Note: Any RED Box is a link to another file or URL. To Access Appendix A or B, Click on the RED box in either the Table of Contents or within the text.

COMPREHENSIVE SCHOOL REFORM AND STUDENT ACHIEVEMENT A Meta-Analysis

Geoffrey D. Borman
University of Wisconsin-Madison

Gina M. Hewes
Laura T. Overman
Johns Hopkins University

Shelly Brown
University of North Carolina, Greensboro

Report No. 59

November 2002

This report was published by the Center for Research on the Education of Students Placed At Risk (CRESPAR), a national research and development center supported by a grant (No. R-117-D40005) from the Office of Educational Research and Improvement (OERI), U.S. Department of Education. The content or opinions expressed herein do not necessarily reflect the views of the Department of Education or any other agency of the U.S. Government. Reports are available from: Publications Department, CRESPAR/Johns Hopkins University; 3003 N. Charles Street, Suite 200; Baltimore MD 21218. An on-line version of this report is available at our web site: *www.csos.jhu.edu*.

THE CENTER

Every child has the capacity to succeed in school and in life. Yet far too many children fail to meet their potential. Many students, especially those from poor and minority families, are placed at risk by school practices that sort some students into high-quality programs and other students into low-quality education. CRESPAR believes that schools must replace the “sorting paradigm” with a “talent development” model that sets high expectations for all students, and ensures that all students receive a rich and demanding curriculum with appropriate assistance and support.

The mission of the Center for Research on the Education of Students Placed At Risk (CRESPAR) conduct the research, development, evaluation, and dissemination needed to transform schooling for students placed at risk. The work of the Center is guided by three central themes—ensuring the success of all students at key development points, building on students’ personal and cultural assets, and scaling up effective programs—and conducted through research and development programs in the areas of early and elementary studies; middle and high school studies; school, family, and community partnerships; and systemic supports for school reform, as well as a program of institutional activities.

CRESPAR is organized as a partnership of Johns Hopkins University and Howard University, and supported by the National Institute on the Education of At-Risk Students (At-Risk Institute), one of five institutes created by the Educational Research, Development, Dissemination and Improvement Act of 1994 and located within the Office of Educational Research and Improvement (OERI) at the U.S. Department of Education. The At-Risk Institute supports a range of research and development activities designed to improve the education of students at risk of educational failure because of limited English proficiency, poverty, race, geographic location, or economic disadvantage.

ABSTRACT

In this meta-analysis, we review the research on the achievement effects of the nationally disseminated and externally developed school improvement programs known as "whole-school" or "comprehensive" reforms. In addition to reviewing the overall achievement effects of comprehensive school reform (CSR), we study the specific effects of 29 of the most widely implemented models. We also assess how various CSR components, contextual factors, and methodological factors associated with the studies mediate the effects of CSR. We conclude that CSR is still an evolving field and that there are limitations on the overall quantity and quality of the research base. The overall effects of CSR, though, appear promising and the combined quantity, quality, and statistical significance of evidence from three of the models, in particular, set them apart from the rest. Whether evaluations are carried out by the developer or by third-party evaluators and whether these evaluators use one-group pre-post designs or control groups are especially important factors for understanding differences in CSR effects. Schools implementing CSR models for five years or more showed particularly strong effects, but the models benefited equally schools of higher- and lower-poverty levels.

A long-term commitment to research-proven educational reform is needed to establish a strong marketplace of scientifcally based models capable of bringing comprehensive reform to the nation's schools.

ACKNOWLEDGMENTS

The authors thank all 33 of the reform developers for their help in accessing studies and information about their CSR models. The authors also thank the three anonymous reviewers of this report for their suggestions and feedback.

CONTENTS

Introduction 1

What is CSR? 2

The Policy Context for CSR 4

Previous Reviews of CSR 6

Objectives and Hypotheses 7

Method 10

Selection of Comprehensive School Reform Models 10

Literature Search Methods 11

Inclusion Criteria 11

Moderator Variables 12

Data Imputation 16

Independence of Observations 16

Characteristics of the Selected Studies 17

Results 18

Computation of Effect Sizes 18

Computation of Variance Components, Weights, and Weighted Effect Sizes within a Random-Effects Model 23

Distribution of Measures of Central Tendency for Effect Size 24

Regression Analysis of Weighted Effect Sizes on Mediating Variables 24

School Poverty Level and Years of Implementation as Moderators of Effect Sizes 26

Analysis of Subject Area as Moderator of Effect Size 27

Evidence of Effectiveness for the 29 CSR Models 28

Discussion 33

Characteristics of the CSR Research Base 33

Overall Effects of CSR 34

Explaining Differences in CSR Effects 36

Conclusion 38

References 40

Notes 44

Appendix A: Studies Included in the Meta-Analysis A.1

Appendix B: CSR Design Characteristics B.1

INTRODUCTION

The latter half of the 20th century was marked by recurring efforts at school reform and improvement in the United States. Yet, as Slavin (1989) observed, this cycle of reforms—like a pendulum swing—has continued to move from one fad to another with little evidence of national progress. As each new reform is widely disseminated and implemented, the research follows closely behind, sometimes weighing in on the issue only after the schools have moved on to the next apparent innovation. Recent national reform and policy movements, though, may halt this frustrating cycle. Indeed, for the first time, Congress and other educational policymakers are making some funding sources available to only those schools that implement educational reforms with high-quality evidence of effectiveness. Most notably, the Comprehensive School Reform Program (CSRP)—formerly known as the Comprehensive School Reform Demonstration (CSRD) program—provides grants to schools to adopt proven comprehensive reforms. With the recent proliferation of externally developed comprehensive school reform (CSR) models, the simultaneous growth in the CSR research base, and the significant public and private financial backing for this reform movement, the potential for spawning a national wave of research-based educational innovation has never been greater.

In addition to their focus on research-based solutions for school improvement, current CSR initiatives help reconcile the two most important recent educational reform movements in the United States. Since the 1980s, competing, and often contradictory, reforms have combined top-down, centralized efforts to improve schools and teaching with efforts at decentralization and school-based management (Rowan, 1990). The general spirit of today's reform efforts continues to articulate top-down standards, which dictate much of the changes in the content of schooling, but fundamentally leaves the process of school change up to the discretion of local educators. The problem is that the complex educational changes demanded by current standards-based reform initiatives, combined with an increasingly heterogeneous student population largely composed of students whom schools have traditionally failed, have pushed the technology of schooling toward unprecedented levels of complexity. In many ways, expecting local educators to reinvent the process of educational reform, school by school, is both unrealistic and unfair. Externally developed CSR models provide a type of top-down direction for designing and supporting the process of school reform. In this case, though, the top-down direction is not in the form of distant legislative mandates, but is, in theory, tangible and accessible support for school change rooted in research and literally packaged and delivered to each school.

In this exhaustive meta-analysis, we review all known research on the achievement effects of the most widely implemented, externally developed school improvement programs known as "whole-school" or "comprehensive" reforms. In addition to reviewing the achievement effects of CSR as a general strategy, we synthesize research on the specific effects of the 29 most widely implemented CSR models.[1] In quantifying the overall, and specific effects, of CSR models, we also assess how the methodological and contextual factors associated with the studies of CSR differ. In addition, we identify common components, such as whether the model specifies and includes a particular curriculum, or whether it specifies and provides a plan for the ongoing professional development of teachers. Using these methodological, contextual, and programmatic factors as predictors of effect size, we assess how they may influence the estimates of the models' effects. The resulting information allows us to examine:

- the general effectiveness of the CSR strategy;
- the effects associated with specific CSR model components;
- the effects of each of the 29 CSR models; and
- the extent to which differences in the methodological and contextual features of the studies mediate the estimated effects.

What is CSR?

The "scale up" of CSR designs is happening at an unprecedented rate, as evidenced by the growing number of externally developed school reform designs (e.g., Accelerated Schools, Core Knowledge, High Schools that Work, Success for All) being implemented in thousands of schools, serving millions of students throughout the United States. CSR focuses on reorganizing and revitalizing entire schools, rather than on implementing a number of specialized, and potentially uncoordinated, school improvement initiatives. In general, the funding sources supporting the implementation of CSR have been targeted toward the schools most in need of reform and improvement: high-poverty schools with low student test scores. According to recent data from the Southwest Educational Development Laboratory (SEDL), schools receiving money to implement CSR models through the CSRP have an average poverty rate of 70%. Further, nearly 40% of schools receiving CSRP funds were identified for school improvement under Title I regulations and more than 25% were identified as low-performing schools by state or local policies.[2]

The other significant funding source for CSR programs has been Title I of the Elementary and Secondary Education Act (ESEA) of 1965, which also aims to expand and improve educational opportunities in the nation's high-poverty schools. In January 2002, with the reauthorization of Title I as the *No Child Left Behind* Act, the CSRP and Title I came together under the same legislation. As Title I, Part F, CSRP has become a significant component of the growing federal movement to support scientifically based efforts to reform low-performing high-poverty schools across the nation.

The U.S. Department of Education defines CSR using 11 components that, when coherently implemented, represent a "comprehensive" and "scientifically based" approach to school reform. Specifically, a CSR program:

1. Employs proven methods for student learning, teaching, and school management that are based on scientifically based research and effective practices, and have been replicated successfully in schools;
2. Integrates instruction, assessment, classroom management, professional development, parental involvement, and school management;

3. Provides high-quality and continuous teacher and staff professional development and training;

4. Includes measurable goals for student academic achievement and establishes benchmarks for meeting those goals;

5. Is supported by teachers, principals, administrators, and other staff throughout the school;

6. Provides support for teachers, principals, administrators, and other school staff by creating shared leadership and a broad base of responsibility for reform efforts;

7. Provides for the meaningful involvement of parents and the local community in planning, implementing, and evaluating school improvement activities;

8. Uses high-quality external technical support and assistance from an entity that has experience and expertise in schoolwide reform and improvement, which may include an institution of higher education;

9. Includes a plan for the annual evaluation of the implementation of the school reforms and the student results achieved;

10. Identifies federal, state, local, and private financial and other resources available that schools can use to coordinate services that support and sustain the school reform effort; and

11. Meets one of the following requirements: the program has been found, through scientifically based research, to significantly improve the academic achievement of participating students; *or* the program has been found to have strong evidence that it will significantly improve the academic achievement of participating children. (U.S. Department of Education, 2002)

Some schools develop their own "home-grown" reform models having these characteristics. As suggested by the eighth component of CSR, though, many educators are turning to external groups, such as universities and educational centers and labs, for assistance in designing whole-school reform models.

Externally developed reform designs are consistent in that they provide a model for whole-school change and attempt to help schools address many, if not all, of the 11 components mentioned previously. At the same time, though, the externally developed designs are remarkably diverse in their analyses of the specific problems in U.S. education, the solutions that they propose, and the processes they propose for achieving those solutions. For example, the Comer School Development Program builds largely around Dr. James Comer's work in community psychiatry, focusing its energy on creating schools that address a wide range of students' health, social, emotional, and academic challenges (Comer, 1988). By contrast, the Core Knowledge reform (Hirsch, 1995, 1996) derives from the developer's experiences as a professor of English and education, and focuses almost entirely on the establishment of a "common core" of knowledge for all children within various subject areas including literature, history, science, mathematics, and the arts. The Coalition of Essential Schools model attempts to create more educationally rich and supportive learning environments through a common adherence to nine, broadly philosophical principles (Sizer, 1992), whereas Success for All (Slavin & Madden, 2001) provides a specific K-6 reading curriculum, professional development sequence, and other schoolwide components.

CSR is expanding rapidly because many models have established development and dissemination infrastructures for replicating and supporting implementations across numerous schools. In other words, the developers can transport their CSR models to schools across the U.S., help local educators understand the tenets of the reform, and teach them how to implement the school organization and classroom instruction that the model suggests. In every case, the developers provide some type of initial training or orientation to help educators to, at least, understand the underlying philosophy of the model. In many circumstances, though, replication also involves a more specific "blueprint" for implementing and sustaining the model. Highly specified models, for instance, often prescribe new curricular materials, new methods of instruction, alternative staffing configurations, and a series of ongoing professional development activities.

The Policy Context for CSR

In addition to the replicable nature of many of the models, expansion of CSR has been fueled by a series of recent national developments: the movement toward systemic and standards-based reform; the establishment of the New American Schools Development Corporation; new federal legislation allowing the use of Title I funds—the primary source of federal assistance to at-risk students from high-poverty schools since 1965—to support schoolwide educational programs in high-poverty schools; and the federal CSRP legislation that provides hundreds of millions of dollars to support the costs of adopting externally developed reform models. Only since the mid-1990s has the idea of schoolwide reform emerged as a prominent strategy for helping improve the outcomes of at-risk students from high-poverty schools. Before then, the school-based services funded through Title I, and other categorical programs for at-risk students, targeted only those students with the lowest test scores. As a result, the vast majority of schools used the funds to develop specialized pullout programs that provided remedial services to the subgroups of students with the greatest academic needs (Borman, Wong, Hedges, & D'Agostino, 2001).

Instead of the seemingly piecemeal and uncoordinated categorical targeted assistance programs that had served Title I schools since the mid-1960s, a growing belief developed that at-risk students and high-poverty schools could be better served by schoolwide reforms. This belief was encouraged by informed opinion (e.g., Rotberg, Harvey, & Warner, 1993), by general findings from the effective schools research tradition (Edmonds, 1979; Teddlie & Reynolds, 2000), and by the concept of systemic reform (e.g., Smith & O'Day, 1991), more than by specific groundbreaking empirical studies. Inspired by the emerging vision of standards-based reform, the 1994 reauthorization of Title I called on states to raise academic standards, to build the capacity of teachers and schools, to develop challenging new assessments, to ensure school and district accountability, to ensure the inclusion of all children, and to develop coordinated systemic reforms.

The new legislation encouraged schoolwide initiatives rather than targeted programs for all schools where at least 50% of the students were poor. These sweeping changes began the transformation of Title I from a supplemental remedial program to the key driver of the standards-based, schoolwide reform movement (Borman, 2000a).

During the 1990s, Title I schoolwide projects proliferated across the country. In 1991, only 10% of the eligible Title I schools operated schoolwide programs, but by 1996, approximately 50% of the eligible Title I schools had implemented them (Wong & Meyer, 1998). Rather than implementing the characteristic Title I pullout programs, educators were granted the flexibility to invent and implement their own reforms designed to upgrade the whole school. A number of studies from the 1990s showed that, in the short-term, these schoolwide efforts did not produce compelling evidence of positive achievement effects and, for the most part, did not result in the desired reforms (Wong & Meyer, 1998, 2001). Also during the 1990s, a more general review indicated that site-based management reforms failed to affect student outcomes positively in large part because the schools failed to develop coherent statements of beliefs or models for guiding the work and decision-making of the school (Murphy & Beck, 1995). These outcomes, combined with new evidence from the Congressionally-mandated Prospects study of the modest overall impacts of Title I services (Borman, D’Agostino, Wong, & Hedges, 1998; Puma, Karweit, Price, Ricciuti, Thompson, & Vaden-Kiernan, 1997), suggested that federal policies for improving education for at-risk students from high-poverty schools were in need of further retooling.

At the same time, the growing research base on several externally developed school restructuring efforts, such as the Comer School Development Program (Comer, 1988; Haynes, Emmons, & Woodruff, 1998) and Success for All (Slavin, Madden, Dolan, Wasik, Ross, Smith, & Dianda, 1996; Slavin & Madden, 2001), seemed to indicate hope for a high-quality education for at-risk students. In addition, the companion study to the national Prospects evaluation of Title I, the Special Strategies Study (Stringfield et al., 1997), indicated that whole-school, externally developed programs funded by Title I appeared more likely to have positive impacts on academic achievements than either traditional Title I pullout programs or locally developed reforms.

Along with the growing policy and research support, in 1991 then-President George Bush announced the creation of a private-sector organization called the New American Schools Development Corporation (NAS), which was intended to support the creation of “break the mold” whole-school restructuring models for the next century (Kearns & Anderson, 1996). Using a business model, NAS turned to the marketplace for proposals for new models of American schools that would enable all students to achieve world-class standards in core academic subjects, operate at costs comparable to current schools after start-up funding, and address all aspects of a school’s operation. After receiving nearly 700 proposals in February 1992, NAS chose 11, and provided funds for a three-year program of development and testing. Since 1995, NAS has continued to focus on “scaling up” seven of the models to thousands of schools nationwide. Providing more than $150 million over the past decade in financial and technical assistance to the reform developers, NAS has helped create a market for CSR and has helped scale up the CSR movement.

In response to the promise of the externally developed programs disseminated by NAS and by other independent model developers, the U.S. Congress has encouraged individual schools to implement “scientifically based” whole-school reforms and to seek the assistance of external groups in developing their school reform plans. In 1998, Congress initiated the CSRP, which encourages schools to develop comprehensive plans for implementing “scientifically based” strategies for school reform. Through a competitive process, CSRP awards a minimum of $50,000 per year for three years to qualifying schools. Since first authorizing CSRP in fiscal year 1998 and allocating $145 million, Congress has steadily increased its support. In fiscal year 2002, allocations for CSRP equaled $310 million—$235 million specifically for Title I schools and $75 million for any schools wishing to apply through the Fund for the Improvement of Education. This initiative, combined with Title I’s continuing focus on schoolwide change and the efforts of NAS and other independent developers, has led to the continuing expansion of externally developed CSR models.

Previous Reviews of CSR

To date, there have been five major practitioner-oriented reviews, or “catalogs,” of CSR models (see Herman et al., 1999; Northwest Regional Educational Laboratory, 1998, 2000; Slavin & Fashola, 1998; Traub, 1999; Wang, Haertel, & Walberg, 1997). Due to the rapid expansion of the CSR movement and the CSR research base, though, these reviews are quickly growing outdated. Although the reviews, most notably the Herman et al., and Slavin and Fashola publications, have provided some appraisals of the effects of the various CSR models, none has offered a comprehensive, quantitative synthesis of the overall effects of CSR or of the effects of the various CSR models. Rather, as Stringfield (2000) suggested, these publications are akin to *Consumers’ Reports* guides for education, offering information that is important for educators to consider when “shopping” for a reform model. The reviews typically contain summaries of the general attributes of the CSR models, appraisals of the level of support that is provided by the developers, the costs associated with implementing the models, and various ratings of the strengths of the research supporting each CSR design.

In addition to these reviews of CSR models, there have been several recent articles critiquing the research supporting particular models and CSR in general. Most notably, these criticisms have suggested that some CSR research may be tainted by the fact that the developers are often also the evaluators (Pogrow, 2000; Walberg & Greenberg, 1999). Another source of controversy involves whether the use of a quasi-experimental, untreated control group design is really preferable to an analysis of pretest-to-posttest gain scores across a large number of sites (Pogrow, 1998; Slavin, 1999). In a sense, this debate has pitted the greater reliability of a large number of gain-score analyses against the greater internal validity of a relatively small number of matched control-group designs when attempting to judge whether an educational intervention has produced “exemplary” effects on student achievement. Despite the controversy and debate, no empirical data from CSR evaluations have been systematically brought to bear on either question.

Beyond these methodological considerations, studies and reviews of CSR and the process of school change have identified several common, substantive factors that have a bearing on the success or failure of externally developed reforms. First is the rather straightforward observation that the quality of the CSR model implementation matters. A number of researchers have demonstrated a strong relationship between reform implementation and positive effects—both qualitative and quantitative—across a variety of reforms (e.g., Berman & McLaughlin, 1978; Crandall et al., 1982; Datnow, Borman, & Stringfield, 2000; Stringfield et al., 1997).

Second, although some reform models have been criticized because their prescriptive designs may suppress teacher creativity and require an inordinate amount of preparation time (Datnow & Castellano, 2000), externally developed reforms that are more clearly defined tend to be implemented with greater fidelity and, in turn, tend to have stronger effects on teaching and learning than reforms that are less clearly defined (Bodilly, 1996, 1998; Nunnery, 1998). Third, well-implemented reforms tend to have strong professional development and training components and effective followup to address teachers' specific problems in implementing change within their classrooms (Muncey & McQuillan, 1996; Nunnery, 1998). Finally, for external models of school change to make an important impact within schools, teachers and administrators must support, "buy into," or even help "co-construct" the reform design (Borman et al., 2000; Datnow & Stringfield, 2000). Although there have been no systematic analyses across a wide range of CSR models, it would seem that those models with clear components addressing each of these issues would tend to result in more reliable implementations and stronger effects than CSR models without such components.

Further, the federal government has detailed 11 clear characteristics, outlined previously, of what it views as a truly comprehensive approach to reform. Not surprisingly, some of these overlap with the components identified in the CSR and school change research literature, including high-quality technical support from the external CSR partner, continuous teacher and staff development and training, and staff support or "buy in" for the reform initiative. The federal recommendations, though, cite several other characteristics that may be associated with effective CSR models, such as evidence that the reform has been replicated successfully, measurable goals for student performance and benchmarks for meeting those goals, and the involvement of parents and the community in the governance of the school and the development of the school improvement plan.

Objectives and Hypotheses

The results from studies of CSR differ in many ways, including (a) who reported the findings (i.e., the developer or someone else); (b) the methods used (e.g., pretest-posttest comparison, experimental comparison, or nonequivalent control-group design); (c) the student and school context (e.g., high-poverty versus lower poverty settings); (d) actual characteristics of the CSR models (e.g., the costs associated with the model, or the level of support for implementation provided by the developer); and (e) indicators of the model's effectiveness (e.g., test scores from reading, math, science, or some other subject). Differences across studies such as these are commonly found in the social sciences, especially in the case of education.

Indeed, given the programmatic, methodological, and contextual diversity of the CSR literature, questions emerge concerning how, or if, we should proceed with a synthesis of its findings. As Borman (2000b) pointed out, there are varying perspectives on what the reviewer should do when confronted by such a variegated literature, in terms of overall research quality and other features, such as the research designs, samples, and the actual circumstances involved. On one hand, Glass (1976) stated, “It is an empirical question whether relatively poorly designed studies give results significantly at variance with those of the best designed studies” (p. 4). On the other hand, Slavin (1986) argued, “Far more information is extracted from a large literature by clearly describing the best evidence on a topic than by using limited journal space to describe statistical analyses of the entire methodologically and substantively diverse literature” (p. 7). Should the researcher combine studies that used varying methods and are characterized by varying substantive characteristics or should one focus only on the “best evidence?”

We believe that there are two important reasons to begin our analysis with a review of the complete CSR literature. First, as Glass (1976) suggested, by empirically examining a diverse range of studies, we may assess how and to what extent methodologic differences across the studies are associated with differences in CSR effects. When outcomes are robust across studies of varying methodologies, one can be more confident in the conclusions. On the other hand, if studies differ in terms of both rigor and results, then one may focus on the subset of more rigorous studies when formulating conclusions. This analysis of the consequences of methodological variations for the estimation of CSR effects, which is unique to the enterprise of meta-analysis, allows methodologists and consumers of the research literature to recognize the biases in the literature and to understand empirically both their frequency and magnitude.

Second, from a practical perspective, relatively little is known about what common components characterize effective CSR models. Well-intended federal policies have outlined the elements that should comprise a school reform that is truly comprehensive. These policies, though, have not benefited greatly from the cumulative knowledge of the CSR research base. By examining how effects vary across models and contexts, it is our hope to provide new evidence of both how and where CSR may make the biggest difference in student achievement. It also may suggest some components or specific models that do not appear to be affecting student outcomes in meaningful ways.

Our meta-analysis begins by assessing these methodological, programmatic, and contextual variations across an extensive collection of all known studies of 29 of the most widely discussed and disseminated CSR models. This preliminary analysis shows how and to what extent the methodologic al, programmatic, and contextual factors shape our understanding of the overall achievement effects of CSR. Specifically, the preliminary analysis empirically identifies and quantifies the potential methodological biases in the literature, reveals the common characteristics of CSR programs that make a difference in terms of student achievement, explores differences in achievement effects associated with varying contexts (e.g., the grade level or the subject area targeted by the reform), and, in general, characterizes the overall quality of the research evidence.

After characterizing the overall CSR research base, and after empirically identifying its potential methodological biases, our second objective is to assess the efficacy of each of the 29 CSR models. Rather than surveying the overall CSR research base and the methodological, programmatic, and contextual factors within it, this phase of our research develops standards for assessing the quality, quantity, and statistical significance of the models' effects on student achievement. In short, we establish the extent to which each of the 29 models is supported by scientifically based research. We address this concern by focusing on only the subgroup of studies that provides the best evidence for evaluating the effectiveness of each of the 29 CSR models. We determine which studies provide the best evidence not by *a priori* judgments or by other potentially subjective criteria, but by our empirical analyses of the CSR literature's methodological biases.

Obviously, our hypotheses concerning the evaluation results require attention to a range of moderating influences that are model-specific, methodological, and contextual in nature. Model-specific influences include those that we identified in our literature review: how tightly prescribed the reform design is, especially as it relates to curriculum and instruction; the extent to which the developer provides ongoing technical support and professional development to address teachers' specific problems in implementing the reform; and the ways in which developers secure teacher support for the reform. They also include various foci suggested by the 11 components identified in the federal definition of CSR. These include having measurable goals for student performance and benchmarks for meeting those goals, incorporating a strong parent-governance component, and providing evidence of successful replication of the model. Though relatively little quantitative research has linked these model-specific influences to student achievement, we hypothesized that CSR models having specific components designed to address the areas identified in our literature review and the 11 federal characteristics of CSR would tend to be better implemented and more comprehensive reforms than CSR models without these components. In turn, we expected the better implemented and more comprehensive models to yield the strongest effects on student achievement.

The two primary methodological characteristics we identified are related to who is doing the research and the general strength, or internal validity, of the study design that the researcher chooses. We hypothesized that evaluations performed by the CSR developer would yield higher estimates of effects than evaluations done by others. In addition, we predicted that studies employing experimental or quasi-experimental treatment-control comparisons would yield lower effect estimates than studies based on analyses of CSR pre-post gain scores. Though imperfectly matched comparison groups could cause positive biases, it is more likely that effect estimates based on simple one-group pre-post designs will yield greater positive biases. Cook and Campbell (1979) note that threats to internal validity, including history, maturation, and regression-to-the-mean effects, are likely to make one-group pre-post designs among the weakest. Also, empirical results from a meta-analysis of Title I program effects by Borman and D'Agostino (1996) illustrated that analyses of pre-post gains resulted in positive biases, relative to studies employing quasi-experimental control group comparisons, of approximately one fifth of one standard deviation.

The contextual factors affecting CSR effects are largely unexplored and are, therefore, less predictable. Our analyses of the relative effects of CSR in reading, math, and other subjects, across various grade levels, and across varying poverty levels are unprecedented. Given the targeting of recent policies, most notably the CSRD program, on scaling up reform within high-poverty contexts, we hoped to find particular benefits for these schools.

METHOD

Selection of Comprehensive School Reform Models

The goal of our analysis was to synthesize the research on widely disseminated, externally developed, CSR, or whole-school, reform models. To be considered for the current study, therefore, a reform model needed to meet four basic criteria: 1) it is a whole-school or schoolwide reform design; 2) it is the subject of at least one prior study, whether positive or negative, on which we could base our review; 3) it is a model that is disseminated by developers external from the schools; and 4) it has been replicated in 10 or more schools. Previous reviews and catalogs of reform models, including the fall 2000 edition of the Northwest Regional Educational Laboratory's (NWREL) *Catalog of School Reform Models* (NWREL, 2000) and *An Educators' Guide to Schoolwide Reform* published by the American Institutes of Research (AIR) (Herman et al., 1999) used similar selection criteria.[3] At the time of our selection, these publications were the only known sources of information available to define the universe of CSR models meeting our criteria.

Therefore, our selection of models drew directly from the previous NWREL and AIR catalogs. Through these sources, we identified 33 CSR models, but only 29 of the models provided at least one report of their achievement effects from which we could calculate effect size estimates. The 33 models originally selected for the present research were implemented in 55.6% of the schools that received CSRP funds for externally developed models, as reported in the SEDL database, and the 29 models ultimately included in this review represented 53.4% of the CSRP implementations. The results of this review should generalize reasonably well to the population of schools implementing CSR models using CSRP and Title I program funds. The review, though, clearly does not represent schools that use these funds to implement "home-grown," non-externally developed CSR designs, or schools that package one or more externally developed, targeted, non-schoolwide interventions to develop their own CSR approaches. Finally, because we cannot review the research for CSR models with no research base, these models are not represented in this synthesis either.

Summary descriptions of each school reform model are presented in Table 1 and further descriptive information about the main features and costs of each model is presented in Appendix B. The descriptive information in the appendix is adapted from the NWREL's *Catalog of School Reform Models* and is supplemented with a narrative description of each reform's research base.

Literature Search Methods

Broad searches of the literature on comprehensive school reform and its effects on student achievement were conducted using several approaches. The preliminary literature review involved computerized searches of the Education Resources Information Center (ERIC) database (1966-2001) and the PsychLit database. A second method used general World Wide Web searches (search engines such as Google), and specific searches of CSR developers' web pages for references to research or any other published or unpublished studies or compilations of data. We also collected all studies referenced in the Herman et al. (1999) and NWREL (1998) reports.

After completing this initial review stage, we compiled separate lists of the references gathered for each of the 33 reform models. We then sent these lists to each of the developers for their review and feedback. All 33 developers responded, either to confirm that our list included all the references known to them, to make suggestions for further references, or to provide studies we were unable to obtain through other sources. The final phase of review involved exhaustive bibliographic reference chasing based on all reports obtained through the computerized databases, via the World Wide Web, and from the developers. After performing this series of search methods, we found no other available evaluations of comprehensive school reform and student achievement outcomes.

The period of aggressive collection of studies began in fall 2000 and concluded at the end of that calendar year. After 2001 began, we no longer conducted an extensive literature search; we did, however, continue to contact reform developers as necessary and followed up with locating articles discovered in the previous round of literature searches and the review of references in articles as they arrived at our facility. Thus, the review includes studies completed through late 2001.

Inclusion Criteria

Liberal inclusion criteria were applied in the preliminary stages of the literature search. All study abstracts provided by the database searches and all evaluations of comprehensive school reform and student achievement that were referenced in the documents were reviewed to ascertain whether any report of student achievement data, in the form of test scores, may have been provided by the studies. If an abstract or study did not suggest these data were reported, the study was excluded from further consideration. More than 800 studies, abstracts, and summaries were read during this preliminary stage of the review process. The vast majority of these studies, though, were not considered beyond this stage, as they typically documented implementation outcomes or the theories supporting the reform model but provided no assessment of the model's achievement effects.

In the second stage, we focused on the subset of studies that provided some form of assessment of the model's effects on students' test scores. From these studies, we chose those that allowed us to generalize to the effects of externally developed whole-school reform models implemented in the United States. In other words, the studies we selected had to help us answer the question: “What would be the expected achievement effects of the reform model if a school or district in the United States chose to contact the developer and arrange to implement the program as a schoolwide intervention?” More specifically, we deemed studies eligible for further consideration based on the following criteria:

- sufficient achievement data for reform participants, and, when applicable, comparison group students, were provided from which effect sizes could be computed;
- the study design involved some form of comparison from which an effect could be determined: either a one-group pre-post design involving treatment students only or a quasi-experimental or experimental treatment-control comparison;
- the sample or data provided were not duplicated in another study accepted for inclusion;
- the sample used in the evaluation was composed of students from a school in the United States; and
- the sample of students was from the school’s regular education program.

Many studies reviewed did not meet these eligibility requirements. This was due, in large part, to insufficient information for calculating effect sizes. The most common reason for excluding studies was the failure to provide a standard deviation or information about the testing instrument from which a standard deviation could be imputed (imputation of data is discussed below in more detail). A substantial number of studies included samples or data that were reported in other studies accepted for inclusion, so they were eliminated. Many other studies used a non-U.S. sample, or a special population, such as special education students. In the end, 232 studies (see Appendix A) met all requirements and were selected for analysis.[4]

Moderator Variables

In addition to collecting the information necessary for calculating effect sizes and weights (e.g., achievement outcomes, standard deviations of the achievement outcomes, and the sample sizes), we coded a number of other characteristics that corresponded to two general areas: contextual information related to the particular implementation that was evaluated, and methodological variables related to the study design. Because studies often reported multiple outcomes from multiple contexts or multiple research designs, the contextual and methodological characteristics were coded at the level of the outcome rather than at the level of the study.

Contextual Variables. Contextual variables helped us examine potential differences in effect size related to the context in which the CSR model was evaluated. The contextual variables included:

- subject area tested;
- grade level evaluated;
- years of CSR model implementation for the results given; and

- the poverty level of the school served by the CSR model. [5]

We identified five major subject areas that were tested and evaluated in the CSR literature: language arts, math, science, social studies, and a general category. These were developed from a wider array of subject areas noted in the studies. Language arts included reading and other literacy-related subcategories such as comprehension, vocabulary, spelling, language, word knowledge, and writing. Math covered subcategories including computation, arithmetic, and math applications. Science included both science and health. Social studies included all social sciences and history. The general category typically consisted of composite scores across subjects or general ability measures. These mutually exclusive categories were coded into five indicator variables.

The grade level tested was a dichotomous variable, where "0" represented elementary school grades (K-5), and "1" represented all other grades (i.e., 6-12 and mixed across levels). If a study listed a range of grades associated with an achievement outcome such that grade levels were mixed across the elementary/middle school break, that outcome was assigned to the "1" middle/high/mixed grades category. For example, if a study provided outcome data for students in grades 4-6, the outcome was assigned "1" on this variable.

A smaller subgroup of studies identified the number of years that the CSR model had been implemented at the school and indicated the poverty level of the school. In all cases, we indexed poverty level by the percent of students at the CSR school who were eligible for the free lunch program. The number of years that the CSR model had been implemented at the school site ranged from 1 school year to 14 school years, with an average of 2.96 years.

Methodological Variables. The methodological variables describing the evaluations included the following:

- type of effect data provided (i.e., correlational, categorical, or mean difference);
- type of research design (i.e., randomized experiment, quasi-experimental matched school design, quasi-experimental covariate-adjusted design, quasi-experimental match to a "similar" school, quasi-experimental comparison to state or district outcomes, or one-group pre-post design);
- whether the study used a longitudinal design or not; and
- whether the study was conducted by the reform developer or not.

Each of these methodological characteristics was represented by an indicator code (0, 1) in our analyses. For type of effect data, we coded an outcome as one that provided correlational effect data when it showed a simple correlation between participation in the reform model and student achievement. Categorical effect data included outcomes that provided a binary achievement result, such as pass/fail or met standard/did not meet standard. The final type of effect data came from mean treatment-control achievement differences or pre-post differences for the treatment group.

We coded six types of research designs, including those that used (a) true random assignment of schools or students to the CSR and control conditions; (b) a quasi-experimental design that included explicit matching of the CSR school (or students) with a comparison school (or students) based on prior achievement levels and student demographics; (c) a covariate-adjusted comparison between the CSR school (or students) and non-CSR school (or students) based on prior achievement levels and, occasionally, student demographics; (d) a comparison of the CSR school (or students) to a non-CSR school (or students) stated to be "similar" based on unspecified criteria; (e) a simple comparison of the CSR school (or students) to all other schools (or students) in the district or state; and (f) a one-group design examining pre-post changes in the CSR school's (or students') achievement outcomes. For our main analyses, we contrasted the one-group pre-post analyses to all of the other designs, which used some form of comparison group.

Third, we coded an indicator variable as "1" for studies that used a true longitudinal design, which tracked the achievement outcomes for the same group of students over time. True longitudinal designs included all outcomes for which there were two or more time points, including simply a pretest and posttest, for the same sample of students at each time point. All other outcomes, including those that contrasted the results for one grade cohort of students in one year to the results for the same grade cohort in a subsequent year, and those that included a simple cross-sectional, posttest-only comparison, were coded as "0." Our original coding scheme provided more detail on the research design, including several distinct types of cohort studies. In analyses not shown, however, all of the non-longitudinal comparisons were found to yield similar effects, or were simply too few in number to stand alone. Consequently, all research involving non-longitudinal designs was pooled and contrasted to true longitudinal designs.[6]

The final methodological characteristic that we coded contrasted evaluations by the CSR developer to those done by others. Those studies that included among its authors the name of any of the CSR model's original developers were coded as "1," and all other studies were coded "0."

Reform Attributes. Separate from the data entry and coding for each study, each of the 29 CSR models was coded for its components by two or three independent coders as to whether each reform model possessed each of the following characteristics:

- required a set of specific curricular materials;
- required replicable pedagogical practices;
- required a faculty vote with at least 75% approval before the reform could be adopted;
- required a specific and replicable component designed to engage parents and the community in the governance of the school and the planning and implementation of the school improvement process;
- required a set of replicable student performance assessment methods and benchmarks that schools may use to track students' progress; and
- required ongoing teacher and staff professional development and training.

Also, for each of the 29 reforms, we documented the number of schools in which the reform had been replicated, the level of technical support the developer provided to schools, and an estimate of the full marginal cost for the first year of implementing the model.

The information for coding these reform attributes came from the Herman et al. (1999) report, the NWREL (1998) catalog, the developers' websites, documents from the developers, and in some cases, direct contact with the developers. The coding relied on interrater agreement among two or three coders, who independently coded the first six (bulleted) attributes. Where the coders did not agree, consensus was met by discussing the reasons for the selected response. If, after this process, there was still no consensus, the CSR developers were contacted to clarify. A single coder derived all cost information, the level of developer support, and the number of replicated schools for the reform models.

Seven of the nine attributes were coded "yes" or "no" for each reform. One attribute, ongoing access to technical support and assistance from the developer, was adapted from the Herman et al. (1999) report, in which it was presented as a scale ranging from 0-4. On this scale, a score of 0 indicated that the developer provided no on-site or other assistance to help schools implement the model, essentially no contact with the school after CSR implementation, and no benchmarks or other useful tools for helping schools assess the progress of their implementation. A score of 4 suggested that the developer provided on-site and other assistance to help schools implement the model, maintained frequent contact with the school after CSR implementation, and provided useful benchmarks and tools for helping schools assess their progress. For reforms not rated in Herman et al. (1999), we used these same criteria to develop ratings on the same scale. There was little variation in reforms' ratings on the 0-4 scale. Most reforms were 3 or 4, with only one being coded as a 2 and none as a 1 or 0. Thus, we recoded this information into an indicator variable where "1" represented the highest support rating of 4 and "0" indicated all other lower ratings.

The number of schools at which the reform was replicated was a continuous variable. This variable was based on the most recent information available regarding the number of schools being served by each of the CSR model developers. The NWREL (2000) catalog provided this information and the date associated with it. When the information was missing, or if the date was before NWREL's most recent update (May 1, 2001), developers were contacted directly for up-to-date information.

Both the full first-year marginal personnel and non-personnel costs for implementing each reform model were estimated. Non-personnel costs included the amount a school would be expected to pay for all materials and services provided by the developer and any additional costs associated with computers, furnishings, and other items demanded by the reform model but not provided through the developer. Personnel costs included the costs of hiring any new staff associated with the reform (e.g., tutors, full-time facilitators, or coaches). In essence, these marginal cost estimates provided a "worse-case scenario" for the costs of the reform. They estimated the total dollar amount of all resources that are demanded by the CSR model, regardless of whether schools could reallocate existing resources to the CSR implementation. For 21 reform models, the total marginal costs were estimated based on information provided in the Herman et al. (1999)

report.[7] For the other eight models, the costs were estimated from information from the developers. All costs were based on a school of 500 students and 25 teachers, and were separated into personnel and non-personnel cost variables.[8]

Data Imputation

To make use of the maximum number of studies possible, we imputed estimates for sample size and standard deviation under a limited range of circumstances. In all cases, outcomes for which data were imputed were flagged with a dummy code. These two imputation dummy codes, for sample size and standard deviation, were included as covariates in our final analyses of effect size.

Sample Size. If the student sample size was not provided, we estimated the number of students involved in the study based on national averages obtained from the National Center for Education Statistics' 1998-1999 Common Core of Data. In addition, this procedure relied on information in the study indicating the grade level of the sampled students and the number of schools included in the analysis. For example, if an evaluation reported data for second graders in one school, but not the actual sample size, we estimated the sample size to be 75, which is the average size of a school's second grade cohort based on national data from the 1998-1999 Common Core of Data.

For studies that used a district or state as the comparison group, we imputed the comparison group sample size as the treatment sample size rather than using the true district or state sample size. We employed this method to avoid dramatically inflating the weights assigned these studies and conferring a level of precision to these results that was not appropriate.

Standard Deviation. If we were not able to obtain the pooled standard deviation from the study, we imputed a standard deviation in one of two ways. First, if the test was a national standardized test, we consulted available norming data from the test developer to obtain a standard deviation. Falling into this first category are situations where Normal Curve Equivalent (NCE) scores were presented without sample standard deviations. In these cases, we imputed the population standard deviation of 21.06 and flagged the case. Second, if the test was a state or local assessment, for which the state or district maintained a web page, we used the overall state or local standard deviation reported for the test, grade, and year that corresponded to our data. These strategies of using national, state, or local population standard deviations are akin to methods outlined by Hedges (1981) for computing effect sizes, namely Cohen's *d* or Hedges's *g*, based on the average, or pooled, standard deviation.

Independence of Observations

There were several situations that threatened the assumption of independence of observations, which is central to most forms of hypothesis testing. The most obvious of these were reports of duplicate samples, which could arise in three ways: a) when researchers included the same sample in multiple studies; b) when researchers presented multiple analyses of the same sample in one or more studies by using somewhat different sets of covariates, for example; and c) when researchers duplicated a sample across a series of studies of multi-year outcomes, for example by reporting first-year results in a preliminary report, and repeating in later reports (along with the outcomes for the second and subsequent years of implementation) analyses of the first-year sample as originally presented, or as the remaining longitudinal sample. In the first two situations, we accepted the first or main analysis of the sample and rejected subsequent reports of duplicate samples: the study with the earliest date, whether published or unpublished, was used for analysis. In the third situation where longitudinal samples were involved, we used only the most recent outcomes for a given sample of students. In this way, we focused on the achievement effects from the longest exposure to the model by the school and students.[9]

Samples were further duplicated when results were reported for both a full student sample and for some clearly defined subsample, such as for a separate racial/ethnic group or for those who were low-achievers at baseline. In these cases, only the full sample was included for our main analyses. These samples best supported our analysis of the schoolwide effects of CSR. The final way in which independence of observations was threatened involved multiple outcomes within a single achievement domain (e.g., language arts) or across two or more achievement domains (e.g., reading and math) for a distinct sample of students. These situations were resolved by taking the mean effect size across all outcomes and/or domains for the main analysis. For example, if the same student sample had outcomes for reading comprehension, reading vocabulary, and math, the mean effect size across the three areas served to represent a single effect size for that sample. For our subanalyses of the outcomes for the separate subject areas, effect sizes for the various achievement domains were disaggregated and were estimated independently as subject-specific CSR effects.

Characteristics of the Selected Studies

From the 232 studies that met all inclusion criteria, 1,111 independent observations were defined. Each of the 1,111 observations was for a distinct CSR model and sample of students from which an effect size was computed. The school was the primary unit of analysis for the meta-analytic findings. It was selected because CSR is designed to affect whole schools and because schools were typically the unit of analysis reported in the primary studies. Key contextual characteristics, including the poverty level and years of CSR implementation, were also school-level features that we hoped to explore as predictors or moderators of effect size.

Reported within-school student sample sizes varied considerably, though. For example, some studies reported achievement data for an entire school, other studies reported data for a single grade level within a school, and still others reported data for a smaller sample of students within a grade level or school. As a result of these differences, we chose to weight all observations based on the student sample. Table 1 presents the number of studies, observations, and treatment and control students involved in the evaluations of all 29 CSR models. This table and Tables 2 and 3 summarize, respectively, the methodological characteristics of the studies and the coded attributes of each of the CSR models. The tables, which list the reforms alphabetically, reveal the diversity of the reform models and studies in the meta-analysis.

The contextual characteristics presented in Table 1 reveal that the number of studies and observations varied widely by reform model, from a low of one study with one observation for Audrey Cohen to a high of 49 studies with 182 outcomes for Direct Instruction. The median number of studies was four and the median number of observations was 23. Overall, these studies involved 145,296 students participating in the CSR schools and 77,660 comparison students. The mean years of implementation across all reforms and studies was 2.96, and, on average, 65.06% of the students in the CSR schools were eligible for the free or reduced price lunch program.

Methodological characteristics are presented in Table 2. Nearly half of the outcomes were derived from one-group pretest-posttest study designs. Nearly half of the observations were from studies conducted by the developers, and about one third were from studies using true longitudinal sample designs. Outcome data were presented as means for most observations, followed by categorical data, and mixed outcome data. Less than 1% of the outcomes relied on correlational data. About three of four outcomes were based on elementary school samples.

The CSR model attributes presented in Table 3 show that there is considerable variety among the 29 models in terms of their general characteristics and the components that they require in typical implementations. For example, 10 of the 29 reforms required specific curriculum materials (34%), and 12 required specific instructional practices (41%). Forty-five percent required a 75% faculty vote; 21% required a parent involvement program; 38% required student assessments and benchmarks; and 34% required ongoing professional development. More than half of the models received the highest rating for ongoing technical support. The number of replication sites varied widely from a low of 15 schools to a high of 1,800. First-year, worst-case scenario costs also varied widely: for personnel, from no cost to $208,361 for Roots & Wings and Success for All, with a median of $13,023; and for non-personnel costs, from $14,585 for Accelerated Schools Project to $780,000 for Montessori, with an overall median of $72,926. Edison Project was assigned the median values for personnel and non-personnel costs because this reform works within a school’s given budget.[10]

RESULTS

Computation of Effect Sizes

Differences in the nature of the outcome data required nine separate methods for computing effect sizes. The nine methods were of three general types: 1) those that used means and standard deviations (six); 2) those that used frequency distributions (two); and 3) those that used correlations (one). For the first and second types, there was a further distinction between effect sizes based on treatment-control comparisons or one-group pre-post designs.

Click here for Tables 1, 2, and 3

The nine different formulas were all algebraically equivalent, and yielded estimates of the standardized mean difference or common effect size index known as Cohen's *d* or Hedges's *g* (Lipsey & Wilson, 2001). This equivalence was of importance, as we intended to combine the effect estimates from the various formulas in our analyses. Three of the six means-based effect size calculations relied on variations of the common formula

$$d = (M_{\mathrm{T}} - M_{\mathrm{C}})/\sigma,$$

where $(M_T - M_C)$ is the difference between the CSR participants' and non-participants' group means, and σ represents the pooled standard deviation. A variation of the formula for *d* involved adjusting for group differences on the pretest. If the two groups were shown to be similar at pretest, or there was some other statement of pre-intervention similarity, or the posttest group means were presented in the report as having been adjusted for pretest differences, then we simply used this common formula. For cases where there were pretest differences between participants and non-participants, but adjusted posttest means were not presented, we adjusted the posttest means ourselves using the pretest group means and the correlation between pretest and posttest.[11]

A second variation of the formula for *d* used participants' and non-participants' gain scores as estimates of means. If a comparison-group design was not used, another variation of this basic formula utilized only the participants' mean gain score in the numerator. In this variation, the participants' pretest in effect serves as the comparison. For both of these variations, the denominator was the pooled or population standard deviation on the posttest itself and not the standard deviation of the gain scores. Finally, three other methods for calculating an effect size used the test statistics *t* and *F* or used a *p* value when the actual group means were not presented in the study.

We used two methods for calculating effect sizes based on categorical outcomes. When results from a χ^2 analysis with $df = 1$ and total sample size (*N*) were presented, we used these data to estimate an effect size directly. In other cases, we approximated an effect size (*d*) based on an arcsine transformation of the proportion (*p*) of successes for each group

$$d = arcsine\ (p_1) - arcsine\ (p_2).$$

Lipsey and Wilson (2001) stated that the arcsine transformation generally produces a more conservative estimate than the probit transformation and suggested that if effect sizes based on frequency distributions are to be included with other effect sizes based on means and standard deviations, as in the present research, a sensitivity analysis should be conducted to determine which to use.

Our sensitivity analysis showed that the arcsine and probit transformations produced similar overall means, but the probit transformation produced longer tails at both ends of the effect size distribution. Furthermore, the effect sizes based on a calculation of means and standard deviations from the actual grouped frequency distributions produced much higher estimates of *d* than either the arcsine or the probit transformation. For these reasons, we used the arcsine transformation for the cases where the outcome variable was non-continuous.

The final method of effect size calculation used correlational data and applied only one formula, which used the correlation between group membership and the outcome variable. Again, this formula produced an effect index that was algebraically equivalent to an effect size based on means and standard deviations.

Computation of Variance Components, Weights, and Weighted Effect Sizes within a Random-Effects Model

From the outset, it was presumed that a random-effects model was most appropriate for the analysis of CSR effects for two reasons. First, the large number of potential methodological, programmatic, and contextual moderators, which were outlined earlier in the introduction, underlies the concept of a study's true effect size as random (Raudenbush, 1994). Second, this set of potential moderators was not considered to be exhaustive. The qualities of instruction in the schools and the characteristics of local implementations, among other program attributes, were all assumed to contribute to the variation in the estimated effect sizes. Thus, it was hypothesized that various reforms, across programs and schools, would not yield the same fixed population effect.

To test whether the true effect size varied, in addition to the variability introduced by sampling variance, or estimation variance, a homogeneity test of the weighted effect-size estimates was performed. Because the value of 10,777.03 for the homogeneity test statistic, *Q*, exceeded the upper-tail critical value of χ^2 at 1110 degrees of freedom ($p < .001$), the observed variance of the effect sizes was significantly greater than that which would be expected by chance if all observations shared the same population effect size. This statistical test confirmed the *a priori* assumption of a random-effects model specification.

The random-effects variance estimates, v^*_i, for the effect sizes for control group comparisons were computed based on the formulas

$$v_i = ((N_T + N_C) / (N_T * N_C)) + (d^2 / (2(N_T + N_C)) \text{ and}$$

$$v^*_i = \sigma_\Theta^2 + v_i,$$

where v_i represents the within-study variance component, and σ_Θ^2 is the between-studies or population variance component, which was calculated based on the method-of-moments procedure explained by Raudenbush (1994). Given that there were no control students for the one-group, pretest-posttest outcomes, the variance formulas were

$$v_i = (1/ N_T) + ((d^2)/(2* N_T)) \text{ and}$$

$$v^*_i = \sigma_\Theta^2 + v_i.$$

Finally, the formula for the computation of the weights, for each observation, *i*, under the assumptions of a random-effects model was

$$w_i = 1/v^*_i.$$

Distribution and Measures of Central Tendency for Effect Size

Our analysis of the effect size data began with an inspection of the distribution of the 1,111 unweighted effect sizes. Applying Tukey's (1977) definition, we identified as statistical outliers any effect sizes that were more than three interquartile ranges above the 75th percentile or below the 25th percentile. Of the 1,111 independent observations, 19, or 1.8%, met this definition.

Similarly, we identified statistical outliers from the distributions of treatment and control sample sizes, with 132, or 12%, of the 1,111 independent treatment student samples meeting the Tukey (1977) criterion for statistical outliers. Of the control sample sizes, 75, or 13%, of the 598 independent samples met the criterion.

Statistical outliers may exert an overly strong influence on the results. Outliers on the dependent variable, effect size, are especially problematic, but outliers on sample size also are of concern. Because sample size plays an important role in weighting each effect size, unusually large samples may have an exceedingly large influence on the outcomes of our analyses. Therefore, we chose to Winsorize both effect sizes and sample sizes that were statistical outliers. In both cases, we set the value for the effect size or sample size to equal the value at three interquartile ranges beyond the 75th percentile or below the 25th percentile. Because some observations had multiple outlier values on these three variables, only 153 cases (13.7%) were involved in the Winsorizing. The 153 Winsorized cases were spread across 20 of the 29 reforms.[12]

Based on the 1,111 unweighted mean effect sizes, an overall weighted effect size was computed. The unweighted average of the 1,111 effect sizes was .15 and the overall weighted value for *d* was also .15. The average weighted effect size, which is equivalent to a pre-post gain or CSR-control difference of 3.16 NCEs, was greater than 0, $Z = 13.11$, $p < .001$. The standard error of the weighted effect size, which is the square root of $v_{.}^{*}$, was .01. This standard error was employed to calculate a 95% confidence interval for the average weighted effect size. The calculation resulted in a confidence interval of .13 to .18, or 2.74 to 3.79 NCEs. However, as Shadish and Haddock (1994) warned, due to the heterogeneity of the effect estimates, the average weighted effect size should not be interpreted as an estimate of a single population effect parameter, but rather simply as describing the mean of the 1,111 observed effect sizes.

Regression Analysis of Weighted Effect Sizes on Mediating Variables

To explain the heterogeneity of the effect sizes, we performed a modified weighted multiple regression analysis using an SPSS macro, METAREG.SPS, provided by Lipsey and Wilson (2001). This macro modifies the output that would result from a regular weighted least squares multiple regression and provides the correct standard errors, confidence intervals, and significance tests for meta-analysis. The modified weighted least squares multiple regression analysis for random effects was performed using weighted effect size as the dependent measure and the moderator variables as predictors. As explained previously, an estimate of the residual variance component was computed as the random-effects variance plus the estimation variance, and weights were defined by the reciprocal of the residual variance component. Table 4 presents the results of the regression analysis.

Click here for Table 4

All moderator variables accounted for 8% of the variance in the weighted effect sizes. Full descriptions of the variables entered into the regression model are provided in the Method section. First, the comparison group indicator contrasted those observations based on a single-group pre-post design to observations that were based on quasi-experimental non-equivalent control-group designs and true randomized designs. The positive coefficient indicated that the one-group comparisons yielded relatively larger mean effect sizes. The magnitude of the coefficient suggested that, after controlling for the other variables in the model, comparisons using control groups produced effect size estimates .08, or about 1.7 NCEs, less than estimates generated by one-group, pre-post analyses of treatment effects.

Second, as expected, the model indicated that effect sizes produced by developers' evaluations were greater than those produced by other researchers' evaluations. The coefficient suggested that, after statistically taking into account the other moderators, evaluations by developers produced effect size estimates .16, or 3.4 NCEs, greater than those produced by external evaluations. Third, use of a longitudinal sample produced a larger effect size than use of other sample types, about .09 greater. This suggests that when researchers measure CSR effects over time on the same longitudinal sample of students the results tend to show stronger achievement effects than when researchers track effects across successive cohorts of students. Fourth, those outcomes that were estimated with imputed standard deviations had smaller effect sizes than those that were based on actual, reported standard deviations.

Finally, only one reform model component was a statistically significant predictor of effect size, and the relationship was in an unexpected direction. Namely, models that required a component designed to involve parents in school governance and improvement had smaller effects on student achievement than models that did not require this form of parent participation.

School Poverty Level and Years of Implementation as Moderators of Effect Size

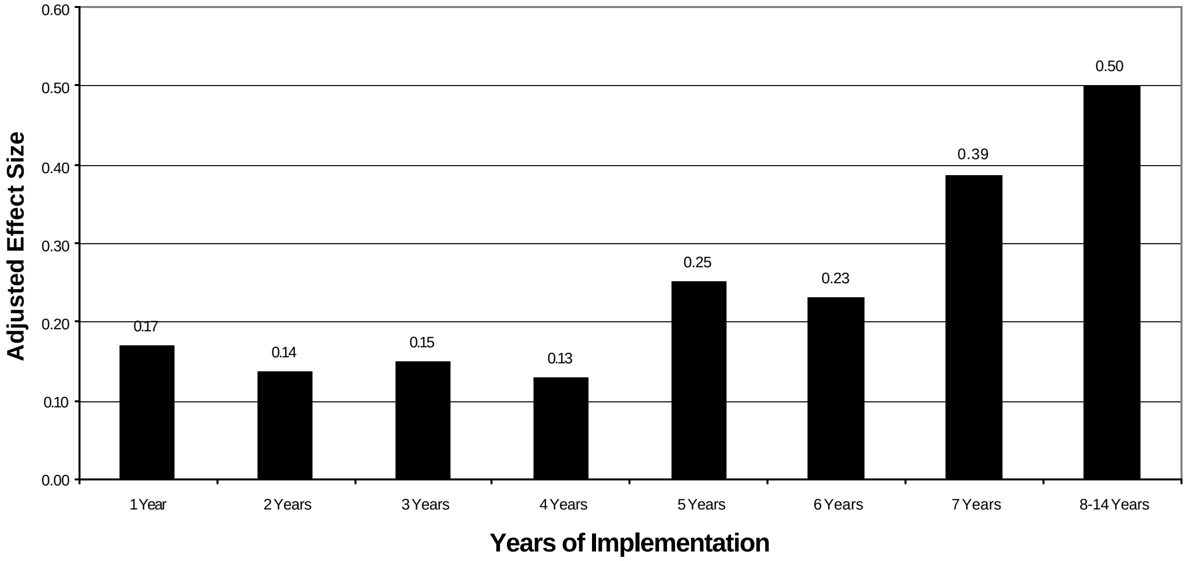
A subset of studies had complete data indicating the CSR school's free or reduced-price lunch participation rate. Of the 1,111 independent observations, 630 (57%) had complete data indicating the poverty level of the CSR school. Similarly, a subset of 975 of the 1,111 observations, or 88%, had complete data indicating the number of years that the CSR model had been implemented at the school.

After regressing weighted effect size on the methodological moderator variables, we obtained the residuals from the regression and added the mean weighted effect size to each observation. In this way, we calculated effect sizes that were statistically adjusted for all of the methodological variables. These adjusted effect sizes became the outcome measures for our subanalyses of the relationship between school poverty and years of implementation and CSR effects.

The weighted regression model using poverty level to predict adjusted effect size revealed that a school's poverty level was not a statistically significant predictor of effect size ($Z = .12$). In other words, across the range of school poverty levels, which tended to be relatively high, CSR was equally effective in relatively lower- and higher-poverty schools.

In a separate weighted regression model, years of implementation was a statistically significant predictor of effect size, with a coefficient of .02 ($Z = 2.82$, $p < .01$). Figure 1 displays the relationship between years of implementation and effect size. This figure shows that the CSR effect size, .17, was relatively strong during the first year of implementation. During the second, third, and fourth years of implementation, though, the effect declined slightly but, essentially, remained the same. After the fifth year of implementation, CSR effects began to increase substantially. Schools that had implemented CSR models for five years showed achievement advantages that were nearly twice those found for CSR schools in general, and after seven years of implementation, the effects were more than two and half times the magnitude of the overall CSR impact of $d = .15$. The small number of schools that had outcome data after 8 to 14 years of CSR model implementation achieved effects that were three and a third times larger than the overall CSR effect.

Figure 1. Adjusted Effect Size by Years of Implementation



Analysis of Subject Area as Moderator of Effect Size

A different level of aggregation of the outcome data was used to analyze the effects for different subject areas. In previous analyses, to retain independent samples of students, we took the mean outcome for students tested across more than one area. For instance, in studies of students attending a CSR school who took both math and reading tests, we aggregated the effects across both subjects and generated a single effect size for the student sample. Our analyses by subject area, though, maintained independence of observations by analyzing the effects in each subject area separately.

All cases had information regarding the subject area evaluated, though some cases presented data for mixed subjects or for more general achievement outcomes. In all other ways, the database used in this analysis was similar to those used for the main analyses and for the subanalyses of school poverty and years of implementation.

The data for these analyses included 1,017 independent samples for reading, 679 for math, 229 for science, 138 for social studies, and 95 cases that could not be grouped into the other subject areas, either because the original research reported results with subjects grouped, or because the achievement test was more general in focus. With a mean effect size of .13 (Z = 10.81, p < .001) for reading, CSR had a statistically significant effect that was somewhat lower than the effect size found for CSR overall. The CSR effect size for math was essentially the same as the overall CSR effect, and slightly higher than the effect for reading, d = .15 (Z = 9.86, p < .001). The CSR effect on science outcomes was somewhat lower than the effects for math and reading, d = .09 (Z = 3.79, p < .001), but was also statistically significant. CSR did not have a statistically significant effect (Z = 0.72) on social studies outcomes. Finally, the cases with outcome data for the general subject area revealed a relatively large CSR effect, d = .20, but also a high standard error (.05) and a wide 95% confidence interval, d = .10 to d = .31. This confidence interval, though, did not include 0 and the result was statistically significant (Z = 3.86, p < .001).

Evidence of Effectiveness for the 29 CSR Models

Table 5 presents the weighted mean effect size, d, the associated significance test, Z, and 95% confidence intervals, which represent the expected range of effects, separately by CSR model. There are three sets of columns in Table 5. The set of columns farthest to the left displays all available evidence concerning the achievement effects of each of the 29 models, regardless of the nature or quality of the study designs. The second set of columns presents results for only those cases that used some form of control group, and the final set of columns shows results for only those cases that were third-party, control-group studies. The latter two, more restrictive presentations of the data provide the best evidence for evaluating the effects of the reform models, in that our prior regression analysis demonstrated that studies performed by the developer and those that used one-group pre-post designs yielded potential biases relative to third-party and control-group comparisons.

The names of the CSR models are listed along the left hand side of Table 5 and are grouped into four categories:

- *Strongest Evidence of Effectiveness*;
- *Highly Promising Evidence of Effectiveness;*
- *Promising Evidence of Effectiveness;* and
- *Greatest Need for Additional Research*

The four categories were established based on a combination of three criteria:

1. *Quality* of the evidence: Does the CSR model have research evidence from the highest-quality studies: control-group studies and third-party control group studies?

2. *Quantity* of the evidence: Does the CSR model have a relatively large number of studies and observations from which one may generalize the findings to the population of schools in the U.S. that are likely to adopt and implement CSR models? (For instance, we used 10 or more studies overall and 5 or more third-party control-group studies as the, arguably arbitrary, standards necessary to be in the top category).

3. *Statistically significant and positive* results: Does the evidence from control-group studies show that the effects of the reform on student achievement are positive and statistically greater than 0?

The notes to Table 5 provide more detailed information about the criteria used to evaluate the quantity of evidence for each of the four categories. Within each of the four categories, the models in Table 5 are listed alphabetically. More information regarding the nature of the reform models along with narrative descriptions of the supporting research base for each may be found in Appendix B.[13]

Strongest Evidence of Effectiveness. CSR models in this category include those that have a large number of studies and observations from schools and students across the United States, such that their outcomes have been replicated in a number of contexts and are reasonably generalizable to the population of U.S. schools that are likely to adopt and implement CSR models. These models also have *statistically significant and positive* achievement effects based on evidence from studies using comparison groups or from third-party comparison designs. Three reforms—Direct Instruction, School Development Program, and Success for All—met the criteria for this category.

Direct Instruction has an overall effect size of $d = .21$ ($Z = 11.61$, $p < .01$), with a 95% confidence interval of $d = .17$ to $d = .25$. The confidence interval expresses the degree of accuracy of the effect size estimate and suggests a range of effects that are likely to be found in similar implementations and studies of the reform model. In this case, similar implementations and studies of Direct Instruction are likely to reveal effects between $d = .17$ and $d = .25$. The effects for Direct Instruction estimated from comparison and third-party comparison designs were somewhat lower than the overall effects, but still positive and statistically significant, $d = .15$ ($Z = 8.40$, $p < .01$) and $d = .15$ ($Z = 7.82$, $p < .01$), respectively.

The School Development Program is another model meeting the highest standard of research evidence, with an overall effect size of $d = .15$ ($Z = 5.48$, $p < .01$) and a 95% confidence interval of $d = .10$ to $d = .20$. As with Direct Instruction, the effect of the School Development Program drops considerably when looking at effects only for comparison or third-party comparison studies: $d = .05$ ($Z = 1.57$, n.s.) and $d = .11$ ($Z =3.23$, $p < .01$), respectively.

Click here for Table 5

Success for All is the third model in the *Strongest Evidence of Effectiveness* category, with an overall effect size of d = .18 (Z = 16.57, p < .01) and a 95% confidence interval of d = .16 to d = .21. The effects are essentially the same when considering only Success for All comparison studies, d = .18 (Z = 15.32, p < .01), as most Success for All evaluations use a comparison group design. The effect estimate from Success for All third-party comparison studies, d = .08 (Z = 5.08, p < .01), is considerably less but still statistically significant.

Highly Promising Evidence of Effectiveness. Models in this category are those that had positive and statistically significant results from comparison or third-party comparison studies, but did not have research bases that were as broad and generalizable as those of the models that met the highest standard. Three reform models met the criteria for this category: Expeditionary Learning Outward Bound, d = .19; Modern Red Schoolhouse, d = .26; and Roots and Wings, d = .38.

Promising Evidence of Effectiveness. Models meeting this standard of evidence were reforms that had more than one study, but still too few to generalize from their results with confidence. All of these CSR models, though, had statistically significant positive effects from comparison or third-party comparison studies. The reforms in this category were: Accelerated Schools, with an overall effect size of d = .09; America's Choice, with an effect size of d = .22; Atlas Communities, d = .27; Montessori, d = .27; Paideia, d = .30; and The Learning Network, d = .22.

Greatest Need for Additional Research. The Greatest Need for Additional Research category included reforms with only one study or those that did not have evidence of statistically significant positive achievement effects from comparison or third-party comparison studies. Seventeen of the 29 CSR models fell into this category. Nearly all of the reforms in this category were there because too few studies have been done to establish statistically reliable and generalizable results. Four of the 17 models had no evidence from either comparison or third-party comparison studies, and another four models lacked evidence from third-party comparison studies. Finally, four CSR models had only a single effect estimate from both comparison and third-party comparison studies. On the other hand, though, there are a number of models, including the Center for Effective Schools, Community for Learning, Co-Nect, Core Knowledge, MicroSociety, Onward to Excellence II, and Talent Development High Schools, that have promising early data but need several more rigorous evaluations to establish a stronger research base.

Two CSR models in this category presented unusual cases that are worthy of discussion. First, the High Schools that Work model has a large research base, composed, almost entirely, of one-group pre-post evaluations performed by its developer. The magnitude of the effect size from these studies, d = .30, is relatively large but the effect size from the one comparison-group study of High Schools that Work actually revealed a statistically significant *negative* effect of the model, d = -.06. This model has been widely replicated and studied and, in many ways, appears to be a promising high school intervention. That the model has been replicated with such success, has been so well supported by the developer, and has accumulated a large number of one-group pre-post evaluations are, indeed, laudable accomplishments. For many schools, this type of evidence may be sufficient to convince decision makers that the model is worthy of adoption. To meet even higher standards of research evidence, though, more research using control groups is needed to help more clearly establish the model's apparent benefits.

Second, though only five studies of the Edison Project have been conducted, they have evaluated the reform in a large number of schools. Taking all of the evidence, Edison appeared to have a statistically significant positive effect size, *d* = .06. When examining the reports of third-party evaluators using comparison groups, though, the results revealed a statistically significant *negative* effect, *d* = -.13. Again, additional studies using comparison groups are needed, from both the developer and from third-party evaluators, to help reconcile these differences.

DISCUSSION

CSR and the CSRP are at the forefront of the national movement to base educational policy and practice on solid research evidence. The recent reauthorization of the Elementary and Secondary Education Act of 1965 and the federal government's single largest investment in America's elementary and secondary schools, the *No Child Left Behind* Act, have similarly required practices based on high-quality research for everything from the technical assistance to schools to the choice of anti-drug-abuse programs. Like a mantra, the *No Child Left Behind* Act repeats phrases such as "scientifically based research" more than 100 times (Olson & Viadero, 2002). This legislation, urging the use of research-based educational practices and procedures in schools receiving federal CSRP and Title I funding, has the potential to revolutionize school improvement in some of the most challenging contexts in the United States.

Does the quantity and quality of the CSR literature provide the scientifically based evidence needed to identify the proven programs and practices that these new policies demand? Our research has sought to understand the CSR research base in various ways. We have described the overall characteristics of the diverse literature; we have identified its biases; and we have empirically established the best evidence that researchers, policymakers, and practitioners should apply to understanding the effects of CSR models. We have estimated the overall effects of the most widely used, nationally disseminated, externally developed CSR models and have gained insight into the overall effects of CSR as a national policy movement. We have also established that there is considerable variation in these effects that is explained by the models themselves, methods used in evaluating the models, and the circumstances in which the programs were implemented. Looking across the 232 studies of CSR and our various analyses of them, the evidence supports six primary findings.

Characteristics of the CSR Research Base

First, *CSR is still an evolving field and there are clear limitations on the overall quantity and quality of studies supporting its achievement effects*. Only 12 reform models are supported by five or more studies of their achievement effects. Only four models have been the subject of five or more third-party studies that used comparison groups. Nearly half of the analyses of CSR effects have been performed by the developers, and about half of the analyses have used some type of quasi-experimental control group. Only seven studies of three CSR models, or about 3% of all studies of the achievement effects associated with CSR, have generated evidence from randomized experiments. These reform models and studies include: the School Development Program (Cook, Habib, Phillips, Settersten, Shagle, & Degirmencioglu, 1999; Cook, Hunt, & Murphy, 1999); Direct Instruction (Crawford & Snyder, 2000; Grossen & Ewing, 1994; Ogletree, 1976; Richardson, Dibenedetto, Christ, Press, & Winsbert, 1978); and Paideia (Tarkington, 1989). In addition to these shortcomings, many of the studies did not present sufficient detail to allow for replication of the findings. For instance, substantial numbers of reports contained no information about student sample sizes and did not provide standard deviations for the outcome measures.

Many of these problems, though, are to be expected given the recent emergence of CSR, in general, and many of the CSR models, in particular. Some models are at an early stage of program development that has not yet demanded third-party evaluations and more costly and difficult control-group comparisons. On the other hand, there are some models that have had relatively long histories, have been replicated in many schools, and should have accumulated this evidence. Still other CSR models are on their way to establishing a strong research base. Three models, in particular, have accumulated enough evidence to meet our highest standard of research evidence. Taken as a whole, there is a sufficient number of reasonably high-quality studies of CSR to evaluate its overall effects and to inform policy.

Overall Effects of CSR

Second, *the overall effects of CSR are statistically significant, meaningful, and appear to be greater than the effects of other interventions that have been designed to serve similar purposes and student and school populations. Overall, students from CSR schools can be expected to score one eighth of a standard deviation, or 2.5 NCEs, higher on achievement tests than control students in non-CSR schools.* Our various analyses suggest that students attending CSR schools can be expected to score between nearly one-tenth and one-seventh of a standard deviation, or between 1.9 NCEs and 3.2 NCEs, higher than control students on achievement tests. The low-end estimate represents the overall effect size of d = .09 for third-party studies using comparison groups and the high-end estimate represents the effect size of d = .15 for all evaluations of the achievement effects of CSR. Using a metric devised by Cohen (1988), U_3, the effect size of d = .12 for all studies using control groups tells us that the average student who participated in a CSR program outperformed about 55% of similar control children who did not attend a CSR school.

How should we interpret this overall effect? Cooper (1981) has suggested a comprehensive approach to effect size interpretation that uses multiple criteria and benchmarks for understanding the magnitude of the effect. First, and most generally, we may compare the overall CSR effect size to Cohen's (1988) definitions of a small effect within the behavioral sciences, d = .20, and a large effect, d = .80. Second, and more specifically, Cohen (1988) pointed out that the relatively small effects of around d = .20 were most representative of fields that are closely aligned with education, such as personality, social, and clinical psychology. Similarly, Lipsey and Wilson's (1993) more recent compendium of meta-analyses concluded that psychological, educational, and behavioral treatment effects of modest values of even d = .10 to d = .20 should not be interpreted as trivial.

Finally, and even more specifically, how do CSR effects compare to previous national efforts to help close the achievement gap and improve the outcomes of large numbers of high-poverty and low-achieving students and schools? The most obvious comparison to the effect of CSR is the effect of traditional Title I programs, which have historically funded targeted remedial interventions, such as pullout programs, and schoolwide programs designed to assist at-risk students. These programs were the subject of Borman's and D'Agostino's (1996) meta-analysis of the achievement effects of Title I programs, which synthesized the results from all federal evaluations conducted between 1965 and 1994. During these years, rather than schoolwide programs and CSR models, the primary methods for upgrading the educational programs of at-risk children were through specialized pullout programs and other targeted assistance. Borman and D'Agostino estimated that the average effect size associated with these efforts was d = .11. The Title I evaluations, though, were almost exclusively based on the less-preferred one-group pre-post design and may overestimate the true Title I effect. Borman and D'Agostino did make an adjustment for regression to the mean effects for all Title I outcomes from one-group pre-post designs. The comparison to this benchmark is suggestive, but because the primary studies and meta-analyses used different methodologies than those reported here, the comparison is imperfect.

A better comparison between CSR and conventional Title I programs may be drawn directly from the current study by estimating the CSR effect size from comparison-group studies in schools of 50% poverty or more. In most of these cases, the comparison schools have such high poverty rates that it is highly likely that they received federal Title I funds. In most cases, these schools implement Title I targeted or schoolwide programs and, in most cases, are not implementing other CSR models. These studies, therefore, provide a relatively good indication of the value-added effects of CSR, above and beyond the effect of traditional Title I programs. Across 346 such comparisons, the effect size, adjusted for methodological characteristics, was d = .12. In other words, despite the fact that the vast majority of these control schools provided their students with extra resources and programs provided through Title I, the average student from a CSR school still outperformed 55% of the children from the Title I schools.

These comparisons and our analyses of the overall effects of the CSR models are valuable for understanding general outcomes. These overall effects, though, are highly variable and should be viewed as averages found across a wide array of reform models and schools that were evaluated in a variety of ways. The overall effect size is a good indicator of the expected effects of CSR across a large number of schools. For instance, we can say with some confidence that policymakers may expect to find CSR effects of between d = .09 and d = .15 across similar studies of national or large district-wide samples of CSR schools. The effects for individual schools and the effects for individual reform models are likely to vary more widely. Our regression analysis and the specific effects of the 29 reform models reveal many reasons for the diverse findings, but a considerable amount of variability is left unexplained.

Explaining Differences in CSR Effects

Third, *the heterogeneity of the CSR effect and the fact that few of the general reform components helped explain this variability suggests that the differences in the effectiveness of CSR are largely due to unmeasured program-specific and school-specific differences in implementation.* Our regression analysis suggested that whether or not a CSR model, in general, requires the following components explains very little in terms of the achievement outcomes the school can expect: a) ongoing staff professional development; b) measurable goals and benchmarks for student learning; c) a faculty vote to increase the likelihood of model acceptance and buy-in; and d) the use of specific and innovative curricular materials and instructional practices designed to improve teaching and student learning. Similarly, the frequency with which the CSR models have successfully replicated their approaches in schools with diverse characteristics; the overall level of external technical support and assistance from the developer, and the general cost of the model do not help us explain a substantial amount of the variability in the CSR effect.

The one reform attribute that was a statistically significant predictor of effect size suggested that CSR models that require the active involvement of parents and the local community in school governance and improvement activities tend to achieve worse outcomes than models that do not require these activities. Taking strong actions to encourage parents to play significant roles in school governance and reform may help the school grow as an institution, but these activities are not likely to have strong impacts on student achievement (Epstein, 1995). In contrast to school-based efforts aimed at helping families enrich their children’s learning opportunities outside of school, which are far more likely to help individual children succeed with specific academic goals, the focus on parent involvement in school governance could sidetrack schools if the immediate priority is to improve student achievement.

The general lack of explanatory power for the required reform characteristics suggests at least two possible interpretations. The first is that these components are not important for promoting student achievement in CSR schools and, therefore, there is no relationship. The second interpretation is that knowing whether or not a CSR model required schools to implement a given component tells us little about whether or not the component actually was implemented. This latter interpretation suggests that some or all of these components may make a difference in terms of student achievement, but school-specific and model-specific differences in the ways that the components are actually implemented explain considerably more than simply knowing whether or not the CSR developer requires them. Consistent with research that has linked the success of school reform to the level and quality of implementation (Berman & McLaughlin, 1978; Crandall et al., 1982; Datnow, Borman, & Stringfield, 2000; Stringfield et al., 1997), the coordination and fit of the model to local circumstances, and the relationship between the CSR developer and the local school and school district (Datnow & Stringfield, 2000), we contend that knowing more about these largely unmeasured and unreported differences in implementation, across both schools and CSR models, would also enrich our understanding of the variability in the CSR effects.

Fourth, *rather than the general programmatic components of the CSR models, the methodological differences across the studies themselves tell us far more about the effects that we* *could expect to find.* Studies performed by the developer yielded considerably stronger effects than studies performed by others. Does this suggest, as Pogrow (2000) and others have implied, that the developers, to use a metaphor, have their thumbs on the scale and are consciously manipulating the evaluation to make the outcomes appear more favorable? This interpretation may have some merit in a few cases, but is probably not a reasonable explanation of the overall trend. Perhaps equally likely is that some third-party researchers may seek to taint a model due to a personal grudge or professional dislike for its particular orientation.

A better explanation for the stronger outcomes we find for the developers' studies is that when developers are more actively involved in the study of their models, they are also more likely to be actively involved in studying a high-quality implementation. After all, why would developers want to study half-hearted implementations of their models? Many of the studies performed by developers may represent what Cronbach et al. (1980) termed the "superrealization" stage of program development. Before broad field trials, interventions are often studied under optimal conditions as assessments of what the program can accomplish at its best. The extent to which the developers' studies and results may generalize across broader implementations of their CSR models, though, is of some concern.

The second key methodological finding was that studies using a one-group, pretest-posttest design produced larger effect sizes than studies using control groups. This is a clear methodological bias that should be addressed in future CSR research. Ideally, evaluations should include randomized designs, which assign schools at random to CSR and control conditions. As Borman (in press) pointed out, innovations should not be forced on schools through random assignment. Schools should be partners in the process of experimentation and should be supportive of the CSR model under study. The only clear trade-off in such studies is that some schools will receive the innovation now and others assigned to the control condition will receive the program later, if it proves to be worthwhile and effective.

High-quality, quasi-experimental control-group designs are also desirable. When comparing directly randomized experiments and quasi-experiments that were designed to answer the same research questions, Lipsey and Wilson (1993) found that quasi-experiments are more highly variable in the results that they produce. As a result, although quasi-experiments may be less expensive than true experiments to conduct in the short run, they are less efficient in the long run because one needs many more of them to arrive at the same conclusion as a randomized experiment. If randomized or matched control groups are not possible, even a comparison of the CSR school's outcomes to district averages will provide some understanding of the value-added effects of the model.

Fifth, *the models meeting the highest standard of evidence, Direct Instruction, the School Development Program, and Success for All, are the only CSR models to have clearly established, across varying contexts and varying study designs, that their effects are relatively robust and that the models, in general, can be expected to improve students' test scores.* As the results in Table 5 demonstrate, the outcomes vary considerably by reform model. In most cases, however, the research base for each CSR model is still too small to generate reliable estimates of the models' expected effects. For instance, it is certainly premature to conclude that the Audrey Cohen CSR model is likely to have a negative effect on student achievement of *d* = -.13 when replicated in schools. It is also too early to say that Integrated Thematic Instruction will likely have a relatively strong positive effect of *d* = .24 when implemented in other schools. In some cases, promising and highly promising models are emerging. Expeditionary Learning Outward Bound, Modern Red Schoolhouse, and Roots & Wings are all on the brink of establishing strong research bases. The models meeting the standard for the *Strongest Evidence of Effectiveness* category are distinguished from these models and others by the quantity and generalizability of their outcomes, the quality of this evidence (for instance, six of the seven randomized experiments and many high-quality quasi-experimental control-group studies have been conducted on the models achieving the highest standard of evidence), and the reliable effects on student achievement.

Sixth, turning to contextual differences that we studied, *the number of years of model implementation has very important implications for understanding CSR effects on student achievement. The strong effects of CSR beginning after the fifth year of implementation may be explained in two ways: a potential cumulative impact of CSR or a self-selection artifact.* Specifically, schools may be experiencing stronger effects as they continue implementing the models, or it may be that the schools experiencing particular success continue implementing the reforms while the schools not experiencing as much success drop them after the first few years. Both explanations seem to have some credence. Nonetheless, it is of note that the average school across all studies reviewed had implemented its CSR model for approximately three years. These studies, therefore, may underestimate the true potential of CSR for affecting change in schools and for improving student achievement. Stronger evidence is needed to understand the linkages between years of implementation and school improvement and, ultimately, its impacts on student outcomes.

We explored the significance of two other important contextual variables for understanding differences in achievement outcomes. The poverty level of the school in which the CSR model is implemented and the subject area that is tested for CSR effects do not explain large differences in the observed effects. All schools, regardless of poverty level, appear to benefit from CSR and most subject areas tested reveal similar reform impacts. Because federal funds for implementation of CSR models target high-poverty schools, this finding is of importance. It suggests that the schools from the most challenging high-poverty contexts are benefiting just as much from CSR as are schools from more advantaged circumstances.

CONCLUSION

Historically, teaching has been fraught with what Lortie (1975) called “endemic uncertainties.” Moreover, Cook and Payne (2002) argued that prevailing theories of evaluation and improvement in education suggest that each district, school, or even classroom, is so complex and distinctive that only highly context-specific change strategies are likely to modify and improve their central functions. The scale-up and early success of CSR, which has broadened the use of replicable technologies driven by scientific knowledge, stands in stark contrast to these beliefs about schools, educational change, and evaluation.

The successful expansion of CSR shows that research-based models of educational improvement can be brought to scale across many schools and children from varying contexts. There are adaptations that are sensitive to context—for instance there is a Spanish version of the Success for All program, *Éxito Para Todos*, for English language learners—but the general models of school improvement also include well-founded and widely applicable instructional and organizational components that are likely to work in a variety of situations. The increasing market place of CSR models and the proven replicability of many of the programs are important developments. To further advance CSR, though, policymakers and educators must demand clear evidence that the reforms will make a difference.

The models meeting our highest standard of evidence have been well researched and have shown that they are effective in improving student achievement across reasonably diverse contexts. These models certainly deserve continued dissemination and federal support through CSRP and Title I. All CSR models—even those achieving the highest standard of evidence—would benefit from more federal support for the formative and summative evaluations that are necessary to establish even more definitively what works, where, when, and how. Rather than approving programs on the basis of the 11 requirements (e.g., parent outreach program, clear goals and benchmarks) that make a model "comprehensive," we suggest that schools and policymakers pay even stronger attention to the models' outputs.

Clear research requirements, ample funding for research and development, and a focus on the CSR models' results may support the transformation of educational research and practices in much the same way that it has helped transform medical research and treatment. Like the series of studies required in the Food and Drug Administration's premarketing drug approval process, a similar set of studies might guide the research, development, and ultimate dissemination of educational programs (Borman, in press). Once a CSR program has met a standard of evidence, then its implementation using federal funds, namely those from CSRP and Title I, should be approved. Before programs have accumulated such evidence, some concern should be shown for the ethics of supporting educational programs with unknown potentials. In medicine, only half of the new treatments subjected to randomized clinical trials actually show benefits beyond the standard treatments patients would have received (Gilbert, McPeek, & Mosteller, 1977). Without the benefit of high-quality evaluation, many widely disseminated educational practices have simply wasted the time of teachers and students. Others, including compensatory education pullout programs and tracking, have been regarded by some scholars as counterproductive and potentially harmful.

At the same time, we do not suggest that schools and policymakers dismiss promising programs before knowing their potential effects. Instead, we challenge the developers and the educational research community to make a long-term commitment to research-proven educational reform and to establish a market place of scientifically based models capable of bringing comprehensive reform to the nation's schools. Similar to Donald Campbell's (1969) famous vision of the "experimenting society," we must take an experimental approach to educational reform, an approach in which we continue to evaluate new programs designed to cure specific problems, in which we learn whether or not these programs make a difference, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available.

REFERENCES

Berman, P., & McLaughlin, M.W. (1978). Federal programs supporting educational change, Vol. VIII. Santa Monica, CA: Rand.

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. (2000a). Title I: The evolving research base. *Journal of Education for Students Placed At Risk, 5*, 27-45.

Borman, G.D. (2000b). The effects of summer school: Questions answered, questions raised. *Monographs of the Society for Research in Child Development, 65,* (1, Serial No. 260).

Borman, G.D., & Hewes, G.M. (2001). *The long-term effects and cost-effectiveness of Success for All*. (Report 53). Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk.

Borman, G.D. (in press). Experiments for educational evaluation and improvement. *Peabody Journal of Education.*

Borman, G.D., & D'Agostino, J.V. (1996). Title I and student achievement: A meta-analysis of federal evaluation results. *Educational Evaluation and Policy Analysis, 18*, 309-326.

Borman, G.D., Rachuba, L., Datnow, A., Alberg, M., MacIver, M., Stringfield, S., & Ross, S. (2000). *Four models of school improvement: Successes and challenges in reforming low-performing, high-poverty Title I schools. CRESPAR Report #48.* Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk.

Borman, G.D., Stringfield, S.C., & Slavin, R.E. (2001). *Title I: Compensatory education at the crossroads.* Mahwah, NJ: Lawrence Erlbaum Associates.

Borman, G.D., D'Agostino, J.V., Wong, K.K., & Hedges, L.V. (1998). The longitudinal achievement of Chapter 1 students: Preliminary evidence from the Prospects study. *Journal of Education for Students Placed At Risk, 3,* 363-399.

Borman, G.D., Wong, K.K., Hedges, L.V., & D'Agostino, J.V. (2001). Coordinating categorical and regular programs: Effects on Title I students' educational opportunities and outcomes. In G.D. Borman, S.C. Stringfield, & R.E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 79-116). Mahwah, NJ: Erlbaum.

Campbell, D.T. (1969). Reforms as experiments. *American Psychologist, 24*, 409-429.

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

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

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton-Mifflin.

Cook, T.D., Habib, F.N., Phillips, M., Settersten, R.A., Shagle, S.C., & Degirmencioglu, S.M. (1999). Comer's School Development Program in Prince George's County, Maryland: A Theory based evaluation. *American Educational Research Journal, 36,* 543-597.

Cook, T.D., Hunt, H.D., & Murphy, R.F (1999) Comer's School Development Program in Chicago: A Theory-Based Evaluation. *American Educational Research Journal, 37*, 535-597.

Cook, T.D., & Payne, M.R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch (eds.), *Evidence matters: Randomized trials in education research* (pp. 150-178). Washington, DC: Brookings.

Cooper, H. (1981). On the effects of significance and the significance of effects. *Journal of Personality and Social Psychology, 41,* 1013-1018.

Crandall, D.P., Loucks-Horsley, S., Baucher, J.E., Schmidt, W.B., Eiseman, J.W., Cox, P.L., Miles, M. B., Huberman, A. M., Taylor, B. L., Goldberg, J.A., Shive, G., Thompson, C.L., & Taylor, J.A. (1982). *Peoples, policies, and practices: Examining the chain of school improvement (vols. 1-10)*. Andover, MA: The NETWORK.

Crawford, D.B., & Snider, V.E. (2000). Effective mathematics instruction: The importance of curriculum. *Education and Treatment of Children, 23(*2), 122-142.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., & Weiner, S.S. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements.* San Francisco, CA: Jossey-Bass.

Datnow, A., Borman, G., & Stringfield, S. (2000). School reform through a highly specified curriculum: A study of the implementation and effects of the Core Knowledge Sequence. *The Elementary School Journal, 101*, 167-192.

Datnow, A., & Castellano, M. (2000). Teachers' responses to Success for All: How beliefs, experiences and adaptations shape implementation. *American Educational Research Journal, 37,* 775-799.

Datnow, A., & Stringfield, S. (2000). Working together for reliable school reform. *Journal of Education for Students Placed At Risk, 5*, 183-204.

Edmonds, R.R. (1979). Effective schools for the urban poor. *Educational Leadership, 37*(1), 15-24.

Epstein, J.L. (1995). School/family/community partnerships: Caring for the children we share. *Phi Delta Kappan*, *76*(9), (701-712).

Gilbert, J., McPeek, B., & Mosteller, F. (1977). Statistics and ethics in surgery and anesthesia. *Science, 198*, 684-689.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3-8.

Grossen, B., & Ewing, S. (1994). Raising mathematics problem-solving performance: Do the NCTM teaching standards help? *Effective School Practices, 13*(2), 79-91.

Haynes, N., Emmons, C., & Woodruff, D. (1998). School Development Program effects: Linking implementation to outcomes. *Journal of Education for Students Placed At Risk, 3*, 71-86.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 106-128.

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

Hirsch, E.D., Jr. (1995). *Core Knowledge Sequence.* Charlottesville, VA: Core Knowledge Foundation.

Hirsch, E.D., Jr. (1996). *The schools we need*. New York: Doubleday.

Kearns, D. & Anderson, J. (1996). Sharing the vision: Creating New American Schools. In Stringfield, S., Ross, S., & Smith, L. (Eds.) *Bold plans for school restructuring* (pp. 9-23). Mahwah, NJ: Erlbaum.

Lipsey, M.W, & Wilson, D.B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

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

Lortie, D.C. (1975). *Schoolteacher.* Chicago: University of Chicago 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: Yale University Press.

Murphy, J., & Beck, L. (1995). S*chool-based management as school reform: Taking stock*. Newbury Park, CA: Corwin.

Northwest Regional Educational Laboratory (1998). *Catalog of school reform models: First edition.* Portland, OR: Author.

Northwest Regional Educational Laboratory (2000). *Catalog of school reform models: Second edition.* Portland, OR: Author. Available at http://www.nwrel.org/scpd/catalog/index.html.

Nunnery, J. (1998). Reform ideology and the locus of development problem in educational restructuring: Enduring lessons from studies of educational innovation. *Education and Urban Society, 30,* 277-295.

Ogletree, E. J. (1976). A Comparative Study of the effectiveness of DISTAR and eclectic reading methods for inner-city children. (ERIC Document Reproduction Service No. ED 146544).

Olson, L., & Viadero, D. (2002). Law mandates scientific base for research. *Education Week,* 21(20), pp. 1, 14, 15.

Pogrow, S. (1998). What is an exemplary program, and why should anyone care? A reaction to Slavin and Klein. *Educational Researcher, 27*(7), 22-29.

Pogrow, S. (2000). Success for All does not produce success for students. *Phi Delta Kappan*, 82, 1 67-81.

Puma, M.J., Karweit, N., Price, C., Ricciuti, A., Thompson, W., & Vaden-Kiernan, M. (1997). *Prospects: Final report on student outcomes.* Bethesda, MD: Abt Associates, Inc.

Raudenbush, S.W. (1994). Random effects models. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Richardson, E., Dibenedetto, B., Christ, A., Press, M., & Winsbert, B. (1978). An assessment of two methods for remediating reading deficiencies. *Reading Improvement, 15*(2), 82-95.

Rotberg, I.C., Harvey, J., & Warner, K.E. (1993). *Federal policy options for improving the education of low-income students. Vol. I, findings and recommendations.* Santa Monica, CA: RAND.

Rowan, B. (1990). Commitment and control: Alternative strategies for the organizational design of schools. In C.B. Cazden (Ed.), *Review of research in education* (pp. 353-389). Washington, DC: American Educational Research Association.

Shadish, W.R., & Haddock, C.K. (1994). Combining estimates of effect size. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285-299). New York: Russell Sage Foundation.

Sizer, T.R. (1992). *Horace's school: Redesigning the American high school.* New York: Houghton Mifflin.

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15*(9), 5-11.

Slavin, R.E. (1989). The PET and the pendulum. *Phi Delta Kappan, 70,* 752-758.

Slavin R.E. (1999). Rejoinder: Yes, control groups are essential in program evaluation: A response to Pogrow. *Educational Researcher, 28*(3), 36-38.

Slavin, R.E., & Fashola, O.S. (1998). *Show me the evidence!* Thousand Oaks, CA: Corwin.

Slavin, R., & Madden, N. (2001). *One million children: Success for All*. Thousand Oaks, CA: Corwin.

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

Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S.H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing, Politics of Education Association yearbook, 1990* (pp. 233-267). London: Taylor & Francis.

Stringfield, S. (2000). A synthesis and critique of four recent reviews of whole-school reform in the United States. *School Effectiveness and School Improvement, 11,* 259-269.

Stringfield, S., Millsap, M., Yoder, N., Schaffer, E., Nesselrodt, P., Gamse, B., Brigham, N. Moss, M., Herman, R., & Bedinger, S. (1997). *Special strategies studies final report.* Washington, DC: U.S. Department of Education.

Tarkington, S.A. (1989). *Improving critical thinking skills using Paideia seminars in a seventh-grade literature curriculum.* Unpublished doctoral dissertation, University of San Diego.

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

Traub, J. (1999). *Better by design? A consumer's guide to schoolwide reform.* Washington, DC: Thomas B. Fordham Foundation.

Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

U.S. Department of Education. (2002). *Comprehensive School Reform (CSR) Program Guidance*. Retrieved 9/16/02, from htth://www.ed.gov/offices/OESE/compreform/guidance2000.html.

Walberg, H. & Greenberg, R. (1999). The Diogenes factor. *Phi Delta Kappan, 81,* 127-128.

Wang, M.C., Haertel, G.D., & Walberg, H. (1997). *What do we know? Widely implemented school improvement programs.* Philadelphia: Temple University Center for Research in Human Development and Education.

Wong, K., & Meyer, S. (2001). Title I schoolwide programs as an alternative to categorical practices: An organizational analysis of surveys from the Prospects study. In G.D. Borman, S.C. Stringfield, & R.E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 195-234). Mahwah, NJ: Lawrence Erlbaum Associates.

Wong, K.K., & Meyer, S.J. (1998). Title I schoolwide programs: A synthesis of findings from recent evaluation. *Educational Evaluation and Policy Analysis, 20* , 115-136.

NOTES

1. Initially, we had identified 33 reform models for possible inclusion in this meta-analysis. Four of the models, though, had no quantitative data on their achievement effects from which we could calculate effect size estimates. These four CSR models were Foxfire Fund, League of Professional Schools, QuEST, and Ventures Initiative and Focus System. The 29 models remaining in the analyses were: Accelerated Schools Project, America's Choice School Design, ATLAS Communities, Audrey Cohen College: Purpose Centered Education, Center for Effective Schools, Child Development Project, Coalition of Essential Schools, Community for Learning, Community Learning Centers, Co-NECT Schools, Core Knowledge, Different Ways of Knowing, Direct Instruction, Edison Project, Expeditionary Learning Outward Bound, High Schools That Work, High/Scope Primary Grades Approach to Instruction, Integrated Thematic Instruction, MicroSociety, Modern Red Schoolhouse, Montessori, Onward to Excellence, Paideia, Roots & Wings, School Development Program, Success for All, Talent Development High Schools with Career Academies, The Learning Network, and Urban Learning Centers.

2. This information was obtained from the Southwest Educational Development Laboratory's CSRD database, which is available on-line at: http://www.sedl.org/csrd/awards.html. The data reported here include all schools receiving CSRD awards that began in 1998, 1999, 2000, and 2001. According to the website, the database from which we derived our information was last updated on November 20, 2001. Not all schools reported whether they had been identified for improvement under Title I, state, or local regulations. Therefore, the percentages that we report are, most likely, underestimates.

3. The two catalogs' inclusion criteria were slightly different, but similar to our goal of including models that were nationally disseminated, externally developed, comprehensive school reform. The AIR catalog based its model selection on five criteria: 1) "promoted by their developers as a means to improve student achievement in low-performing schools;" 2) "mentioned by name in federal [CSRD] legislation;" 3) "used in many schools and districts;" 4) "have obtained national visibility in the educational and national press;" and 5) "there is some research evidence about their effects on students and/or implementation in schools." Any reform meeting the second criterion was included automatically, but other reforms had to meet at least three of the other four criteria (Herman et al., 1999, p. 7).

   Models for the first edition of the NWREL catalog were selected through an open application process. Any developer requesting an application packet was sent one. NWREL then chose from among the submitted applications, based on criteria similar to the Herman et al. (1999) set: "Criteria for selecting models included evidence of effectiveness in improving student academic achievement, extent of replication, implementation assistance provided to schools, and comprehensiveness" (NWREL, 2000). The selection process for NWREL's second edition was by invitation only. Developers of models that met criteria of adoption by five or more schools receiving CSRD funds, were nominated by a state or regional lab CSRD manager, or acknowledged by one of several national educational organizations, were asked to submit applications. Submitted applications were then reviewed based on the criteria outlined previously.

4. Despite all efforts to obtain the studies from traditional sources (e.g., libraries and ERIC), the model developers, and the authors of the studies, there were 10 publications that we did not obtain. Because we had no opportunity to review these studies, we were not able to establish whether they would have met our requirements for inclusion in the synthesis. These 10 studies were distributed across the following CSR models: Accelerated Schools Project (1); Co-NECT Schools (1); Direct Instruction (2); High Schools that Work (1); Paideia (1); School Development Program (2); and Success for All (2).

5. Perhaps the most important contextual information, the level or quality of the model's implementation, was rarely provided. This is one of the most important deficits in the research literature on CSR.

6. The separate types of cohort designs initially coded included: a) comparing the outcomes for one grade level (e.g., third graders) in one year to the outcomes for the same grade level (e.g., third graders) in a subsequent year; b) comparing the outcomes for one grade level in one year (e.g., third graders in 1999) to the outcomes for the same student cohort in a subsequent year (fourth graders in 2000); or c) comparing the outcomes for several grade levels (e.g., third through fifth graders) in one year to the outcomes for the same grade levels (third through fifth) in a subsequent year. "True" longitudinal designs are distinguished from all of these in that they track the same sample of students across each time point. In contrast, the cohort designs have different, but often overlapping, samples of students at each time point.

7. To achieve greater consistency between the cost estimates provided by select developers during 2001 and the cost estimates for other models based on data in the Herman et al. (1999) report, we adjusted the latter cost estimates to constant 2001 dollars using gross domestic product implicit price deflators.

8. By assuming the same number of students and teachers for each model, we were able to gain greater consistency in the cost estimates. Nevertheless, the estimated marginal costs of implementing the reform models may vary widely by school, depending on a variety of other factors. Rather than relying on these general estimates to project costs for implementing a reform in a particular school, we suggest contacting the developer directly to obtain specific cost estimates.

9. We did not include long-term effects of the models that are sustained after discontinuation of the program. We confronted one such example for Success for All, which has been shown to have sustained effects through the end of eighth grade (Borman & Hewes, 2001). This analysis, though, estimates the sustained effect beyond the discontinuation of this elementary school program in fifth grade. This type of analysis, though highly important, offers a different type of information than that offered by the other analyses that are the focus of this review.

10. In some cases, Edison pays additional start-up costs that are above and beyond the district's or school's per-pupil allowance. Because these are not marginal costs incurred by the schools or districts, they are not included in our estimate of the cost of implementing Edison.

11. When the correlation between pretest and posttest was not provided, we imputed a pre-post correlation of 0.80. These cases were so few that we did not include a flag to indicate an imputed value.

12. We also ran the analyses with the original non-Winsorized values and obtained similar results. In the regression analysis, there were some minor changes in the magnitudes of coefficients, but not in the direction or level of statistical significance of the results. In the reform-specific analyses, again, all changes were inconsequential, except for three models whose effect size estimates were somewhat larger with the non-Winsorized values. The three models were Direct Instruction, whose estimated effect sizes were 0.06 greater using non-Winzorized values; Paideia, 0.03 greater for all cases and 0.05 greater for comparison-group-only cases and for third-party comparison-group cases only; and Expeditionary Learning Outward Bound, 0.03 greater for comparison-group-only cases.

13. Although four CSR models, Foxfire Fund, League of Professional Schools, QuEST, and Ventures Initiative and Focus System, were dropped from our study for lack of research evidence amenable to analysis, they could be considered among the models for which there is the *Greatest Need for Additional Research*.