Special Issue on

The Elementary and Secondary Education Act at 40: Reviews of Research, Policy Implementation, Critical Perspectives, and Reflections

29

2005

REVIEW OF RESEARCH IN EDUCATION

Chapter 1

National Efforts to Bring Reform to Scale in High-Poverty Schools: Outcomes and Implications

GEOFFREY D. BORMAN
University of Wisconsin at Madison

Education in the United States is a decentralized system composed of highly variable practices, programs, and school contexts. The primary technology of education, teaching, is highly complex and is typically designed and implemented by teachers who have traditionally enjoyed a great deal of autonomy and independence from regular inspection. The principal goals and products of education and whether they should be centered around creativity, knowledge of basic facts, sound moral judgment, or something else are constantly open to differing opinions and debate. Can such a diffuse system with uncertain technology and goals be served by centralized efforts to implement educational reform at scale? Furthermore, how can educational research support the scale up of promising programs and practices?

This chapter discusses some of the ways in which recent national efforts to reform the country's high-poverty schools inform these questions. I begin by tracing the recent history of these reform efforts. In addition to considering the evolution of the interventions themselves, I consider how the national research and evaluation agenda has evolved, and can continue to evolve, to help advance the development of replicable programs and evidence-based educational policy. Finally, applying the lessons learned from two syntheses of the federal Title I and comprehensive school reform research literature, I offer four conclusions for methodologists, policymakers, and reform developers to consider when conducting research, crafting policy, and refining educational programs to support the scale up of educational innovations.

THE RECENT HISTORY OF NATIONAL REFORM EFFORTS

Since the advent of a national effort to improve the United States' most challenged high-poverty elementary and secondary schools, the capacity, technology, and policy to support the scale up of school reform have expanded dramatically. The roots of this movement can be traced back to 1965, when Title I of the Elementary and Secondary Education Act was implemented as a centerpiece of Lyndon B. Johnson's "War on Poverty." The goal was "to provide financial assistance to . . . local educational agencies serving areas with concentrations of children from low-income families to expand

and improve their educational programs by various means . . . which contribute particularly to meeting the special educational needs of educationally deprived children" (79 Stat. 27, 27). Along with the emerging system of social programs of the 1960s, Title I was the major educational initiative designed to close the achievement gap between poor children and their more advantaged peers and, ultimately, break the vicious cycle of poverty.

PROBLEMS IN IMPLEMENTATION

The early years of the nation's efforts to bring reform to scale in high-poverty schools were largely sabotaged by ineffectual policy and a nearly nonexistent knowledge base regarding how to improve schools for the disadvantaged. McLaughlin (1976) noted that the original program mandates were ambiguous concerning the proper and improper uses of Title I funds, and the guidelines and intent of the law were open to varying interpretations. Varying local interpretations of the law rather than clear and uniform federal mandates guided the use of federal funds.

Also, in 1965, the research base and practitioner knowledge base for developing effective compensatory education programs were extremely limited. The majority of local administrators and teachers lacked the experience and understanding for developing, implementing, and teaching compensatory programs. Although research provided some basic descriptions of exemplary practices in select sites (Hawkridge, Campeau, DeWitt, & Trickett, 1969; Hawkridge, Chalupsky, & Roberts, 1968; Wargo, Campeau, & Tallmadge, 1971), there were no clear, replicable programs that could be scaled up to serve large numbers of schools.

Although the federal dollars provided localities an incentive to improve education for the disadvantaged, a viable intergovernmental compliance system was not in place. Without effective regulation, receipt of funds did not depend on meeting the letter or the spirit of the law. Responding to local self-interests, and using Title I dollars for established general aid policies, was an easier option than the new and more complicated task of implementing effective programs for poor, low-achieving students.

Despite early resistance by most federal policymakers to restrictions in local control, the continued misuse of Title I funds by various states, districts, and schools, along with growing pressures exerted by local poverty and community action groups, prompted the U.S. Office of Education to reconsider the legislative and administrative structure of Title I (Jeffrey, 1978; Kirst & Jung, 1982). During the 1970s, Congress and the U.S. Office of Education established more prescriptive regulations related to school and student selection for services, the specific content of programs, and program evaluation, among other things (Herrington & Orland, 1992). These additional responsibilities placed greater administrative demands on local school systems. Funded in part by federal dollars, larger and more specialized state and district bureaucracies grew to monitor local compliance. State and local compliance was confirmed through periodic site visits and program audits by the Office of Education and the Department of Health, Education, and Welfare. As noted by D. K. Cohen (1982) and Meyer, Scott, and Strang (1986), the Title I legislation of the 1970s, along with the proliferation of other state and federal educational mandates, promoted the expansion and increased bureaucratization of local educational agencies.

REFORM THROUGH BUREAUCRATIZATION

As the 1970s progressed, the bureaucratic organization of Title I became institutionalized across the country, and services were delivered to the children targeted by the law (Peterson, Rabe, & Wong, 1986). Rather than a heavy federal presence and intergovernmental conflict, implementation of Title I became a cooperative concern and professional responsibility of local, state, and federal administrators. In addition, as noted by Peterson et al., Title I had inspired greater local concern for, and attention to, the educational needs of children of poverty. Therefore, in marked contrast to the first decade of the program, during the latter half of the 1970s and throughout the 1980s the specific legislative intents, and the desired hortatory effects, were achieved on a far more consistent basis.

Although the program was reaching the students it had targeted during this era, the actual practices were driven more by bureaucratic regulations than by any research-based or practitioner-developed model of what constituted effective education services for disadvantaged children. One of the most important regulations affecting program delivery had been the provision that the compensatory services provided through Title I must supplement, not supplant, the regular educational programs provided to eligible students. In the case of program audits, and to clearly account for federal money, educators and administrators needed to show that the targeted Title I programs actually provided something "extra" and that they were not merely replacing services the students would have received through the regular school program.

This regulation led to widespread use of the "pullout model" as a means for delivering supplemental compensatory services to eligible Title I students. Most often, the students who qualified for services were taken out, or "pulled out," of their regular classrooms each day for 30 to 40 minutes of remedial instruction in reading and mathematics. This arrangement had the advantage of making it clear that the funds were providing something separate from the regular school program, as special teachers, books, and other materials were clearly allocated to the pulled-out Title I students and not to their regular classroom peers. Despite some research suggesting that pullout programs stigmatized children and provided few, if any, academic benefits (Glass & Smith, 1977), approximately three of four Title I schools during the 1970s, 1980s, and much of the 1990s used the pullout model to deliver supplemental services.

COMBINING FLEXIBILITY WITH ACCOUNTABILITY FOR IMPROVEMENT

Instead of the seemingly piecemeal and uncoordinated categorical targeted assistance programs that had served Title I schools since the mid-1960s, in the late 1980s and 1990s a growing belief developed that at-risk students and high-poverty schools could be better served by school-wide 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 school-wide initiatives rather than targeting programs toward all schools where at least 50% of the students were poor. Also, it encouraged schools to use the funds with greater flexibility to support ongoing school-based reform efforts or initiate new ones to help address the educational needs of all children from high-poverty schools. These sweeping changes began the transformation of Title I from a supplemental remedial program to the key driver of the standards-based school-wide reform movement (Borman, 2000).

During the 1990s, Title I school-wide projects proliferated across the country. In 1991, only 10% of eligible Title I schools operated school-wide programs; by 1996, however, approximately 50% of such schools had implemented these programs (Wong & Meyer, 1998). A number of studies from the 1990s showed that, in the short term, these school-wide 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 schools failed to develop coherent statements of beliefs or models for guiding their work and decision making (Murphy & Beck, 1995). These outcomes, combined with new evidence from the congressionally mandated Prospects study of the modest overall effects of Title I services (Borman, D'Agostino, Wong, & Hedges, 1998; Puma et al., 1997), suggested that federal policies aimed at improving education for at-risk students from high-poverty schools were in need of further retooling. Despite the new flexibility afforded by the law, the largely locally inspired school-wide reforms did not yield the desired effects on educational practices and outcomes.

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, 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 effects on academic achievement than either traditional Title I pullout programs or locally developed reforms.

SCALING-UP REFORM WITH REPLICABLE PROGRAMS

Along with 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, would operate at costs comparable to those of current schools after startup funding, and would address all aspects of a school's operation. After receiving nearly 700 proposals in February 1992, NAS chose 11 and provided funds for a 3-year program of development and testing. Since 1995, NAS has continued to focus on scaling up 7 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 comprehensive school reform (CSR) and has helped scale up the CSR movement.

In response to the promise of the externally developed programs disseminated by NAS and other independent model developers, the U.S. Congress also 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 Comprehensive School Reform Program (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 3 years to qualifying schools. Since first authorizing CSRP in fiscal year 1998 and allocating a total of $145 million, Congress has steadily increased its support. In fiscal year 2002, allocations for the CSRP equaled $310 million. This figure included $235 million set aside specifically for Title I schools and $75 million available to any school wishing to apply through the Fund for the Improvement of Education.

The other significant funding source for CSR programs has been Title I. 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. This federal support, combined with the efforts of NAS and other independent developers, has led to the continuing expansion of externally developed CSR models.

Since the early 1990s, the scale up of CSR designs has taken place 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, schools receiving money to implement CSR models through the CSRP have an average poverty rate of 70%. Furthermore, nearly 40% of schools receiving CSRP funds have been identified for school improvement under Title I regulations, and more than 25% have been identified as low-performing schools according to state or local policies.[1]

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 founded 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 school-wide 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 supporting and sustaining the school reform effort.
11. Meets one of the following two 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 possessing these characteristics, but, as suggested by the eighth component of CSR, many educators are turning to groups external to schools, 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 just mentioned. 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 they propose, and the processes through which they propose that schools may achieve those solutions. For example, the Comer School Development Program builds

largely around Dr. James Comer's work in community psychiatry and focuses its energy on creating schools that address a wide range of students' health, social, emotional, and academic challenges. By contrast, the Core Knowledge reform (Hirsch, 1995, 1996), derived from the developer's experiences as a professor of English and education, 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 common adherence to nine broadly philosophical common principles (Sizer, 1992), whereas Success for All (Slavin & Madden, 2001) provides a specific K–6 reading curriculum, professional development sequences, and other school-wide 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 United States, 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, at the least, help educators understand the underlying philosophy of the model. In many circumstances, 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.

FOUR STAGES OF DEVELOPMENT IN THE NATIONAL REFORM MOVEMENT

This series of initiatives in the national movement to bring reform to scale in high-poverty schools has a clear developmental trajectory that can be summarized by four distinct stages. First, the early implementation of Title I was characterized by intergovernmental conflict, poor implementation, and a lack of research-based and practitioner-based knowledge of how to develop effective educational interventions for disadvantaged students. A second stage, during the 1970s and 1980s, was marked by the development of increasingly specific policies to guide the Title I program's implementation, growing bureaucratic cooperation between federal and local authorities in implementing the policies, and improved access for disadvantaged students to the supplemental resources and instruction offered by the program.

Rather than simple access to supplemental services, during the late 1980s and 1990s new Title I legislation stressed reform and improvement of the program. The emphasis on emerging national education standards and systemic reform supplanted many of the earlier concerns about fiscal and procedural accountability, as this latter type of accountability was all but taken for granted. In keeping with the national trends toward site-based decision making and decentralization, Title I afforded schools

greater flexibility to serve disadvantaged students, so long as their test scores improved. For the most part, though, this flexibility did not prompt schools to develop new visions for reform. Aside from some tinkering around the edges, the administration and operation of Title I remained fairly stable.

Beginning in the 1990s, the current stage emerged in which the scale up of research-proven programs and practices has been increasingly regarded as key to improving the effectiveness of high-poverty schools. As in the 1980s and 1990s, the general spirit of today's reform efforts continues to articulate top-down education standards, which dictate many of the changes in the content of schooling. However, the process of reform is in marked contrast to the earlier stages of Title I. Rather than policy mandates or flexibility alone, a growing constellation of replicable programs is becoming the primary lever through which educational practices and the processes of school change are shaped.

In many ways, this recent focus on replicable programs helps 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 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 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, however, the top-down direction is not in the form of distant legislative mandates but represents, in theory, tangible and accessible support for school change rooted in research and literally packaged and delivered to each school's door.

EVIDENCE OF EFFECTS ON ACHIEVEMENT OUTCOMES

Given the apparent progress made in scaling up reform in high-poverty schools, it should come as little surprise that recent evidence suggests that these efforts to meet the needs of disadvantaged children have helped the United States make strides toward greater educational equality. Long-term trend data from the National Assessment of Educational Progress indicate tremendous progress, beginning in the 1970s and 1980s, in closing the persistent achievement gaps separating poor and more advantaged children and African American and White students (Grissmer, Kirby, Berends, & Williamson, 1994; Smith & O'Day, 1991). For instance, during this period the gaps between African American and White children shrank by about two grade levels. The reasons for this rather remarkable trend are open to some debate, but Grissmer and his colleagues asserted that Title I and the other social and educational programs that were first introduced during the "War on Poverty" of the mid-1960s surely had something to do with it.

META-ANALYSIS OF TITLE I EFFECTS

Supporting this assertion, a comprehensive meta-analysis, or quantitative review, of the results of 17 federal evaluations conducted between 1966 and 1993 indicated that the 1970s and early 1980s were also the periods of the greatest improvements in Title I students' math and reading achievement outcomes (Borman & D'Agostino, 1996, 2001). During the early years of Title I, in the late 1960s, the program was not effective in closing the gap because it simply was not implemented as intended by Congress. As the regulations and knowledge base for implementing Title I programs came into clearer focus during the 1970s and 1980s, the intended recipients of the program's services, largely poor and African American children, began to show clear benefits from Title I, and the nation's achievement gaps began to close.

Although it is not possible to establish a true cause-effect relationship between these closing gaps and improvements in Title I students' outcomes, two points are clear. First, Borman and D'Agostino's meta-analysis suggests that the children served by Title I would have been worse off academically without the program. Second, the fact that such tremendous national progress was made in closing the achievement gaps demonstrates that educational inequality can be overcome and potentially eliminated in a relatively short period of time when new policies and funding sources are targeted toward improving education and other services for disadvantaged children and their families. Indeed, these outcomes suggest that scale up of programs for high-poverty schools can contribute to widespread effects on student outcomes.

Beginning in the late 1980s, however, the important gains made by African American and poor children began to slow and even erode somewhat (Grissmer et al., 1994). Once Title I was effectively implemented as intended by Congress during the late 1970s and early 1980s, the promising gains made by participating children also plateaued (Borman & D'Agostino, 2001). After statistically taking into account a variety of programmatic and methodological moderators that influenced the estimates generated by national evaluations of the Title I effect size during the years 1965 through 1994, Borman and D'Agostino (2001) obtained the residuals from the regression analysis. When the average Title I effect size (d) of .11 is fit to each residual, the resulting adjusted effect size by year of implementation scatterplot displayed in Figure 1 provides a visual representation of how Title I effects have changed over the years after statistically taking into account the differences across evaluations.

Figure 1 contains 657 data points, each representing an independent estimate of the Title I effect derived from 17 national studies and including the test scores of more than 41 million Title I students. The line of best fit through the data points indicates a somewhat nonlinear relationship between adjusted effect size and year of implementation. Specifically, Figure 1 shows a linear improvement in program effects from 1966 to the early 1980s, increasing from an effect size of about 0 in 1966 to an effect of nearly .15 in the early 1980s. This suggests that when localities implemented programs of variable but generally poor quality during the 1960s, the effects were, on average, essentially zero. Improved implementation led to improvements in the effectiveness of the

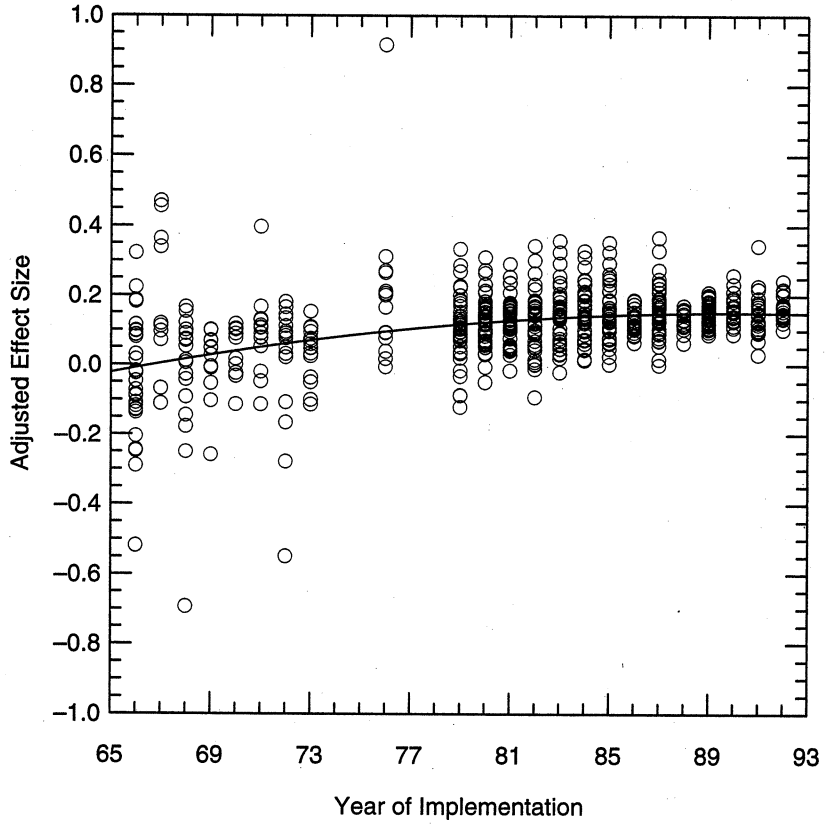

FIGURE 1 Adjusted Effect Sizes by Year of Title I Implementation

program during the 1970s. However, beginning in the 1980s, the effects plateaued, remaining at approximately .15 throughout most of the 1980s and the early 1990s.

This pattern of improvement in Title I effects suggests that once the program was implemented as intended by Congress during the late 1970s and early 1980s, the effects reached a peak that has not changed substantially. The pattern of variability in program effects also supports this conclusion. The wide variation in program effects during the 1960s and early 1970s appears to reflect the variability of local program implementation and evaluation. However, once implementation and accountability requirements became more uniform and established throughout the late 1970s and 1980s, this led not only to increased effectiveness but to more consistent effectiveness. One might conclude that this result suggests an effect of .15 is the best we can do given the current federal funding commitment. Alternatively, it could be taken as a sign that the standardized, and modestly effective, procedures of Title I's recent history require substantial reform to promote continued improvement.

META-ANALYSIS OF CSR EFFECTS

With the No Child Left Behind Act and the movement toward CSR, this reform is under way. A meta-analysis conducted by Borman, Hewes, Overman, and Brown (2003) that synthesized evidence regarding the achievement effects of 29 widely replicated CSR models provides insight into the achievement effects associated with these recent national reform efforts. The 29 models selected for the research synthesis were implemented in 55.6% of the schools that received CSRP funds for externally developed models, as reported in the Southwest Educational Development Laboratory database. Therefore, the results of the review generalize reasonably well to the population of U.S. schools implementing CSR models with CSRP and Title I program funds.

So how do CSR effects compare with the 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 with the effect of CSR programs is the effect of the traditional Title I programs that preceded them, which were the subject of Borman and D'Agostino's (1996) earlier meta-analysis. The overall mean weighted CSR effect size of .15 compares favorably with the overall average weighted Title I effect of .11; however, because the primary studies and the two meta-analyses involved somewhat different methodologies, the comparison is imperfect.

A better comparison between CSR and conventional Title I programs may be drawn directly from the Borman et al. (2003) meta-analysis by examining the CSR effect sizes estimated from the comparison-group studies in schools of 50% poverty or more. In most of these cases, the comparison schools had such high poverty rates that it was very likely they were receiving federal Title I funds. In most cases these schools implemented Title I targeted or school-wide programs, and in most cases they were not implementing other CSR models. These studies, therefore, provided 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, statistically adjusted for methodological characteristics, was .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 CSR school still outperformed 55% of the Title I schools.

Evidence from the National Assessment of Educational Progress and from meta-analytic estimates of the effects of Title I and CSR allows at least two points to come into clearer focus. First, there appears to be national progress in scaling up improved educational outcomes for students and schools in disadvantaged circumstances. This is marked by progress in closing the achievement gaps separating African Americans and Whites and poor and nonpoor students. It is also distinguished by the trend of growing achievement effects associated with national efforts to reform high-poverty schools through Title I and the CSRP. These outcomes suggest that national efforts to scale up reform in high-poverty schools are capable of producing widespread improvements in educational outcomes. In the aggregate, though, these national effects are somewhat modest, amounting to no more than effect sizes ranging from .11 to .15.[2] However, as suggested by the great variability in schooling across the diverse contexts in which it is carried out,

variations in the effects of scaling up reform are often a more significant part of the story than the aggregate effects.

EXPLAINING THE VARIABILITY OF EFFECTS

Perhaps the most salient theme of the meta-analyses of Title I and CSR research is that the overall effects of these national efforts to bring reform to the nation's high-poverty schools are marked by considerable heterogeneity. Rather than a distinct and replicable model for reform, Title I is better understood as a funding mechanism that allows for extensive variation, both across and within schools, in design and implementation. Some schools operate Title I programs that serve all students school-wide, whereas others operate programs that target only the lowest-achieving students within the school. Some schools may also, for example, spend all of their Title I funds on helping kindergartners and first graders learn to read, while other schools may channel their resources toward helping students in the upper elementary grades master math concepts and applications. As a consequence, and as the results from Borman and D'Agostino's (1996, 2001) meta-analytic work suggest, any overall "treatment effect" is best viewed as random rather than fixed, in that a single estimate of the population effect of Title I is not likely to generalize across schools and programs.

Across the 29 CSR models, as one might expect, there is also a considerable amount of variability in effects. As one might expect as well, there is less variability across schools implementing any one of the 29 models because, in contrast to Title I, each of the 29 models offers a relatively distinctive and replicable model of school reform. There are also a number of discrete features of CSR programs, either those listed among the 11 components called for by the U.S. Department of Education or those that have been the topic of previous research that one may identify as key ingredients of reform across the 29 models. However, when modeled as predictors of CSR effect sizes, whether or not the reform model required various components, such as the following, explained very little in terms of the achievement outcomes the school could 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 to explain a substantial amount of the variability in CSR effects.

The general lack of explanatory power for these CSR characteristics suggests at least two possible interpretations. The first is that these components are not important in promoting student achievement in CSR schools, and therefore there is no relationship. The second interpretation is that knowing whether or not a CSR model generally 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. Previous research 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). Knowledge of these factors, which have been largely unmeasured and unreported in evaluations of the achievement effects of CSR programs, would enrich our understanding of the variability in CSR effects.

Indeed, with respect to the variability of outcomes found for both Title I and CSR, one of the most convincing findings from both meta-analyses is simply that implementation matters. The history of Title I has shown that there is a strong relationship between implementation and program effects, as measured by students' achievement outcomes. The best available measure of level of implementation from the meta-analysis of CSR research—the number of years a CSR model was implemented at a school—exhibits a similar outcome. Figure 2, which combines evidence from across the 29 CSR models, displays effect sizes by number of years of CSR program implementation. The finding across the 29 models is consistent in showing an increasing effect on achievement outcomes associated with a greater number of years of implementation.

The figure shows that the CSR effect size, .17, was relatively strong during the first year of implementation. Then, perhaps reflecting the "implementation dip" that Fullan (1991) noted from his conversations with principals and teachers, there appears to be a tendency for new CSR initiatives to worsen before they improve. This is reflected by

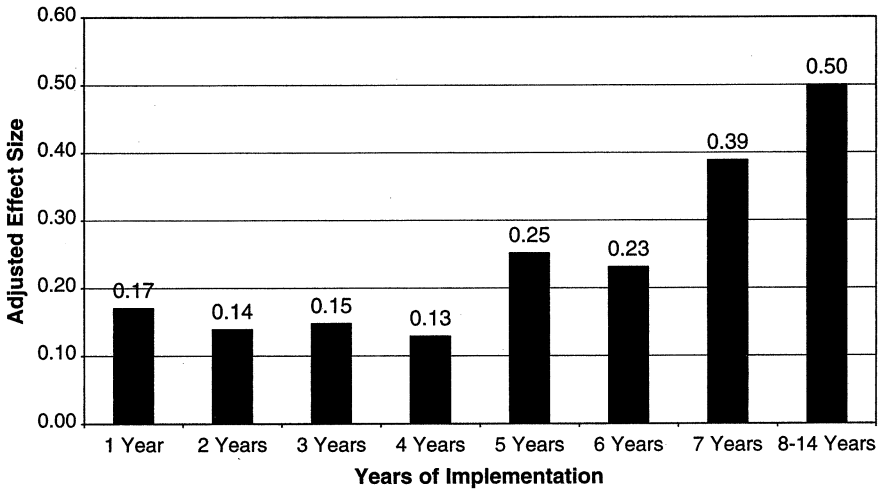

FIGURE 2 Adjusted Effect Sizes by Number of Years of Comprehensive School Reform Model Implementation

the slight decline in effect sizes during the second, third, and fourth years of implementation. After the fifth year of implementation, however, the CSR effects began to increase substantially. Schools that had implemented CSR models for 5 years showed achievement advantages that were nearly twice those found for CSR schools in general, and after 7 years of implementation the effects were more than two and a half times the magnitude of the overall CSR impact (*d*) of .15. The small number of schools that had outcome data available after 8 to 14 years of CSR model implementation achieved effects that were three and a third times larger than the overall CSR effect.

These strong effects of CSR that begin 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 could 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 are plausible. Nonetheless, it is of considerable significance that the average school across all studies reviewed in the meta-analysis had implemented its CSR model for approximately 3 years. The typical study in the meta-analysis, therefore, may have underestimated the true potential of CSR to effect change in schools and improve student achievement.

METHODOLOGICAL CONSIDERATIONS

Beyond implementation and programmatic features of the reforms, differences in the methods researchers have used in evaluating Title I and CSR have had a great deal to do with the variability of the estimates of their achievement effects. Two key differences stand out as important moderators of effect size. First, beginning with the evaluations of most Title I programs during the 1970s, a large amount of research on both Title I and CSR has been based on students' pretest-to-posttest change scores from various norm-referenced achievement tests administered on either a fall-to-spring or annual testing cycle. According to this norm-referenced evaluation model, or one-group pretest-posttest design (Cook & Campbell, 1979), if the mean change score of participating students within a school is greater than zero normal curve equivalents (NCEs) (normalized percentile scores with a mean of 50 and a standard deviation of 21.06) the program is said to be effective. A mean gain greater than zero NCEs has been interpreted as evidence of programmatic impact, on the assumption that in the absence of intervention students tend to remain at the same national percentile rank over time—the "equipercentile assumption" (Tallmadge & Wood, 1981).

After statistical control for other methodological moderators of effect size, one-group pretest-posttest designs yield Title I and CSR effect sizes that are, respectively, one fifth and about one twelfth of a standard deviation higher than those estimated with quasi-experimental and experimental designs that involve control groups. This finding offers an empirical result supporting the suggestions of Cook and Campbell (1979), who noted that effect estimates based on simple one-group pretest-posttest designs are likely to yield greater positive biases owing to history, maturation, and regression-to-the-mean effects.

As a result of these general threats to internal validity and the observed biases in the Title I and CSR effect estimates, the one-group pretest-posttest design may be regarded as among the weakest for measuring educational effects. Ideally, future evaluations would include randomized designs in which schools are assigned at random to CSR and control conditions. When randomized field trials are not feasible, high-quality quasi-experimental control-group designs also may be desirable. However, when comparing directly randomized experiments and quasi-experiments designed to answer the same research questions, Lipsey and Wilson (1993) found that quasi-experiments are more highly variable in the results they produce. As a result, although quasi-experiments may be more convenient and less expensive to conduct in the short run than true experiments, they can be 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 with district averages will provide some understanding of the value-added effects of the model.

In addition, there have been several recent articles suggesting that the CSR research base may be tainted by the fact that the developers are often also the evaluators of their programs (Pogrow, 2000; Walberg & Greenberg, 1999). The meta-analytic findings did suggest that studies performed by the developers yielded effects that were considerably larger ($d = .16$) 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 third-party researchers may seek to taint a model as a result of a personal grudge or professional dislike for its particular orientation. A more plausible source of developer bias is a variant of the so-called "file-drawer" problem, which involves the tendency for researchers to publish or otherwise disseminate their statistically significant findings but consign their nonsignificant findings to a dusty filing cabinet. In this case, CSR developers may selectively report the positive outcomes of their models and file away null and negative findings.

Rather than overt bias or selective reporting, another explanation for the stronger outcomes for the developers' studies is that when developers are more actively involved in the study of their models, they also are 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? Furthermore, if developers found that they were studying half-hearted implementations, they would be in the best position of anyone to help the school improve the quality of its implementation. 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 replications of their CSR models, though, is of some concern.

IDENTIFYING REPLICABLE STRATEGIES WITH THE STRONGEST EVIDENCE OF EFFECTIVENESS

When attempting to determine a single practice or program to implement in a school or a scale up to serve multiple schools, one must weigh considerations regarding the costs, replicability, and quality of the evidence supporting the approach. Few research studies, or even whole bodies of evidence supporting a particular intervention, provide educational policymakers and practitioners with all three pieces of information. But this combination of evidence is essential for good decision making. An intervention backed by solid research demonstrating its effectiveness is worthless if it is too costly or too difficult to implement and scale up. Furthermore, interventions that produce somewhat slighter educational benefits can be preferable over others with evidence of greater benefits if the latter interventions are more expensive and more difficult to replicate. Reflecting on some examples from my work on CSR, I discuss how one may consider factors beyond a simple effect size when attempting to decide on the best available model for a given context.

Costs

Cost analyses and cost-effectiveness analyses in education help decision makers ascertain which program or combination of programs can achieve particular objectives at the lowest cost. As Levin (1995) noted, the underlying assumption is that different alternatives are associated with different costs and different educational results. By choosing those with the least cost for a given outcome, society can use its resources more effectively. When more cost-effective approaches are selected, those resources that are saved can be devoted to expanding programs. In this way, a systematic consideration of both costs and effects can help further the scale-up process.

In deciding whether or not to make a transition from a Title I school-wide or targeted intervention model to an externally developed CSR model, a policymaker or practitioner may ask: Are the benefits of implementing the CSR model worth their seemingly high costs? Borman et al. (2003) found that, on average, CSR programs have first-year costs of approximately $85,000, including both personnel and nonpersonnel expenditures (e.g., training and materials). However, some developers have argued that schools with concentrations of poor children generally are able to garner sufficient resources to implement CSR models by simply reallocating existing supplemental funds and personnel from federal and state Title I programs, special education, desegregation settlements, and other sources (Slavin et al., 1994). In this way, many schools can afford even the high-priced CSR models by simply trading in their largely remedial approaches of the past, most often represented by federal and state Title I programs, for new designs that will enable them to implement research-based school-wide reform programs. As Odden and Archibald (2000) have argued, this method of "resource reallocation" can make implementations of CSR programs essentially "costless."

There are indeed clear challenges in determining the relative costs and benefits of CSR models (Levin, 2002), but if one assumes that implementations in high-poverty

schools generally have few additional costs, the benefits suggested by the CSR meta-analysis are obviously well worth these modest investments. There is some research evidence to suggest that even if one does not assume that CSR implementations are "costless," high-quality models are capable of yielding cost-benefit ratios that equal or exceed those found for other noted educational interventions, including the Tennessee Student/Teacher Achievement Ratio (STAR) class-size reduction effort (Borman & Hewes, 2003). Furthermore, the analyses of Borman and Hewes revealed that a CSR model that focuses on early intervention and prevention actually may save schools investments in the costly remedial practices of special education referrals and retentions in grade, which can alone offset the costs of implementing CSR models. Although this evidence is important, much more cost-effectiveness research is needed for a wider range of CSR models and for a broader array of educational interventions in general.

Replicability

Obviously, if one is concerned with implementing a promising program or practice in a school or scaling it up to serve a large number of schools, one must also consider the replicability of the program and its effects. Borman and Hewes (2003), for instance, considered the replicability of four interventions with strong evidence of educationally meaningful effects on students' short- and long-term outcomes: Success for All, the Perry Preschool, the Abecedarian Preschool, and the reductions in class size of the STAR study.

Success for All and Perry Preschool are the two interventions of the four that are available as nationally disseminated models. Studies from diverse localities suggest that the educational effects of the original Success for All pilot programs tend to be replicated with a good deal of consistency but that these effects depend on the quality of the implementation (Slavin & Madden, 2001). Implementation is not a trivial matter, as Success for All requires educators throughout a school to rethink and actively change many of their practices. After all, it is a whole-school reform model. If teachers do not accept the changes the model suggests, it is not likely to succeed in improving practices and is not likely to affect student outcomes. Before Success for All is adopted, the developer requires that 80% of the faculty agree, by secret ballot, to follow through with the implementation. If this support wanes, or if systemic support through the district or state tails off, the reform is likely to fail.

This has typically been the case in circumstances in which Success for All has failed, including the Memphis, Tennessee, school district, which recently dropped Success for All from more than 40 of its schools, and the Miami–Dade County school district, which dropped the program from all but 7 of the 45 schools that once ran it. The overall quality of implementation, though, clearly is helped by the Success for All Foundation's growing national infrastructure for supporting schools that adopt the model and by recent federal policy changes that make more supplemental resources available to finance comprehensive school reform programs such as Success for All.

Similarly, the educational approach used in the Perry Preschool classrooms and home visits is widely implemented today, primarily through the use of federal Head Start

funds, as the High/Scope Curriculum (Epstein, 1993). Unfortunately, though, the significant resources necessary to replicate the Perry Preschool program, as it was originally designed in Ypsilanti, typically have not been available through public programs (Barnett, 1995; Kagan, 1991). There are other recent examples of high-cost, high-impact preschool programs, including the Chicago Child Parent Centers (Reynolds, Temple, Robertson, & Mann, 2001), that have shown enduring effects on achievement and other important student outcomes. Examples such as these are significant in showing that the general concept of the intensive and relatively costly Perry Preschool model can be successfully funded and replicated. More public commitment, through programs such as Head Start and Title I, or private support, through community organizations and foundations, is needed to establish the large-scale national replication of the pilot program's effects.

Widespread efforts to deliver the Abecedarian model of highly intensive health, educational, and social services to children beginning shortly after birth have not been fully realized either. The Abecedarian project did inspire the U.S. Congress, in its reauthorization of the Head Start Act in 1994, to develop the Early Head Start program, which covers the first 3 years of life. Since its inception, Early Head Start has grown to a nationwide effort of 635 community-based programs serving 45,000 children. However, similar to the comparison between Head Start and Perry Preschool, the Early Head Start program has not provided the same high-intensity services that the Abecedarian children received. Again, although the research evidence from the Abecedarian project clearly demonstrates that highly intensive early intervention can make a profound and enduring difference for the children who participate, the considerable monetary investments and capacity-building efforts necessary to establish a similarly intensive national network of programs have not been undertaken.

On the surface, the reductions in class size modeled by the Tennessee STAR study would seem to be the most easily replicated intervention of the four. In recent years, the federal government has made available billions of dollars to reduce class sizes in the early grades. State-led efforts, such as California's massive initiative, also have begun recently. At least two noteworthy differences, though, set apart the Tennessee STAR model from these national and state-level initiatives. First, the Tennessee STAR class-size reductions occurred in only those schools that had the facilities to accommodate the new classrooms needed to reduce class sizes. Second, the experiment operated in a relatively small number of schools and, therefore, did not create tremendous demands for new teachers. As suggested by California's recent statewide initiative, scaling up class-size reductions to larger numbers of schools has resulted in higher than anticipated costs, shortages of classroom space and qualified teachers, and smaller than anticipated achievement effects (Bohrnstedt & Stecher, 1999). In addition, rather than improving equality of opportunity, Bohrnstedt and Stecher reported that the California effort has exacerbated disparities between districts serving many minority and poor students and districts serving few minority and poor students. Therefore, in areas that require considerable capital improvements to make available the additional classroom space needed to reduce class sizes and where there are potential shortages of qualified teachers, class-size reduction policies may not enjoy the level of success experienced in Tennessee.

Practical matters, including cost and the likelihood that an intervention's effects can be replicated and scaled up, should be considered along with careful analyses of the local context in which the program is to be implemented. Here I have mentioned some contextual factors that may hinder the replication of four model programs. These factors, along with cost information and general evidence of an intervention's replicability, should be considered by local policymakers when choosing among alternative approaches to improving the education of educationally at-risk children. For instance, local funding shortfalls would prevent faithful replication of the two preschool programs. Teacher shortages and a lack of additional classroom space might complicate class-size reductions. Finally, a lack of commitment among teachers and principals to altering their practices and reforming their schools might derail attempts to implement Success for All. All of these contextual issues, among many others that may be specific to the intervention or the locale in which it is to be implemented, may compromise the replication of promising interventions that have otherwise provided good evidence that their results can be replicated with relative consistency and reasonable monetary investments.

Evidence of Effectiveness

Finally, in identifying strategies for scale up, one must simultaneously consider the overall quality, quantity, and effect size of the intervention. In reviewing the research base for replicable CSR programs, we (Borman et al., 2003) developed appraisals of the evidence supporting 29 models for reforming high-poverty schools. We defined four categories of the relative strength of evidence supporting each of the 29 CSR models: *strongest evidence of effectiveness, highly promising evidence of effectiveness, promising evidence of effectiveness,* and *greatest need for additional research.*

With respect to quality of evidence, we sought to identify interventions that had the clearest causal relationships to student achievement outcomes. The level of confidence that the CSR model led to an improvement in student achievement depended on our ability to rule out other explanations for the increase in student achievement. Similar to Cook and Campbell (1979), we deemed the experimental and quasi-experimental research designs as among the most appropriate methodologies for ruling out alternative explanations. In addition to the suggestions of Cook and Campbell, we based this decision on our empirical results. That is, we found clear biases of one-group pretest-posttest designs relative to those studies that included experimental and quasi-experimental control groups.

The second key consideration when assessing the evidence base for an intervention, especially with regard to scale up, is that there is a relatively large number of studies and observations from which one may generalize the findings for the intervention to the population of U.S. schools likely to adopt and implement it. Establishing how many studies is enough to support claims that an educational program or practice is truly scientifically based is a bit more open to debate than decisions regarding the quality of the studies. In the meta-analysis of CSR effects, we used standards of 10 or more studies overall and 5 or more third-party control-group studies as the (arguably arbitrary) standards necessary for inclusion in the top category.

Finally, in establishing the strength of an intervention's evidence base, one must attempt to understand whether the outcomes are statistically significant, educationally meaningful, and, of course, positive. In the context of the meta-analysis of CSR effects, we asked "Does the evidence from control-group studies show that the effects of the reform on student achievement are positive and statistically greater than zero?" In establishing whether or not the effects were educationally meaningful, we compared the effects sizes for the CSR models with the effect sizes for various other existing standards and competing interventions. In the concluding section of the chapter, I return to this topic in attempting to understand the magnitudes of the CSR effects by comparing them with different benchmarks.

The CSR 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, in general, they can be expected to improve students' test scores. The models meeting the *strongest evidence of effectiveness* standard are distinguished from other available CSR designs 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. These programs are among the best examples of reforms capable of being brought to scale that are likely to make a difference across large numbers of high-poverty schools.

CONCLUSION

The CSRP and Title I of the No Child Left Behind Act are at the forefront of the national movement to base the scale up of educational reform on solid research evidence. This legislation, urging the use of research-based educational practices and procedures in schools receiving federal funding, has the potential to revolutionize school improvement in some of the most challenging contexts in the United States. Do the quantity and quality of the research on CSR and Title I provide the scientifically based evidence needed to identify the proven programs and practices that these new policies demand? Furthermore, what lessons might researchers, policymakers, and program developers learn from the preceding review of recent national efforts to bring reform to scale? Four clear implications emerge from the review of the findings of the two meta-analyses and the general history of national efforts to scale up reform.

First, *ironically, the two educational policy areas, CSRP and Title I, most recently and most strongly tied to higher standards of evidence have clear limitations in terms of the overall quality and quantity of studies supporting their achievement effects.* Despite annual expenditures of approximately $10 billion and a history of nearly 40 years, Title I itself has never been subjected to randomized trials (Borman & D'Agostino, 1996). Large-scale evaluations of Title I typically have provided nationally representative survey data describing the characteristics of Title I and non–Title I schools, the characteristics of Title I and non–Title I students, and the achievement outcomes of participants and non-

participants. Quasi-experimental comparisons of outcomes among Title I and non–Title I students have provided some insights into the potential achievement effects of the program. However, the results of these previous national evaluations ultimately suggest that researchers should focus less on attempting to generate national estimates of the program's characteristics and effectiveness and more on studying the effectiveness of specific interventions that could be funded under Title I.

Title I clearly is not a unique, supplemental, or uniform program. It is a funding mechanism designed to support a range of whole-school reform models, various instructional programs and practices, and school organizational and structural changes. Therefore, much more may be learned by studying the effects of an array of replicable programs and practices. For example, in some states, it may be possible to permit a random sample of Title I schools to use their funds to reduce class sizes. Likewise, high-quality data on the effects of various whole-school reform models (e.g., Core Knowledge, Comer's School Development Program, or Success for All) could be generated by randomly selecting control and treatment sites from statewide lists of schools interested in implementing specific reform models. Another experimental strategy could involve multiple small-scale experiments, allowing for the investigation of multiple treatments. The evidence provided by randomized field trials such as these could advance Title I research and policy in unprecedented ways.

The research on CSR focuses on clear and replicable programs. The results, therefore, provide more direct implications for the scale up of reform. CSR, however, is still an evolving field. Twelve of 29 reform models are supported by five or more studies of their achievement effects, and only 4 models have been the subject of five or more third-party studies involving comparison groups. More than 40% of the analyses of CSR effects have been performed by the developers, and about half of the analyses have involved some type of quasi-experimental control group. Only seven studies of 3 CSR models, or about 3% of all studies of the achievement effects associated with CSR, have generated evidence from randomized experiments.[3]

Many of these problems 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. In contrast, some models 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 a relatively high standard of research evidence.

Second, *this history of national efforts suggests a clear developmental trajectory from 1965 to the present that has resulted in historical improvements in disadvantaged students' outcomes.* The implication seems to be that policy mandates and flexibility alone are less likely to produce educational reform and improved achievement outcomes than provider-based assistance in implementing clear and replicable strategies for school change. Although clearer federal mandates were associated with improved implementation and effects of Title I, these efforts were capable of producing no more than modest

effects on student achievement outcomes. Despite continued efforts to tweak federal policy and provide greater flexibility to support school-based reform efforts, the lack of a clear vision or model for reform led to most schools not capitalizing on this flexibility.

In contrast, the most successful CSR models have enjoyed sustained periods of development, evaluation, and refinement and provide clear and replicable strategies for reforming schools. Despite being known as "comprehensive" models, the three most successful models focus on improvement in one rather discrete core area. Success for All and Direct Instruction have very clear instructional technologies that relate, most importantly, to improving literacy instruction. The School Development Program \mic success. In addition to a clear focus on improvement in a discrete area that the developer understands well, the models provide ongoing professional development and site-based assistance to help ensure the success of the reforms. These clear, focused, and well-supported school-based models of improvement are in stark contrast to top-down direction and flexibility for educational reform.

Third, *the results from these national efforts suggest that large-scale reform is capable of producing widespread, but modest, achievement effects.* Historically, teaching has been fraught with what Lortie (1975) called "endemic uncertainties." Moreover, Cook and Payne (2002) argued that the dominant perspectives on evaluation and improvement in education suggest that the context of each district, school, and classroom is so distinctive that only highly specific change strategies mapped to site-specific circumstances are likely to modify and improve central functions. The continued growth and early success of CSR, which has advanced the application of replicable technologies based on scientific knowledge, provide a clear contrast to these long-standing theories and 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 (e.g., 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 capable of being brought to scale across a large number of schools. The increasing marketplace of CSR models and the proven replicability of many of the programs are important developments. To further advance CSR, policymakers and educators must demand clear evidence that the reforms will make a difference.

The results from the meta-analyses suggest that the achievement effects associated with Title I and CSR are statistically significant and meaningful and that they appear to have increased in magnitude as the policies and programs have been better implemented. Our various analyses suggest that Title I and CSR schools can be expected to score between nearly one tenth and one seventh of a standard deviation, or between 1.9 and 3.2 NCEs, higher than control schools on achievement tests. The low-end estimate represents the overall CSR effect size of .09 for third-party studies involving comparison groups, and the high-end estimate represents the effect size of .15 for all evaluations of the achievement effects of CSR. When U_3, a metric devised by J. Cohen (1988), is used, the effect size of .12 for all studies involving control groups tells us that the average school implementing a CSR program outperformed about 55% of similar control schools that did not implement a CSR model.

How should we interpret this overall effect? Cooper (1981) suggested a comprehensive approach to effect-size interpretation in which multiple criteria and benchmarks are used to understand the magnitude of the effect. First, and most generally, one may compare the overall CSR effect size with Cohen's (1988) definitions of small ($d = .20$) and large ($d = .80$) effects within the behavioral sciences. Second, and more specifically, Cohen (1988) pointed out that relatively small effects (approximately .20) were most representative of fields closely aligned with education, such as personality, social, and clinical psychology. Similarly, in their more recent compendium of meta-analyses, Lipsey and Wilson (1993) concluded that psychological, educational, and behavioral treatment effects of modest values of even .10 to .20 should not be interpreted as trivial. Finally, and even more specifically, the effects of recent CSR models appear somewhat stronger than the effects of the extra resources and programs provided through Title I.

Finally, *better evidence is needed to provide both summative and formative appraisals of current and future national efforts to scale up reform in high-poverty schools.* Some models 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 CSR programs on the basis of the 11 requirements (e.g., parent outreach program, clear goals and benchmarks) that make a model "comprehensive," schools and policymakers should pay even stronger attention to model outputs.

Clear research requirements, ample funding for research and development, and a focus on CSR models' results may support the transformation of educational research and practices in much the same way that similar factors have helped transform medical research and treatment. A series of studies similar to the series required in the Food and Drug Administration's premarketing drug approval process might guide the research, development, and ultimate dissemination of educational programs (Borman, 2003). Once a CSR program has met a standard of evidence, its implementation via federal funds, namely those derived 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, Gilbert, McPeek, and Mosteller (1977) noted that only half of the new treatments subjected to randomized clinical trials actually showed benefits beyond the standard treatments patients would have received. Without the benefit of high-quality evaluation, many widely disseminated educational practices may simply waste the time of teachers and students or, potentially, do harm.

At the same time, schools and policymakers should not dismiss promising programs before knowing their potential effects. Instead, developers and the educational research community need to make a long-term commitment to research-proven educational reform and to establish a marketplace 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 address 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 and according to the imperfect criteria available.

NOTES

[1] 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 CSRP awards that began in 1998, 1999, 2000, and 2001. According to the Web site, the database on which this information is based 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 reported are, most likely, underestimates.

[2] Note that these achievement effects are also fairly consistent with experimental estimates from the recent Tennessee STAR study of the educational outcomes of the statewide scale up of reform through reductions in class size (Finn & Achilles, 1999).

[3] These reform models and studies include the School Development Program (Cook, Habib, et al., 1999; Cook, Hunt, & Murphy, 2000), Direct Instruction (Crawford & Snider, 2000; Grossen & Ewing, 1994; Ogletree, 1976; Richardson, Dibenedetto, Christ, Press, & Winsbert, 1978), and Paideia (Tarkington, 1989).

REFERENCES

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *Future of Children, 5*(3), 25–50.

Berman, P., & McLaughlin, M. W. (1978). *Federal programs supporting educational change: Vol. 8. Implementing and sustaining innovation.* Santa Monica, CA: RAND.

Bohrnstedt, G. W., & Stecher, B. M. (Eds.). (1999). *Class-size reduction in California: Early evaluation findings, 1996–1998.* Palo Alto, CA: American Institutes for Research.

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

Borman, G. D. (2003). Experiments for educational evaluation and improvement. *Peabody Journal of Education, 77*(4), 7–27.

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, 4,* 309–326.

Borman, G. D., & D'Agostino, J. V. (2001). Title I and student achievement: A quantitative synthesis. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 25–58). Mahwah, NJ: Erlbaum.

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., & Hewes, G. M. (2003). The long-term effects and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24,* 243–267.

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

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

Cohen, D. K. (1982). Policy and organization: The impact of state and federal education policy on school governance. *Harvard Educational Review, 52,* 474–499.

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. (2000). 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 Institution.

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., et al. (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,* 122–142.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements.* San Francisco: 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. *Elementary School Journal, 101,* 167–192.

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, A. S. (1993). *Training for quality: Improving early childhood programs through systematic inservice training.* Ypsilanti, MI: High/Scope Press.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21,* 97–109.

Fullan, M. G., with S. Stiegelbauer. (1991). *The new meaning of educational change.* New York: Teachers College Press.

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

Glass, G. V., & Smith, M. L. (1977). *"Pullout" in compensatory education.* Boulder: University of Colorado, Laboratory of Educational Research.

Grissmer, D. W., Kirby, S. N., Berends, M., & Williamson, S. (1994). *Student achievement and the changing American family.* Santa Monica, CA: RAND.

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

Hawkridge, D. G., Campeau, P. L., DeWitt, K. M., & Trickett, P. K. (1969). *A study of further selected exemplary programs for the education of disadvantaged children.* Palo Alto, CA: American Institutes for Research.

Hawkridge, D. G., Chalupsky, A. B., & Roberts, A. O. H. (1968). *A study of selected exemplary programs for the education of disadvantaged children.* Palo Alto, CA: American Institutes for Research.

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.

Herrington, C. D., & Orland, M. E. (1992). Politics and federal aid to urban school systems: The case of Chapter 1. In J. Cibulka, R. Reed, & K. Wong (Eds.), *The politics of urban education in the United States* (pp. 167–179). Washington, DC: Falmer Press.

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.

Jeffrey, J. R. (1978). *Education for children of the poor: A study of the origins and implementation of the Elementary and Secondary Education Act of 1965.* Columbus: Ohio State University Press.

Kagan, S. L. (1991). Excellence in early childhood education: Defining characteristics and next-decade strategies. In S. L. Kagan (Ed.), *The care and education of America's young children: Obstacles and opportunities* (90th yearbook of the National Society for the Study of Education, pp. 237–258). Chicago: National Society for the Study of Education.

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

Kirst, M., & Jung, R. (1982). The utility of a longitudinal approach in assessing implementation: A thirteen-year review of Title I, ESEA. In W. Williams, R. F. Elmore, J. S. Hall, R. Jung, M. Kirst, S. A. MacManus, et al. (Eds.), *Studying implementation* (pp. 119–148). Chatham, NJ: Chatham House.

Levin, H. M. (1995). Cost-effectiveness analysis. In M. Carnoy (Ed.), *International encyclopedia of economics of education* (2nd ed., pp. 381–386). Oxford, England: Pergamon Press.

Levin, H. M. (2002). *The cost effectiveness of whole school reforms.* New York: Teachers College, Columbia University.

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.

McLaughlin, M. W. (1976). Implementation of ESEA Title I: A problem of compliance. *Teachers College Record, 77,* 397–415.

Meyer, J. W., Scott, W. R., & Strang, D. (1986). Centralization, fragmentation, and school district complexity. *Administrative Science Quarterly, 32,* 186–201.

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

Odden, A., & Archibald, S. (2000). *Reallocating resources: How to boost student achievement without asking for more.* Thousand Oaks, CA: Corwin Press.

Ogletree, E. J. (1976). *A comparative study of the effectiveness of DISTAR and eclectic reading methods for inner-city children.* Chicago: Chicago State University.

Peterson, P. E., Rabe, B. G., & Wong, K. W. (1986). *When federalism works.* Washington, DC: Brookings Institution.

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.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest—A 15-year follow-up of low-income children in public schools. *Journal of the American Medical Association, 285,* 2339–2346.

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. 1. 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.

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

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

Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S. M., & Smith, L. M. (1994). 'Whenever and wherever we choose': The replication of 'Success for All.' *Phi Delta Kappan, 75,* 639–647.

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

Stringfield, S., Millsap, M., Yoder, N., Schaffer, E., Nesselrodt, P., Gamse, B., et al. (1997). *Special strategies studies final report.* Washington, DC: U.S. Department of Education.

Tallmadge, G. K., & Wood, C. T. (1981). *User's guide to the ESEA Title I evaluation and reporting system.* Mountain View, CA: RMC Research.

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

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

U.S. Department of Education. (2002). *Comprehensive School Reform (CSR) program guidance.* Retrieved March 17, 2003, from http://www.ed.gov/offices/OESE/compreform/chiefltr.html

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

Wargo, M. J., Campeau, P. L., & Tallmadge, G. K. (1971). *Further examination of exemplary programs for educating disadvantaged children.* Palo Alto, CA: American Institutes for Research.

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.

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: Erlbaum.