-level aggregate achievement data were collected across 5 years from 1999 to 2003. This includes information about the districts from before the survey was distributed and during the 3 years the survey was administered. The achievement measures consisted of standardized test scores, which vary by state. To reduce the variation in score matrices, test scores were standardized within school districts into *z*-scores (with mean of 0 and standard deviation of 1). This procedure converts test scores from different matrices and scales into a common scale. The value in the new score matrix represents a school's position within the district relative to the district's mean, as measured by the district's standard deviation, independent of the original scale. The changes in *z*-scores of school-level achievement from year to year reflect the gains that schools made in their position ranked within the district. For example, in adjacent years, if the *z*-score of a school in the second year is higher than the first year, it means that this school performed better in the second year relative to other schools in the district, compared to its first-year score relative to the score of other schools in the district. The gain of standardized score in 2 years for a school reflects the change this school made in student achievement relative to the district average, regardless of its ranking position. Of course, for schools at different positions, the efforts to make the same magnitude of gain are not equal; that is, it may be more difficult for a high-achieving school to make the same amount of improvement as a low-achieving school. However, this is less likely to be a problem in this study because both CSR and comparison schools were low-achieving schools at the beginning of the study.



During the years of school-level achievement data collection, some states tested only a few grades, and the grades tested varied across states. Thus, picking any single tested grade as the measure of student achievement would be perceived as arbitrary. Because the average of all tested grades provides more reliable school-level student achievement than any single tested grade, the average test score of grades 2–5 was used as the measure for elementary schools and grades 6–8 as the measure of middle schools. When a score was reported in more than one format, an appropriate test was selected based on our understanding of the score property and the availability of the score across years. The score format in order of preference was scale scores, normal curve equivalents or percentile rank, and percentage of students passing the lowest proficiency level. The proficiency scores usually were reported at three levels. The lowest proficiency was used out of a concern that most CSR schools were low-performing schools. Thus, the largest gains they would make should be more apparent in the improvement, based on the percentage of students meeting the lowest level of proficiency. The expectation was that the average *z*-score would be negative, indicating a rank below the district average. Test scores used in this study are listed by state in Appendix A.

Analytical Approaches

The analysis consists of four parts, each corresponding with the four research questions. To examine the first research question (Is the level of implementation a function of the number of years of implementation?), mean differences in implementation scores over time were calculated, and the scores were tested to determine whether the mean differences were significantly different from zero. A positive value that was significantly different from zero indicated an increase in implementation level. The differences were tested using a univariate *t* test for each implementation score and a multivariate *T*-square test for the combined mean of multiple implementation scores. It is possible that not all implementation scores increased over time. We expected the combined average difference to be positive and significant, indicating improvement of implementation over time.

Because the standardizing procedure is a linear transformation and not an equating procedure, the standardized score is only valid for comparing schools within districts, not across them. For this reason, the most straightforward and relevant way to address the second research question (Is growth in student achievement a function of the number of years of implementation?) is to conduct analysis within districts and use meta-analysis to summarize the results. This approach ensures that the estimated average changes in student achievement will be free of error variance given the variety of test formats across districts. There is, however, a trade-off for the within-the-district analysis in that a small sample size will yield a small power of analysis. Because of the small sample size of each district, the chance of detecting a change in student achievement from year to year is low. Therefore, to improve power rankings, the sample size must be increased by pooling schools together across districts. The matched-pair design provides one possible solution, by pooling all schools together and creating a variable that measures the performance of CSR schools relative to their paired school counterparts.

In the first step of the analysis, the differences in *z*-scores were calculated between CSR schools and their matched schools from 1999 to 2003. Taking the difference between the matched-pair schools minimizes the variance among districts, given the varied tests among them. This property of the difference in *z*-score permits the pooling together of schools among districts and at the same time controls for some school-level background variables.

Two approaches are often used to model achievement data in hierarchical linear modeling (HLM). One is value-added and another is time-series. These two approaches address different questions. The value-added approach models the academic gains of students from one grade to the next or from 1 year to the next. The time-series approach models the improvement of a particular school or grade in a selected time series. This study uses a two-level, time-series HLM approach—setting year at level 1 and schools at level 2—to measure CSR schools making improvement in student achievement in 4 academic years from 2000 to 2003, relative to matched comparison schools. The equations of this model are presented below.

Time-Level Model

$$\eta_{ij} = \beta_{0j} + \beta_{1j} Score99_{ij} + \beta_{2j} Year01_{ij} + \beta_{3j} Year02_{ij} + \beta_{4j} Year03_{ij} + \beta_{5j} SchoolSize_{ij} + \beta_{6j} Pct\_Minority_{ij} + \beta_{7j} Pct\_Freelunch_{ij} + \beta_{8j} Student\_teacher\_ratio_{ij} + \beta_{9j} SchoolSize\_Dif_{ij} + \beta_{10j} Pct\_Minority\_Dif_{ij} + \beta_{11j} Pct\_Freelunch\_Dif_{ij} + \beta_{12j} Student\_teacher\_ratio\_Dif_{ij} + \varepsilon_{ij}$$

where:

η_{ij} is the difference in *z*-score between CSR school *i* and its paired school in year *j*; and

$Score99_{ij}$ is the difference in *z*-score between school *j* and its pair in 1999.

The following control variables are standardized by the grand mean and standard deviation:

$Year01_{ij}$ is coded 1 for 2001 and 0 for other years.

$Year02_{ij}$ is coded 1 for 2002 and 0 for other years.

$Year03_{ij}$ is coded 1 for 2003 and 0 for other years.

$SchoolSize_{ij}$ is the average number of students enrolled in school *j*.

$Pct\_Minority_{ij}$ is the average percentage of minority students enrolled in school *j*.

$Pct\_Freelunch_{ij}$ is the average percentage of students eligible for free lunch in school *j*.

$Student\_teacher\_ratio_{ij}$ is the average student–teacher ratio in school *j*.

The following control variables are not centered:

SchoolSize_Dif$_{ij}$ is the difference in the average number of students enrolled between paired schools.

Pct_Minority_Dif$_{ij}$ is the difference in the average percentage of minority students between paired schools.

Pct_Freelunch_Dif$_{ij}$ is the difference in the average percentage of students eligible for free lunch between paired schools.

Student_teacher_ratio_Dif$_{ij}$ is the difference in the average student–teacher ratio between paired schools.

ε_{ij} is the probability of CSR schools improving more than matched-pair comparison schools.

School-Level Model

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ . \\ . \\ \beta_{12j} \end{bmatrix} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ . \\ . \\ \gamma_{120} \end{bmatrix} + \begin{bmatrix} \nu_{0j} \\ 0 \\ . \\ . \\ 0 \end{bmatrix}$$

where:

γ_{00} is the average intercept;

γ_{i0} i = 1 to 12 are average slope; and

ν_{0j} is a random error term representing unmeasured factors related to the intercept of the growth curve for schools *j*.

The primary hypotheses of interest in the model described above concern differences in achievement trajectories of *z*-score differences between CSR and comparison schools over time from 2000 to 2003. The intercept and three time variables of Year01, Year02, and Year03 test these year-specific hypotheses. The intercept reflects the *z*-score made by CSR schools relative to matched-pair comparison schools in 2000, and the three time variables reflect the three changes made in the following 3 years. The hypothesized effects the model seeks are reflected by the coefficient of the intercept and the three time variables. If CSR implementation has a positive effect on achievement, then we would expect that the intercept or coefficients of time variables would be positive. The coefficients of time variables Year01, Year02, and Year03 test the changes in the difference between pairs across years. A positive value indicates improvement in achievement compared with the base year (2000). When testing the growth rate in the line of implementation years, the variable measures of the number of implementation years replace the dummy-year variables.

The third research question examines whether high levels of implementation and changes over time are positively associated with growth in student achievement (see the description of statistical model in Appendix B). To perform this analysis, two variables that measure implementation level as independent variables were added to the model described above. The first variable measures the implementation level in 2002. It indicates whether CSR schools with a higher implementation level in 2002 had larger academic achievement gains than matched comparison schools. The second variable measured the changes in implementation level from 2002 to 2004. It indicates whether CSR schools that exhibited improved implementation levels from 2002 to 2004 were more likely to make more growth in student achievement than matched comparison schools.

Results

Table 1 reports descriptive statistics of background variables of the sampled schools. Sampled schools were rematched after the first wave of survey data was collected in 2002. [1] The rematching resulted in 395 pairs: 39 CSR Model B, 19 CSR Model A, 30 CSR Model C, 135 CSR Model F, 171 schools implementing other models, and one school implementing an unknown model. In 2002, the average length of CSR implementation among the sampled schools was approximately 3.5 years. CSR schools and matched comparison schools were very similar in terms of the percentage of students who were eligible for free or reduced-price lunch, the percentage of minority students, school size, and student–teacher ratio. The similarities reflect the success in the matched-pair design.

Table 1. Descriptive Statistics of School Background Variables

Variable	CSR schools		Comparison schools		Difference
	Mean	SD	Mean	SD	Mean
Percentage of students who were eligible for free or reduced-price lunch	0.81	0.18	0.78	0.20	0.03
Percentage of minority students	0.90	0.17	0.90	0.18	0.00
School size (unit per 1,000 students)	0.63	0.32	0.64	0.36	-0.01
Average student–teacher ratio (unit of 10 students per teacher)	1.66	4.40	1.77	4.90	-0.11

Note. n = 395.

[1] Because some comparison schools adopted CSR in the years after we completed sampling selection, some of the CSR schools lost matched-paired comparison schools as some comparison schools became CSR schools. When calculating the difference in *z*-score for pairs, the average *z*-score of comparison schools within the district was used for CSR schools without pairs.

Research Question 1: Is the Level of Implementation a Function of the Number of Years of Implementation?

This question examines whether the level of implementation is a function of implementation length. To examine this question, implementation scores from 2002 were compared with those from 2004. Kurki et al. (2005) developed a set of implementation scores for each of these years. An increase in the implementation scores from 2002 to 2004 indicates that the implementation level is a function of implementation length. Table 2 presents the average principal-reported implementation levels on five variables: shared decision making among principals and teachers, school organization and governance, inclusion of special education students, student grouping, and parent–community involvement. All five implementation measures changed significantly. There was a large increase from 2002 to 2004 in implementing component of inclusion. The implementation level on shared decision making and student grouping increased slightly, and the relationship between level of implementation on school organization and parent involvement slightly decreased. The large standard deviation of paired-difference indicates that schools may have made uneven changes from 2002 to 2004, even though the average change was only 2–3% on some components.

Table 2. Average Principal-Reported Implementation Levels on Five Variables in 2002 and 2004

	Reported in 2002			Reported in 2004			Change	
Variable	N	Mean	SD	N	Mean	SD	Average paired-difference	Std
Shared decision making among principals and teachers	214	0.84	0.11	189	0.86	0.09	168	0.02**	0.12
School organization and governance	223	0.71	0.17	196	0.58	0.18	174	-0.13***	0.22
Inclusion of special education students	223	0.63	0.19	195	0.78	0.17	173	0.15***	0.23
Student grouping	223	0.60	0.16	195	0.63	0.14	173	0.03**	0.18
Parent–community involvement	223	0.82	0.18	196	0.80	0.22	174	-0.03**	0.18

Note. The multivariate analysis of variance (MANOVA) paired T^2 test on the difference resulted in 0.56 for Wilks' lambda and $p < .001$, indicating that the combined mean difference is significantly different from zero.

**p*-value significant at 0.1 level. ***p*-value significant at 0.05 level. ***p-value significant at 0.01 level from *t* test for paired difference.

Measures of implementation for instruction, curriculum, and professional development were derived from answers to questions on the teacher survey. Table 3 presents the average teacher-reported implementation levels on eight variables in 2002 and 2004: use of technology in instruction, inclusion of special education students, student grouping, parent–community involvement, curriculum, teaching time, type of professional development, and pedagogy. The large standard deviation of the changes indicates a large variation in changes of implementation level among the sampled schools. It provides evidence that CSR implementation influences schools, but the pace of change and the time of change varies greatly, or the implementation level is a function of implementation length, with uneven changes of pace. Teacher-

reported implementation of parent–community involvement increased (see Table 3), but principal-reported implementation of parent–community involvement decreased (see Table 2). This inconsistency could reflect differences between principals and teachers as a result of their perspectives and roles.

Table 3. Average Teacher-Reported Implementation Levels on Six Variables in 2002 and 2004

	Reported in 2002			Reported in 2004			Change		
Variable	N	Mean	SD	N	Mean	SD	N	Difference	SD
Use of technology in instruction	235	0.54	0.14	234	0.56	0.13	229	0.02*	0.16
Inclusion of special education students	226	0.72	0.21	215	0.77	0.15	211	0.05***	0.17
Student grouping	251	0.87	0.09	249	0.91	0.08	245	0.04***	0.10
Parent–community involvement	226	0.56	0.10	225	0.63	0.11	221	0.07***	0.10
Curriculum	216	0.82	0.08	206	0.82	0.08	199	0.00	0.08
Teaching time	251	0.87	0.12	249	0.87	0.14	245	0.00	0.07
Type of professional development	251	0.75	0.13	240	0.75	0.15	245	0.00	0.16
Pedagogy	235	0.89	0.09	225	0.90	0.10	217	0.01	0.06

Note. The MANOVA paired T^2 test on the difference resulted in 0.65 for Wilks' lambda and $p < .001$, indicating the combined mean difference is significantly different from zero.

*p-value significant at 0.1 level. **p-value significant at 0.05 level. ***p-value significant at 0.01 level, from t test for paired difference.

Research Question 2: Is Growth in Student Achievement a Function of the Number of Years of Implementation?

Assuming, upon CSR adoption, that a relationship exists between CSR implementation and growth in student achievement from year to year, CSR schools should be expected to experience greater gains in student achievement relative to matched comparison schools. From 1999 to 2003, z-scores were calculated, and the differences between the CSR schools and matched comparison schools were computed. Table 4 displays the average z-scores of CSR schools and matched comparison schools and the differences between them. The number reported in the column "Difference" indicates how CSR schools were ranked in terms of academic performance relative to matched comparison schools within districts. This is the dependent variable used in all regression models reported hereafter. A positive number means that CSR schools were ranked higher than matched comparison schools. A number close to zero means that CSR schools performed similar to matched comparison schools. In spring 1999, which is the baseline in the time series, CSR schools were ranked lower than matched comparison schools in both mathematics and reading. In spring 2002, the achievement level of CSR schools was still lower than matched comparison schools. However, the gap was closed slightly, indicating that CSR schools had accomplished larger gains in academic achievement than matched comparison schools from year to year. Whether this difference is larger at a significant level is subject to statistical tests.

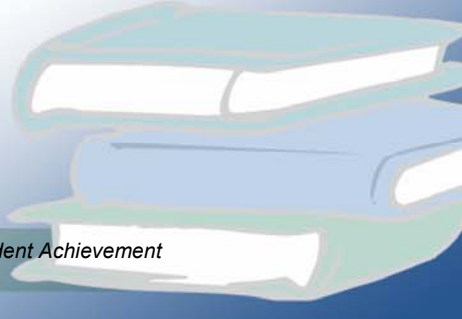


Table 4. Average Mathematics and Reading *z*-Scores Between CSR and Matched Comparison Schools by Year

	CSR schools		Comparison schools		Difference	
Variable	Mean	SD	Mean	SD	Mean	SD
Mathematics *z*-score, spring 1999	-0.46	0.75	-0.15	0.66	-0.32	0.88
Mathematics *z*-score, spring 2000	-0.47	0.77	-0.10	0.68	-0.37	0.88
Mathematics *z*-score, spring 2001	-0.40	0.83	-0.19	0.70	-0.21	0.96
Mathematics *z*-score, spring 2002	-0.41	0.76	-0.13	0.69	-0.28	0.92
Mathematics *z*-score, spring 2003	-0.44	0.81	-0.16	0.73	-0.28	0.98
Reading *z*-score, spring 1999	-0.47	0.74	-0.18	0.68	-0.29	0.88
Reading *z*-score, spring 2000	-0.45	0.74	-0.20	0.69	-0.25	0.85
Reading *z*-score, spring 2001	-0.41	0.77	-0.22	0.72	-0.19	0.90
Reading *z*-score, spring 2002	-0.48	0.76	-0.14	0.70	-0.33	0.95
Reading *z*-score, spring 2003	-0.47	0.79	-0.25	0.73	-0.22	0.96

Note. Only paired schools with data available in all 5 years are reported. n = 272 (schools).

Table 5 reports HLM results of estimating differences in mathematics and reading between CSR and matched comparison schools from 2000 to 2003, for schools with data available over year and controlling for variables of school background. The results control for *z*-score differences in 1999 and four school background variables that could be related to student performance, including percentage of students who are eligible for free or reduced-price lunch, percentage of minority students, school size, and student–teacher ratio. These four background variables are presented in two formats. One represents the average level of pairs, and the other represents the difference between the pairs. Because the dependent variable is the paired difference, the average of pairs and difference between pairs for the controls was used, rather than the measures for the CSR schools only. The variables of average and year indicators were grand mean centered, and the intercept reflects the average *z*-score differences between CSR and matched comparison schools during a 4-year period. The coefficients of year indicators note the changes in the *z*-score difference from year to year. The coefficients of control variables that measure the average of paired schools estimate how much the changes in *z*-score difference are related to background measures of pairs. The coefficients of control variable that measures the difference of paired schools estimate the effect contributed by the difference in background between pairs.

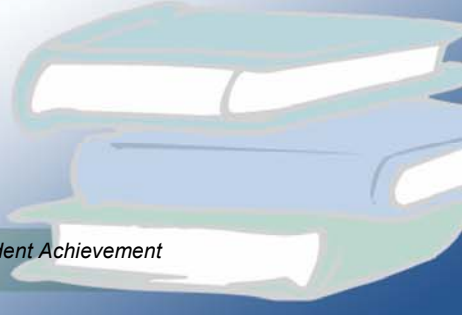
Table 5. Estimated Mathematics and Reading *z*-Score Differences Between CSR and Matched Comparison Schools From 2000 to 2003

Effect	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Intercept	-0.104***	0.039	-0.072*	0.038
Year indicator: 2001	0.063***	0.022	0.028	0.021
Year indicator: 2002	0.034	0.022	-0.024	0.021
Year indicator: 2003	0.039*	0.022	0.017	0.021
z-score difference in 1999	0.526***	0.045	0.537***	0.044
Average percentage of students who are eligible for free or reduced-price lunch	-0.031	0.036	-0.028	0.035
Average percentage of minority students	0.061	0.038	0.034	0.037
Average school size (unit per 1,000 students)	-0.039	0.039	-0.017	0.038
Average student–teacher ratio (unit of 10 students per teacher)	0.021	0.052	0.028	0.050
Average difference in percentage of students who are eligible for free or reduced-price lunch	-0.761**	0.326	-0.413	0.319
Average difference in percentage of minority students	0.355	0.435	-0.219	0.425
Average difference in school size (unit per 1,000 students)	-0.030**	0.014	-0.021	0.014
Average difference in student–teacher ratio (unit of 10 students per teacher)	-0.012	0.015	0.003	0.015

*p-value significant at 0.1 level. **p-value significant at 0.05 level. ***p-value significant at 0.01 level.

The negative intercepts from 2000 to 2003 indicate that on average achievement levels were lower in sampled CSR schools than in matched comparison schools. The positive coefficients of year indicators show that CSR schools were closing the gap, but most of them were not doing so at a significant level. CSR schools made significantly larger achievement gains in mathematics in years 2001 and 2002, for example, but they dropped in reading in 2002. Overall, student achievement in CSR schools showed more growth than matched comparison schools, and the growth in mathematics appeared to be more stable than in reading.

Some control variables show a significant impact in mathematics and reading achievement gains. The prior achievement level, which is measured by the *z*-score difference from 1999, has a large coefficient in predicting such growth. The positive coefficient means that CSR schools achieved higher gains than matched comparison schools in 1999 and maintained that lead. School background variables were also shown to have various impacts on mathematics and reading. The pairs in the study were matched by measures on the percentages of minority students and students who were eligible for free or reduced-price lunch. This process made CSR schools very similar to matched comparison schools. Additionally, because CSR schools and their counterparts served high poverty students, the similarity in the backgrounds left very small variance for explaining the differences in the growth of student achievement.



This is indicated by the insignificant coefficients of controls taking the average. The matching process does not mean, however, that the pairs had identical values for all background measures. The unmatched portions in their backgrounds also had an impact on predicting growth in student achievement. CSR schools that were larger in size than matched comparison schools, for example, were less likely than comparison schools to exhibit gains in mathematics achievement scores.

Because the number of implementation years in sampled CSR schools ranged from 0 to more than 10 years, the growth in student achievement could be underestimated, especially if we considered it too early to expect CSR to affect student achievement in young CSR schools. It is also possible that CSR's effects could have occurred in earlier years and are no longer observable in older CSR schools. For these reasons, the analyses reported in Table 5 were reconducted by replacing the dummy-year variables with variables that measure the age group of CSR schools and the number of implementation years. Two indictors for CSR age were created: One indicates CSR in middle age (i.e., 3–5 years of implementation), and the other indicates CSR in old age (i.e., more than 5 years of implementation). The variable of implementation years indicates the actual number of years of implementation. Thus, the implementation variable becomes a time-varying variable because young-aged CSR schools will always move into the middle-age group after 2 years of implementation. Adding the middle- and old-age dummy variables into the model leaves CSR schools in the young-age group as the reference group.

Overall, the number of implementation years is not a significant predictor of growth in student achievement because growth is mainly explained by the variables of age (Table 6). The coefficient of middle-aged CSR schools is positive and marginally significant in predicting mathematics achievement, and the coefficient of old-aged CSR schools is negative. These numbers show that CSR schools in the middle-aged group experienced faster growth in student achievement gains than CSR schools in the young-aged group (the reference group in the model). In the meantime, the old-aged CSR schools had slower growth compared with young-aged CSR schools. This pattern is consistent with the hypothesized growth shown in Figure 2. The pattern for reading is similar but not significant.

Table 6. Estimated *z*-Score Differences in Mathematics and Reading Between CSR and Matched Comparison Schools, by Implementation Year From 2000 to 2003

Effect	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Intercept	-0.174*	0.070	-0.088	0.069
Number of implementation years	0.008	0.023	0.000	0.023
Middle-age indicator (3–5 years of implementation)	0.047**	0.025	0.011	0.025
Old-age indicator (more than 5 years of implementation)	-0.062	0.039	-0.010	0.039
z-score difference in 1999	0.527*	0.045	0.538*	0.046
Average percentage of students who were eligible for free or reduced-price lunch	-0.022	0.036	-0.033	0.036
Average percentage of minority students	0.059	0.038	0.037	0.038
Average school size (unit per 1,000 students)	-0.046	0.040	-0.023	0.040
Average student–teacher ratio (unit of 10 students per teacher)	0.027	0.052	0.028	0.052
Average difference in percentage of students who were eligible for free or reduced-price lunch	-0.743***	0.328	-0.459	0.328
Average difference in percentage of minority students	0.309	0.435	-0.189	0.436
Average difference in school size (unit per 1,000 students	-0.028**	0.014	-0.015	0.014
Average difference in student–teacher ratio (unit of 10 students per teacher)	-0.016	0.015	-0.002	0.015

*p-value significant at 0.1 level. **p-value significant at 0.05 level. ***p-value significant at 0.01 level.

Research Question 3: Is a High Level of or Positive Change in Implementation Over Time Positively Associated With Growth in Student Achievement?

The best evidence for establishing an empirical inference that CSR implementation improves growth in student achievement would be to demonstrate that CSR schools with improvements in implementation level over time also make larger growth in student achievement. As shown in Tables 2 and 3, sampled CSR schools made significant improvements in their implementation levels on some components. Although the analyses on the growth of *z*-scores in mathematics and reading found that CSR schools, in general, do not always improve more significantly than matched comparison schools, it is very possible that only the CSR schools that make a large improvement in implementation will also experience larger gains than matched comparison schools. If that is the case, then it should be true that such CSR schools will make larger student achievement in mathematics and reading after controlling for implementation level and changes made in implementation level over time. To test this relationship, the same variables that were reported to measure implementation level in Table 5 were added in HLM. In the analysis, only CSR schools with 3–5 years of implementation (i.e., middle-aged schools) in 2002 were kept, because up until the third year of implementation, growth in student achievement lagged behind growth in CSR

implementation. In addition, substantial levels of implementation of the model were not reached until the third year. Thus, attempting to link academic performance to implementation prior to the third year would be pointless. Table 5 does show, however, that CSR schools with 3–5 years of implementation did perform better than other CSR schools relative to their comparison schools. Keeping the middle-aged CSR schools in 2002 only ensures that most CSR schools started implementation in at least 2000.

As reported in Tables 2 and 3, data from principals and teachers were collected for measuring implementation of several components of CSR. These measures were also used as predictors to investigate a possible relationship between teacher- and principal-reported implementation level growth and student achievement. Because the latest year of the achievement data available for this study came from 2003 and the last implementation scores were from 2004, it becomes problematic if the changes in implementation scores from 2002 to 2004 are used to predict student achievement before 2004. In spite of this challenge, implementation data from 2004 can still be used to study the association between implementation score and growth in student achievement if it is used as an indicator of implementation status only for grouping CSR schools. That is, 2002 implementation scores were used as the measures of implementation level, and the differences between 2004 and 2002 implementation scores were used to indicate the status of implementation from 2002 to 2004. This indicator informs whether a CSR school continuously improved its implementation level after 2002. The indicator can also be used to examine how the implementation level status relates to growth in student achievement in the previous years.

Table 7 reports the results of using implementation measures on the component of student grouping to predict *z*-score differences in mathematics and reading between CSR schools and matched comparison schools. The analysis model used for this table is similar to the preceding one but it included three new predictors: the implementation level in 2002 and implementation status from both 2002 and 2004.[2] The implementation year variable is assumed to have random effect on growth in student achievement, which means the average *z*-score difference changed by implementation year. The coefficient of implementation year in predicting mathematics is positive and significant, indicating that CSR schools with longer implementation years had larger achievement gains in mathematics. The coefficient of implementation level in 2002 is also positive and significant, indicating that CSR schools with higher implementation scores in student grouping in 2002 had larger achievement gains in mathematics than other CSR schools. The positive and significant coefficient of the dummy indicator of implementation status indicates that CSR schools that continually improved implementation level from 2002 to 2004 also experienced larger gains in achievement in mathematics than other CSR schools from 2000 to 2003.

The results of the analysis model also yielded two interactions. The first one was the interaction between number of implementation years and implementation status after 2002. It explored whether CSR schools with different implementation statuses after 2002 experienced similar growth rates in student achievement when they were of similar CSR age. The second one was the interaction between implementation level in 2002 and status after 2002. It investigated whether the association between the continuity of implementation after 2002 and the growth in student achievement depends on the implementation level in 2002. The results of the interactions revealed that growth rate was similar among CSR schools of similar

[2] There are two indicators for the status of implementation level from 2002 to 2004. One indicates whether 2004's level is higher than 2002's level. Another one indicates the status is unknown because of the unknown implementation level in either 2002 or 2004.

age but had different implementation statuses from 2002 to 2004. The negative and significant coefficient of the interaction between implementation level in 2002 and status from 2002 to 2004 revealed that among CSR schools that improved the same amount of implementation level from 2002 to 2004, the schools that started at a lower level in 2002 made larger growth in student achievement than those that started at a higher level. In other words, CSR schools that had more room to improve in implementation level experienced higher gains in student achievement. Once they improved the implementation level of their design, as discussed previously, CSR schools with high implementation levels showed larger gains in student achievement. This indicates that CSR schools that reached a high implementation level in 2002 experienced large gains in student achievement from 2000 to 2003. However, the rate of improvement in implementation actually began to decelerate (once it peaked in 2002) compared with previous years. This pattern was consistent after 5 years of implementation, as shown in Figure 2.

The association between implementation score on student grouping and growth in reading is not as apparent when predicting mathematics. When CSR has an effect on improving student achievement, the extent of the effect can vary by CSR component. For this reason, implementation component measures may not always reveal significant effects on improvements in academic success. With this in mind, the general expectation is still that the direction of the effect is expected to be positive. In other words, a high level of implementation or improvement in implementation should be traceable to large gains in student achievement. The results of this analysis reveal that high implementation of the specific core components of inclusion of special education students and parent–community involvement will mostly yield a similar pattern in predicting high mathematics and reading gains but not all coefficients reach significant levels (Appendix B, Table B1).

Table 7. Effect of Principal-Reported Implementation of Student Grouping on Growth in Mathematics and Reading From 2000 to 2003

	Mathematics		Reading	
Effect	Estimate	SE	Estimate	SE
Intercept	-1.784***	0.538	-0.965	0.632
Number of implementation years	0.112***	0.042	0.007	0.040
Implementation level in 2002: grouping students	2.046***	0.764	1.169	0.903
Implementation status indicator: level in 2004 higher than in 2002	2.257***	0.751	1.917**	0.881
Implementation status indicator: unknown level in 2004 or 2002	0.402	0.642	0.196	0.752
Interaction terms between number of implementation years and indicator of implementation status	-0.096	0.072	0.048	0.068
Interaction terms between 2002 implementation level and status from 2002 to 2004	-3.003**	1.213	-3.201**	1.437

*p-value significant at 0.1 level. **p-value significant at 0.05 level, and ***p-value significant at 0.01 level.

Table 7. Effect of Principal-Reported Implementation of Student Grouping on Growth in Mathematics and Reading From 2000 to 2003 (continued)

Effect	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-0.788	0.979	0.004	1.148
z-score difference in 1999	0.379***	0.069	0.222***	0.081
Average percentage of students who were eligible for free or reduced-price lunch	0.036	0.055	0.032	0.065
Average percentage of minority students	0.165***	0.054	0.203***	0.064
Average school size (unit per 1,000 students)	-0.018	0.049	-0.062	0.058
Average student–teacher ratio (unit of 10 students per teacher)	0.005	0.073	0.053	0.085
Average difference in percentage of students who were eligible for free or reduced-price lunch	-1.140**	0.512	-0.943	0.598
Average difference in percentage of minority students	1.403**	0.684	-0.074	0.803
Average difference in school size (unit per 1,000 students)	-0.018	0.021	-0.013	0.025
Average difference in student–teacher ratio (unit of 10 students per teacher)	-0.058**	0.027	-0.002	0.032

*p-value significant at 0.1 level. **p-value significant at 0.05 level. ***p-value significant at 0.01 level.

The relationship between teacher-reported implementation level and growth in student achievement is similar to that of principal-reported implementation level. Professional development is a core component in CSR implementation that is directly related to instruction and curriculum. Table 8 reports the effect of teacher-reported implementation of professional development on predicting gains in mathematics and reading achievement. The results showed that the effect of implementation level was positive and significant on mathematics, demonstrating that CSR schools with higher levels of implementation on this component (professional development) in 2002 experienced larger achievement gains in mathematics than CSR schools with lower levels of implementation, compared with matched comparison schools. In addition, schools with improved implementation levels from 2002 to 2004 also experienced greater reading gains than schools with smaller levels of implementation improvement from 2002 to 2004, relative to matched comparison schools. Finally, with regard to professional development, although most of the remaining teacher-reported implementation measures positively predicted mathematics and reading gains, most of these coefficients were not significant (Appendix B, Table B2).

Table 8. Effect of Teacher-Reported Implementation Measures of Professional Development Type on *z*-Score Differences in Mathematics and Reading From 2000 to 2003

Effect	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Intercept	-1.425***	0.507	-0.529	0.498
Number of implementation years	0.093**	0.044	0.030	0.043
Implementation level in 2002: type of professional development	1.537**	0.652	0.539	0.639
Implementation status indicator: level in 2004 higher than in 2002	1.337**	0.721	0.432	0.708
Implementation status indicator: unknown level in 2004 or 2002	2.135	5.329	0.990	4.427
Interaction terms between number of implementation year and indicator of implementation status	-0.028	0.067	-0.047	0.064
Interaction terms between 2002 implementation level and status from 2002 to 2004	-1.904*	0.964	-0.397	0.952
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-3.748	7.255	-0.731	6.146
z-score difference in 1999	0.508***	0.072	0.491***	0.073
Average percentage students who were eligible for free or reduced-price lunch	-0.002	0.058	-0.045	0.058
Average percentage of minority students	0.134**	0.057	0.158***	0.057
Average school size (unit per 1,000 students)	-0.039	0.053	-0.028	0.052
Average student–teacher ratio (unit of 10 students per teacher)	0.026	0.080	0.056	0.080
Average difference in percentage of students who were eligible for free or reduced-price lunch	-0.628	0.517	-0.308	0.527
Average difference in percentage of minority students	0.587	0.687	-0.952	0.687
Average difference in school size (unit per 1,000 students)	-0.020	0.022	-0.013	0.022
Average difference in student–teacher ratio (unit of 10 students per teacher)	-0.041	0.030	-0.036	0.030

*p-value significant at 0.01 level. **p-value significant at 0.05 level. ***p-value significant at 0.1 level.

Research Question 4: How Are CSR Implementation and Growth in Student Achievement Associated Across CSR Models?

As mentioned earlier, sampled CSR schools in this study were implementing several CSR models in 2002, including CSR Model A, CSR Model C, CSR Model F, CSR Model E, and CSR Model D. Of interest is whether the association of implementation and student achievement varies by specific CSR designation. To examine this possible variation, four dummy variables were added to compare schools across CSR models. These dummy variables represent CSR Model B, CSR Model A, CSR Model C, and

other CSR models. Because the number of schools in the sample implementing CSR Model E and CSR Model D was small, these schools were combined into a new exclusive category labeled "other CSR." CSR Model F schools were the largest group in the sample, so they were used as the reference group.

Table 9 displays the relationship between teacher-reported average implementation level and the *z*-score differences in mathematics and reading. The effect of average implementation level was better at predicting reading than mathematics. This model specifically focuses on the coefficients of dummy variables and their interaction terms with the implementation level. The negative coefficient of CSR Model B, CSR Model A, and CSR Model C in predicting reading shows that schools implementing these three models experienced smaller gains in reading achievement scores than schools that were implementing CSR Model F. The positive and significant coefficients of interaction terms between dummy variables of these three models and implementation level indicate that implementation in CSR Model B, CSR Model A, and CSR Model C schools seem not to be as high as CSR Model F schools. However, the effect of implementation level in CSR Model B, CSR Model A, and CSR Model C schools was higher than CSR Model F schools. When CSR Model B, CSR Model A, and CSR Model C schools improved to the same unit of implementation level, the coefficient could predict larger achievement gains in the schools using the respective models than in CSR Model F schools. In other words, the relationship between implementation level and growth in reading is stronger in CSR Model B, CSR Model A, and CSR Model C schools than in CSR Model F schools. This finding is evident from the variation in the association of CSR implementation average level and student achievement across CSR models.

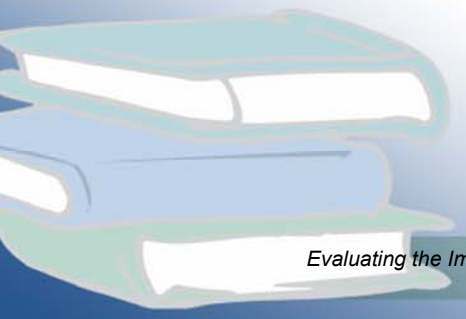


Table 9. Effect of Teacher-Reported Average Implementation Level in 2002 on *z*-Score Differences in Mathematics and Reading From 2000 to 2003

Effect	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Intercept	-0.605	1.675	2.071	1.804
Number of implementation years	0.086**	0.033	0.010	0.032
Average implementation level in 2002	0.349	2.178	-2.844	2.346
z-score difference in 1999	0.463***	0.074	0.307***	0.079
CSR Model B	-12.941	11.358	-22.586*	12.301
CSR Model A	1.076	2.884	-7.287**	3.097
CSR Model C	-2.327	5.288	-10.369*	5.644
Other CSR models	0.516	1.852	-0.955	1.996
Interaction: CSR Model B with implementation level	18.539	16.154	32.177*	17.491
Interaction: CSR Model A with implementation level	-1.277	4.064	10.399**	4.362
Interaction: CSR Model C with implementation level	2.949	7.380	14.744*	7.875
Interaction: other CSR models with implementation level	-0.735	2.437	1.439	2.628
Average percentage of students who were eligible for free or reduced-price lunch	-0.029	0.064	0.016	0.070
Average percentage of minority students	0.134**	0.063	0.159**	0.068
Average school size (unit per 1,000 students)	-0.009	0.059	-0.052	0.064
Average student–teacher ratio (unit of 10 students per teacher)	-0.009	0.089	0.081	0.097
Average difference in percentage of students who were eligible for free or reduced-price lunch	-0.616	0.562	-0.656	0.607
Average difference in percentage of minority students	0.207	0.711	-1.177	0.767
Average difference in school size (unit per 1,000 students)	-0.012	0.024	-0.023	0.026
Average difference in student–teacher ratio (unit of 10 students per teacher)	-0.040	0.035	0.062*	0.036

*p-value significant at 0.05 level. **p-value significant at 0.1 level. ***p-value significant at 0.01 level.

Conclusion

The results of this study confirm that implementing CSR is a complex procedure and its effect on student achievement is still uncertain in some cases. The uncertainty makes it difficult to deliver a simple and direct conclusion on the association between CSR implementation and student achievement. Without measures of implementation level or implementation length, the CSR effect on student achievement may be indeterminable. This is understandable because of the reality of CSR implementation. That is, CSR schools often implement components at various levels and paces, not just across schools and CSR models but also in different stages (i.e., young-, middle-, and old-age) of implementation. The sampled CSR schools in this study had various starting levels and increased CSR implementation at different rates and paces. Some schools significantly improved implementation in some components from 2002 to 2004. Other schools, however, did not make progress or had even lower levels of implementation gains in later years. On average though, CSR schools improved the fidelity of their implementation over time. Among schools, the large variation of growth in CSR implementation from 2002 to 2004 indicates an uneven pace in implementation.

Duration of implementation of a specific CSR model does not guarantee an effect on student achievement, as measured by the growth in *z*-score made by CSR schools relative to matched comparison schools. Overall results reveal that CSR schools did not experience larger achievement gains in either mathematics or reading compared with matched comparison schools. However, this is not for lack of evidence of a CSR effect. Results of various analyses conducted in this study reveal that schools implementing certain components of CSR for 3–5 years were definitely likely to experience significant academic gains. Schools that implemented CSR at high levels were more likely to experience large academic gains than matched comparison schools and CSR schools that implemented CSR at lower levels. Some CSR schools with lower implementation levels also made larger growth in student achievement in mathematics or reading but only when they improved implementation level over time. Moreover, the impact of CSR implementation on student achievement varied by CSR model and the subjects being tested (i.e., mathematics and reading). These findings indicate that the impact of CSR implementation on student achievement is conditional of implementation level, components, number of implementation years, and specific CSR model. As such, studies that do not take into account all conditions will result in mixed findings regarding the relationship between CSR implementation and student achievement.

Discussion

For schools in this study, CSR was found to have an effect on student achievement when CSR design implementation level progressively improved to a high level alongside implementation length. The student achievement gains were more likely to be identified if the implementation situation of CSR schools was known in advance and samples of CSR schools with frequent and well-implemented components were selected based on this knowledge. This may have been the case for studies conducted by CSR developers that resulted in positive findings. Conversely, studies using samples of CSR schools with a broad range of implementation levels and CSR models would be expected to less likely detect the growth in student achievement, unless schools within the study are grouped by implementation measures and CSR models. This may have been the case for studies conducted by independent researchers, for which results were frequently misled. From a methodological perspective, both approaches share the same assumptions—that implementation of CSR is important—but reach different conclusions about CSR

effects on student achievement. When one considers only those CSR schools with successful implementation of most components, CSR does demonstrate an effect on student achievement. However, as has been revealed in this study, not all CSR schools make progress in implementation at the same pace, and even if they do, the numbers of implementation years is not always uniform. Given these considerations, issuing a blanket statement about the effects of CSR on academic achievement would be erroneously misleading. Rather, researchers are encouraged to focus on middle-aged CSR schools when attempting to explain the relationship between implementation and growth in student achievement. This approach should not be viewed as a way of incorporating self-selection bias into the analysis. Rather, this approach better enables researchers to have a more precise estimated measurement of the effects of CSR, especially given what has been learned about the realities of the relationship between implementation level and academic achievement.

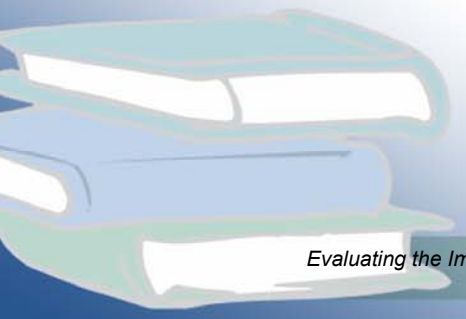
Conducting studies on the effect of CSR presents several challenges. The unperfected solution to these challenges brings limitations to such studies. The first challenge is attempting to create a psychometrically valid, reliable, and measurable definition of level of implementation. The pace of implementation varies by school and CSR model, and the number of years and level of implementation can also vary across CSR schools. Therefore, it would be important to distinguish between level and length of implementation, because the two do not measure the same thing. Currently, there is no psychometrically sound way to conceptualize and measure implementation. Thus, the validity of implementation measures currently in use is often being challenged by the striking similarities between instructional practices of CSR schools and comparison schools (in this study, non-CSR schools). In some cases, the measure of implementation from CSR schools could be lower than matched comparison schools (Kurki et al., 2005). The similarity may be explained in three possible ways:

- First, many CSR schools might implement some but not all basic components of their CSR designs, and they may not reach full implementation (Berends, 2000; Bodilly, 1998).
- Second, CSR schools have to abide by other priorities such as district policies (Bodilly, 1996; Smith et al., 1997).
- Third, principals and teachers in comparison schools could have possibly learned some ideas endorsed by CSR models from noncomprehensive school reforms (e.g., 90-minute reading blocks). For this reason, comparison schools may have being implementing CSR-like activities, for which may not differ remarkably. If CSR and non-CSR schools change in similar ways, then it becomes difficult to expect CSR schools to perform significantly different than comparison schools.

The second challenge is the complexity of establishing and exploring the relationship between implementation and growth in student achievement. This complexity is mainly reflected in two scenarios. The first scenario is the uneven rate of implementation during various time periods of implementation. As displayed in Figure 2, the rate of student achievement gain varies with implementation length. In the second scenario, the mobility in the relationship between schools and CSR developers also affects implementation. During the time of data collection for this study, a few sampled CSR schools ended their connection with the developer each year; thus, the implementation level steadily declined. No single factor can fully predict such abandonment in implementation or adoption (Taylor, 2005). Rather, there are several reasons why CSR schools end implementation: lack of funds, launch of a new program, change in school personnel, or reaching a high level of implementation. In general, one could expect less growth in

student achievement in schools in which implementation has ended because they may have failed to complete successful implementation. It becomes problematic when we expect schools that have dropped the program because they have attained their highest possible level of implementation, to experience academic achievement gains that are directly connected to increasingly high levels of program implementation. One should expect the implementation level in these schools to begin to drop, although the pattern of student achievement may remain the same for a while. In fact, the relationship between improvement in implementation level and growth in student achievement could become negative, and thus it becomes irrelevant to use the level or change of implementation to predict growth in future years. Therefore, predicting growth in student achievement should be based on implementation from previous years.

How to exactly measure the comprehensiveness of implementation is another challenge. The difficulty lies in integrating separate measures of CSR implementation components into a comprehensive measure on the dynamic process of model implementation. This was explored in this study by examining whether implementation scores improved from 2002 to 2004. A better measure should consider not only the number of components being improved but also the relevant order and level of improvement. This combination of factors would help to truly capture the infrastructure of good implementation. The key to a successful implementation may be in setting up a mechanism that makes the components work together in schools, rather than just implementing individual components separately. In other words, the magnitude and the extent to which model components are being implemented may not be the most crucial part of a successful implementation. Instead, a mechanism that can integrate the successive change made by each individual component to eventually create a comprehensive combination that works for the school may be more significant. This means that in some cases, seemingly trivial changes that an individual school may make to individual components could eventually have large effects on the cohesiveness and the capacity to successfully implement the model in that school or setting. Studies should also focus on components that measure depth, sustainability, spread, and shift in reform ownership rather than the number of existing model components (Coburn, 2003). Building such a mechanism in CSR schools may be the most critical part of understanding the key to successfully implementing CSR models, given that implementation continues to be one of the greatest challenges faced when attempting to measure the effects of CSR.



References

Berends, M. (2000). Teacher-reported effects of New American Schools design: Exploring relationships to teacher background and school context. *Educational Evaluation and Policy Analysis, 22*(1), 65–82.

Berends, M., Bodilly, S., & Kirby, S. N. (2002). *Facing the challenges of whole-school reform: New American Schools after a decade.* Santa Monica, CA: RAND.

Berends, M., Kirby, S.N., Naftel, S., & McKelvey, C. (2001). Implementation and performance in New American Schools: Three years into scale-up. Santa Monica, CA: RAND.

Bloom, S. H., Rock, J., Ham, S., Melton, L., & O'Brien, J. (2001). *Evaluating the Accelerated Schools approach: A look at early implementation and impacts on student achievement in eight elementary schools.* New York, NY: MDRC.

Bodilly, S. J. (1996). *Lessons from New American Schools Development Corporation's demonstration phase*. Santa Monica, CA: RAND.

Bodilly, S. J. (1998). *Lessons from New American Schools' scale-up phase: Prospects for bringing designs to multiple schools*. Santa Monica, CA: RAND.

Borman, G. D., & Hewes, G. M. (2002). Long-term effects and cost effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24*(4), 243–266.

Borman, G. D., Hewes, G., Overman, L., & Brown, S. (2003). Comprehensive school reform and student achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125–230.

Borman, G.D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (in press). Success for All: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis.*

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). New York, NY: Greenwood Press.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94,* S95–S120.

Comer, J. P. (1984). Home-school relationships as they affect the academic success of children. *Education and Urban Society, 16,* 323–337.

Comer, J. P. (1988). Educating poor minority children. *Scientific American, 259*(5), 42–48.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Cook, T. D., Habbib, F., Phillips, M., Settersten, R. A., Shagle, S. G., & Degirmencioglu, S. M. (1999). Comer's School Development Program in Prince Georges' County, Maryland: A theory-based evaluation. *American Education Research Journal, 36*(3), 543–597.

Cook, T. D., Hunt, H. D., & Murphy, R. F. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal,* 37(2), 535–597.

Coburn, C. (2003). Rethinking scale: Moving beyond numbers to deep and lasting changes. *Educational Researcher, 32*(6), 3–12.

Cuban, L. (1984). Transforming the frog into a prince: Effective schools research, policy, and practice at the district level. *Harvard Educational Review, 54*(2), 129–151.

Cuban, L. (1990). Reforming again, again, and again. *Educational Researcher, 9*(1), 3–13.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5–31.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research, 72*(3), 433–479.

Desimone, L. (2000). *Making comprehensive school reform work: Lessons from implementation research.* New York, NY: ERIC Clearinghouse on Urban Education, Teacher College, Columbia University.

Epstein, J. L., & Hollifield, J. H. (1996). Title I and school-family-community partnerships: Using research to realize the potential. *Journal of Education for Students Placed at Risk, 1*(3), 263–278.

Dewey, J. (1966). *Democracy and education.* New York, NY: The Free Press.

Fast, E. F., Aladjem, D., Shive, J., Herman, R., & Carter, K., (2001, November). *A review of the design of eight comprehensive school reform models.* Paper presented at the Annual Meeting of the American Evaluation Association, St. Louis, MO.

Glennan, T. K., Jr. (1998). New American schools after six years. Santa Monica, CA: RAND.

Garet, M., Zhang, Y., & O'Day, J. (2003, April). *Immediate Intervention/Underperforming Schools Program of the California Public Schools Accountability Act of 1999: A preliminary analysis of achievement outcomes.* Paper presented at 2003 annual meeting of the American Educational Research Association, Chicago, IL.

Herman, R., Aladjem, D., McMahon, P., Mulligan, I., Smith-O'Malley, A., Quinones, S., Reeve, A., & Woodruff, D. (1999), *An educator's guide to schoolwide reform*. Washington, DC: American Institutes for Research.

Herman, R., Carl, B., Lampron, S., Sussman, A., Berger, A., & Innes, F. (2000). *What we know about comprehensive school reform models*. Washington, DC: American Institutes for Research.

Kurki, A., Aladjem, D., & Carter, K.R. (2005). *Implementation: Measuring and explaining the fidelity of CSR implementation.* Paper presented at the 2005 annual meeting of the American Educational Research Association, Montreal, Canada.

Levin, H. M. (1989). Financing the education of at-risk students. *Educational Evaluation and Policy Analysis, 11*(1), 47–60.

Light, J. L., & Pillemer, D. B. (1984). *Summing up the science of reviewing research*. Cambridge, MA: Harvard University Press.

Muncey, D. E., & McQuillan, P. J. (1996). *Reform and resistance in schools and classrooms: An ethnographic view of the Coalition of Essential Schools*. New Haven, CT: Yale University Press.

Murphy, J., & Hallinger, P. (Eds.). (1993). *Restructuring schooling: Learning from ongoing efforts*. Thousand Oaks, CA: Corwin Press.

Newmann, B., Smith, B. A., Allensworth, E., & Bryk, A. S. (2002). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23*(4), 297–321.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Department of Education.

Orfield, G. (1999). Strengthening Title I: Designing a policy based on evidence. In G. Orfield & E. H. DeBray (Eds.), *Hard work for good schools: Facts, not fads, in Title I reform* (pp. 1–20). Cambridge. MA: Harvard University, Civil Rights Project.

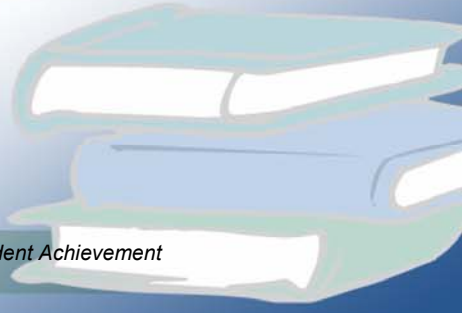
Peak, L. (1996). *Pursuing Excellence: A study of U.S. eighth-grade mathematics and science teaching, learning, curriculum, and achievement in international context.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Pong, S. (1998).The school compositional effect of single parenthood on 10th-grade achievement. *Sociology of Education, 71*(1), 24–43.

Porter, A. C. (1994). National standards and school improvement in the 1990s: Issues and promise. *American Journal of Education, 102*, 421–449.

Ross, S. M., & Lowther, D. L. (2003). Impacts of Co-nect school reform design on classroom instruction, school climate, and student achievement in inner-city schools. *Journal of Education for Students Placed at Risk, 8*(2), 215–246.

Ross, S. M., Troutman, A., Horgan, D., Maxwell, S., Laitinen, R., & Lowther, D. (1997). The success of schools in implementing eight restructuring designs: A synthesis of first-year evaluation outcomes. *School Effectiveness and School Improvement, 8*(1), 95–124.



Ross, S. M., Stringfield, S. S., William, L. S., & Wright, S. P. (2003). Inside systemic elementary school reform: teacher effects and teacher mobility. *School Effectiveness & School Improvement, 14*(1), 73–110.

Slavin, R. (2001). The facts about comprehensive school reform. *Educational Leadership, 51*(1), 84–85.

Slavin, R. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21.

Slavin, R. E., & Madden, N. A. (2000). Research on achievement outcomes of Success for All: A summary and response to critics. *Phi Delta Kappan, 82*(1), 38–40.

Slavin, R. E., & Fashola, O. S. (1998). *Show me the evidence! Proven and promising programs for America's schools.* Thousand Oaks, CA: Corwin Press.

Slavin, R.E., Madden, N. A., Dolan, L. J., Wasik, B. A., Smith, L. M., Ross, S. M., & Dianda, M. (1996). Success for All: A summary of research. *Journal of Education for Students Placed At Risk, 1*, 41–76.

Slaton, D., Atwood, V., Shake, M., & Hales, R. (1997). Experienced teachers' reactions to mandated reform and primary school programs. *Journal of Research in Childhood Education, 12*, 5–15.

Smith, J., Maxwell, S., Lowther, D., Hacker, D., Bol, L., & Nunery, J. (1997). Activities in schools and programs experiencing the most, and least, early implementation successes. *School Effectiveness and School Improvement, 8*(1), 125–150.

Smith, S., & O'Day, J. (1991). System school reform. In R. Elmore & S. Fuhrman (Eds.), *The governance of curriculum: 1990 yearbook of the politics of education association* (pp. 233–267). London, UK: Falmer Press.

Southwest Educational Development Laboratory. (n.d.). *CSR awards database.* Retrieved March 22, 2005, from http://www.sedl.org/csr/summary.html

Taylor, J. E. (2005). *Sustainability: Examining the survival of schools' comprehensive school reform efforts.* Paper presented at the 2005 annual meeting of the American Educational Research Association, Montreal, Canada.

Teddlie, C., & Reynolds, D. (2000). The international handbook of school effectiveness research. London, UK: Falmer Press.

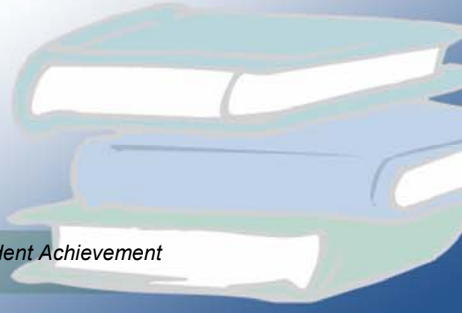
U.S. Department of Education. (2002a, updated August). *Guidance on the Comprehensive School Reform Program.* Retrieved March 21, 2005, from http://www.ed.gov/programs/compreform/guidance/index.html

U.S. Department of Education. (2002b). *Public Law 107–110, The No Child Left Behind Act of 2001.* Retrieved March 21, 2005, from http://www.ed.gov/policy/elsec/leg/esea02/index.html

U.S. Department of Education, Office of the Under Secretary, Planning and Evaluation Service, Elementary and Secondary Division. (2000). *Early implementation of Comprehensive Reform Demonstration (CSRD) Program.* Washington, DC: Author.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1997). *What do we know: Widely implemented school improvement programs.* Philadelphia, PA: Laboratory for Student Success, The Mid-Atlantic Regional Educational Laboratory at Temple University, Center for Research in Human Development and Education.

Wasik, B., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly, 28,* 178–200.



Appendix A

Table A1. Measures of Student Achievement

State	Test Type	Scoring Type	Grade
CA	SAT-9 until 2003; CAT6 in 2002–2003	Mathematics average scaled score	2–8
		Reading average scaled score	2–8
DC	SAT-9	Mathematics NCE Score	2–8
		Reading NCE Score	2–8
FL	FCAT	Mathematics % at or above level 2	5 and 8
		Reading % at or above level 2	4 and 8
IL	ISAT	Mathematics [ISAT] % meeting standards & up	3, 5, and 8
		Reading [ISAT] % meeting standards & up	3, 5, and 8
	IGAP	Mathematics [IGAP] average scaled score	3, 6, and 8
		Reading [IGAP] average scaled score	3, 6, and 8
KY	CTBS	Mathematics average scaled score	5 and 8
		Reading average scaled score	4 and 7
MD	MSPAP 1999–2002; MSA 2002–2004	Mathematics % satisfactory/proficient +	3, 5, and 8
		Reading % satisfactory/proficient +	3, 5, and 8
MA	MCAS	Mathematics % proficient and higher	4 and 8
		ELA % proficient and higher	4 and 8
MI	MAEP	Mathematics % moderate+/% basic+	4 and 7
		Reading % moderate+/% basic+	4 and 7
MO	MAP	Mathematics % at or above progressing (level 2)	4 and 8
		Reading % at or above satisfactory	3 and 7
NJ	GEPA (grade 8); ESPA (grade 4) in 2002; ASK4 in 2003	Mathematics % proficient and above	4 and 8
		Reading % proficient and above	4 and 8
NV	TerraNova in 1999–2002; ITBS in 2003	Mathematics percentile rank	4 and 8
		Reading percentile rank	4 and 8

Table A1. Measures of Student Achievement (continued)

State	Test Type	Scoring Type	Grade
OH	PSA	Mathematics % scoring at or above proficient on primary state assessment (PSA) or mathematics % passing for all students	4 and 6
		Reading % scoring at or above proficient or reading % passing for all students	4 and 6
PA	PSSA	Mathematics average scale score	5 and 8
		Reading average scale score	5 and 8
TX	TAAS from 1999–2002; TAKS in 2003	Mathematics [Eng] % passing for all students	3–8
		Reading [Eng] % passing	3–8
UT	SAT-9	Mathematics percentile rank	5 and 8
		Reading percentile rank	5 and 8
WA	WASL	Mathematics percent at or above level 2	4 and 7
		Reading percent at or above level 2	4 and 7

Appendix B

Description of growth model for modeling implementation effect on growth in student achievement:

Time-Level Model

$$\eta_{ij} = \beta_{0j} + \beta_{1j} Score99_{ij} + \beta_{2j} Implemenation\_year_{ij} + \beta_{3j} Year03_{ij} + \beta_{4j} SchoolSize_{ij} + \beta_{5j} Pct\_Minority_{ij} + \beta_{6j} Pct\_Freelunch_{ij} + \beta_{7j} Student\_teacher\_ratio_{ij} + \beta_{8j} SchoolSize\_Dif_{ij} + \beta_{9j} Pct\_Minority\_Dif_{ij} + \beta_{10j} Pct\_Freelunch\_Dif_{ji} + \beta_{11j} Student\_teacher\_ratio\_Dif_{ij} + \varepsilon_{ij}$$

where:

η_{ij} is the difference in *z*-score between CSR school *i* and its paired school in year *j;* and

$Score99_{iij}$ is the difference in *z*-score between school *j* and its pair in 1999.

The following control variables are grand mean centered:

$Implementation\_year_{ij}$ is the number of implementation years for school *i* in year *j*.

$SchoolSize_{ij}$ is the average number of students enrolled in school *j*.

$Pct\_Minority_{ij}$ is the average percentage of minority students enrolled in school *j*.

$Pct\_Freelunch_{ij}$ is the average percentage of students who were eligible for free lunch in school *j*.

$Student\_teacher\_ratio_{ij}$ is the average student–teacher ratio in school *j*.

The following control variables are not centered:

SchoolSize_Dif$_{ij}$ is the difference in the average number of students enrolled between paired schools.

Pct_Minority_Dif$_{ij}$ is the difference in the average percentage of minority students between paired schools.

Pct_Freelunch_Dif$_{ij}$ is the difference in the average percentage of students eligible for free lunch between paired schools.

Student_teacher_ratio_Dif$_{ij}$ is the difference in the average student–teacher ratio between paired schools.

ε_{ij} is the probability of CSR schools improving more than matched-pair comparison schools.

School-Level Model

$$\beta_{00} = \gamma_{00} + \gamma_{01} * implementation\_year_{0j} + \gamma_{02} * implemenation\_status_{0j} + \gamma_{03} * interaction + \nu_{0j}$$

$$\beta_{10} = \gamma_{10} + \gamma_{11} * implemenation\_status_{1j} + \nu_{1j}$$

$$\beta_{20} = \gamma_{20}$$

...

$$\beta_{11} = \gamma_{11}$$

where:

γ_{00} is the average intercept;

γ_{i0} i = 1 to 12 are average slope;

Implemantion_year is the number of implementation years for school *j;*

Implemantion_status is the indicator of implementation in 2004 compared with 2002 for school *j;*

Interaction is the interaction between implementation year and status; and

ν_{0j} and ν_{1j} is a random error term representing unmeasured factors related to the intercept of the growth curve for school *j*.

Table B1. Effect of Principal-Reported Implementation Level on Paired-Difference in Mathematics and Reading Gains From 1999 to 2003

Effect of implementation	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Shared decision making among principals and teachers				
Number of implementation years	0.092*	0.049	0.022	0.042
Implementation level in 2002: shared decision making among principals and teachers	-2.098	1.674	-1.949	1.763
Implementation status indicator: level in 2004 higher than in 2002	-2.731	1.669	-1.906	1.755
Implementation status indicator: unknown level in 2004 or in 2002	-2.246	1.837	-3.263*	1.935
Interaction terms between number of implementation years and indicator of implementation status	-0.012	0.072	0.043	0.063
Interaction terms between 2002 implementation level and status from 2002 to 2004	2.975	1.849	2.131	1.949
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	1.976	2.043	3.611*	2.153
School organization and governance				
Number of implementation years	0.098**	0.039	0.009	0.036
Implementation level in 2002: school organization and governance	-0.244	0.588	0.073	0.681
Implementation status indicator: level in 2004 higher than in 2002	0.105	0.643	-0.803	0.729
Implementation status indicator: unknown level in 2004 or in 2002	-1.205*	0.692	-0.703	0.800
Interaction terms between number of implementation years and indicator of implementation status	-0.068	0.085	0.109	0.077
Interaction terms between 2002 implementation level and status from 2002 to 2004	0.180	0.960	1.311	1.111
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	1.329	0.931	1.050	1.075

Table B1. Effect of Principal-Reported Implementation Level on Paired-Difference in Mathematics and Reading Gains From 1999 to 2003 (continued)

Effect of implementation	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Inclusion of special education students				
Number of implementation years	0.172***	0.057	0.066	0.053
Implementation level in 2002: inclusion of special education students	0.282	1.034	-0.459	1.182
Implementation status indicator: level in 2004 higher than in 2002	0.988	0.887	0.548	1.009
Implementation status indicator: unknown level in 2004 or in 2002	-0.091	0.888	0.497	1.014
Interaction terms between number of implementation years and indicator of implementation status	-0.141**	0.071	-0.054	0.067
Interaction terms between 2002 implementation level and status from 2002 to 2004	-0.732	1.156	-0.398	1.323
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-0.079	1.194	-0.588	1.363
Parent–community involvement				
Number of implementation years	0.115**	0.041	0.083**	0.037
Implementation level in 2002: parent–community involvement	0.640	0.528	0.037	0.661
Implementation status indicator: level in 2004 higher than in 2002	1.318*	0.706	0.147	0.871
Implementation status indicator: unknown level in 2004 or in 2002	0.918	0.584	0.743	0.734
Interaction terms between number of implementation years and indicator of implementation status	-0.120	0.076	-0.186***	0.069
Interaction terms between 2002 implementation level and status from 2002 to 2004	-1.562*	0.824	0.256	1.032
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-1.579**	0.690	-0.952	0.867

Note. This is the same model as reported in Table 7, and only coefficients of implementation variables are reported.

*p-value significant at 0.1 level. **p-value significant at 0.05 level. ***p-value significant at 0.01 level.

Table B2. Effect of Teacher-Reported Implementation Level on Paired-Difference in Mathematics and Reading Gains From 1999 to 2003

Effect of implementation	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Use of technology in instruction				
Number of implementation years	0.007	0.048	0.011	0.046
Implementation level in 2002: use of technology in instruction	0.562	0.792	-0.034	0.766
Implementation status indicator: level in 2004 higher than in 2002	0.025	0.576	0.180	0.560
Implementation status indicator: unknown level in 2004 or in 2002	-0.573	3.505	1.280	2.899
Interaction terms between number of implementation years and indicator of implementation status	0.137**	0.069	0.008	0.065
Interaction terms between 2002 implementation level and status from 2002 to 2004	-0.920	0.949	-0.559	0.926
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-0.066	10.209	-2.294	8.683
Inclusion of special education students				
Number of implementation years	0.139**	0.055	0.043	0.049
Implementation level in 2002: inclusion of special education students	1.741	1.424	1.470	1.408
Implementation status indicator: level in 2004 higher than in 2002	2.077*	1.238	1.420	1.207
Implementation status indicator: unknown level in 2004 or in 2002	0.594	1.980	4.368**	1.958
Interaction terms between number of implementation years and indicator of implementation status	-0.108	0.072	-0.044	0.064
Interaction terms between 2002 implementation level and status from 2002 to 2004	-1.943	1.470	-1.410	1.439
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-0.729	2.435	-5.539**	2.403

Table B2. Effect of Teacher-Reported Implementation Level on Paired-Difference in Mathematics and Reading Gains From 1999 to 2003 (continued)

Effect of implementation	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Student grouping				
Number of implementation years	0.017	0.053	-0.063	0.050
Implementation level in 2002: use of technology in instruction	2.806	1.477	0.001	1.348
Implementation status indicator: level in 2004 higher than in 2002	2.638	1.685	2.772*	1.544
Implementation status indicator: unknown level in 2004 or in 2002	-4.713	90.263	-14.153	73.002
Interaction terms between number of implementation years and indicator of implementation status	0.111	0.067	0.126**	0.064
Interaction terms between 2002 implementation level and status from 2002 to 2004	-3.348	1.889	-3.641**	1.734
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	4.286	97.881	15.795	79.239
Parent-community involvement				
Number of implementation years	0.087	0.067	0.021	0.060
Implementation level in 2002: inclusion of special education students	2.254*	1.159	-2.406**	1.097
Implementation status indicator: level in 2004 higher than in 2002	1.454*	0.827	-0.896	0.772
Implementation status indicator: unknown level in 2004 or in 2002	-0.486	0.570	0.486	0.539
Interaction terms between number of implementation years and indicator of implementation status	-0.015	0.080	0.011	0.071
Interaction terms between 2002 implementation level and status from 2002 to 2004	-2.403*	1.393	2.005	1.311
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	0.000	—	0.000	—
Curriculum				
Number of implementation years	0.157***	0.044	0.021	0.045
Implementation level in 2002: inclusion of special education students	-3.507**	1.750	-1.394	1.636
Implementation status indicator: level in 2004 higher than in 2002	-1.047	1.797	-1.076	1.680
Implementation status indicator: unknown level in 2004 or in 2002	5.012	5.691	5.836	5.345
Interaction terms between number of implementation years and indicator of implementation status	-0.107*	0.064	0.006	0.065
Interaction terms between 2002 implementation level and status from 2002 to 2004	1.576	2.166	1.395	2.024
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-6.028	6.604	-6.869	6.205

Table B2. Effect of Teacher-Reported Implementation Level on Paired-Difference in Mathematics and Reading Gains From 1999 to 2003 (continued)

Effect of implementation	Mathematics		Reading	
	Estimate	SE	Estimate	SE
Teaching Time				
Number of implementation years	0.114**	0.052	-0.007	0.050
Implementation level in 2002: use of technology in instruction	0.497	0.681	-0.684	0.644
Implementation status indicator: level in 2004 higher than in 2002	0.765	1.191	0.953	1.125
Implementation status indicator: unknown level in 2004 or in 2002	0.334	3.575	-0.653	2.841
Interaction terms between number of implementation years and indicator of implementation status	-0.049	0.068	0.038	0.065
Interaction terms between 2002 implementation level and status from 2002 to 2004	-0.642	1.350	-1.218	1.281
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-1.016	4.028	1.198	3.286
Pedagogy				
Number of implementation years	-0.034	0.029	-0.016	0.025
Implementation level in 2002: inclusion of special education students	0.581	0.403	0.444	0.338
Implementation status indicator: level in 2004 higher than in 2002	0.680	0.521	0.351	0.450
Implementation status indicator: unknown level in 2004 or in 2002	0.665	0.909	0.555	0.833
Interaction terms between number of implementation years and indicator of implementation status	0.061	0.041	0.065*	0.035
Interaction terms between 2002 implementation level and status from 2002 to 2004	-0.942	0.580	-0.661	0.504
Interaction terms between 2002 implementation level and unknown status from 2002 to 2004	-0.820	0.989	-0.739	0.907

Note. This is the same model as reported in Table 8, and only coefficients of implementation variables are reported.

*p-value significant at 0.1 level. **p-value significant at 0.05 level. ***p-value significant at 0.01 level.