RUNNING HEAD: Success for All

Final Reading Outcomes of the National Randomized Field Trial of Success for All

Geoffrey D. Borman

University of Wisconsin—Madison

Robert E. Slavin

Johns Hopkins University

Alan Cheung

Hong Kong Institute of Education

Anne Chamberlain, Nancy Madden, Bette Chambers

Success for All Foundation

Borman, G., Slavin, R.E., Cheung, A., Chamberlain, A., Madden, N.A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal, 44* (3), 701-731.

Final Reading Outcomes of the National Randomized Field Trial of Success for All

Abstract

This article reports the final third-year outcomes of the national randomized evaluation of Success for All, a comprehensive reading reform model. Using a cluster randomization design, schools were randomly assigned to implement Success for All or control methods. The final analyses assess literacy outcomes for a three-year longitudinal sample of children, who participated in the Success for All or control condition from kindergarten through second grade, and a combined longitudinal and in-mover student sample, both of which were nested within 35 schools. Hierarchical linear model analyses for both samples revealed statistically significant school-level effects of assignment to Success for All on all three literacy outcomes measured. These effects were as large as one third of a standard deviation on the Word Attack outcome. The results correspond with the Success for All program theory, which focuses on both comprehensive school-level reform and targeted student-level achievement effects through a multi-year sequencing of intensive literacy instruction.

The gaps in reading achievement between minority and white children and poor and more affluent children are among the most important of all educational problems in the United States. According to our nation's report card, the National Assessment of Educational Progress (NAEP) the achievement disparities between fourth grade African American and white, Hispanic and white, and poor and non-poor children are the equivalent of 2½ to nearly 3 years worth of learning (U.S. Department of Education, 2005). Only 13% of fourth grade African Americans and 16% of Hispanics scored at the proficient level on the NAEP, compared to 41% of whites, and just 16% of those eligible for free lunch scored at the proficient level, compared to 42% of non-eligible students. Indeed, the national movement to improve early elementary literacy instruction and learning has left most poor and minority children behind.

The particular importance of literacy can be understood through research that has demonstrated that reading skills provide a critical part of the foundation for children's overall academic success (Whitehurst & Lonigan, 2001). Children who read well read more and, as a result, acquire more knowledge in various academic domains (Cunningham & Stanovich, 1998). Differences in early-elementary outcomes typically become differences in high school graduation, college attendance, and ultimately socioeconomic status That is, the inequalities in our society may be traced in large part to differences that begin with reading in kindergarten and first grade (Entwisle & Alexander, 1989).

Many solutions have been proposed to improve the reading achievement of disadvantaged and minority children. In recent years, education policymakers have increasingly promoted the idea that the best way to improve overall achievement is to

have schools implement programs that have been validated by rigorous, scientific research. This concept is mentioned more than 100 times in the No Child Left Behind Act, and is appearing routinely in education legislation and policies of all kinds . In particular, "scientific evidence" has been defined as experimental evidence from studies in which participants were assigned at random to treatment and control groups (Shavelson & Towne, 2002; Mosteller & Boruch, 2002; Slavin, 2003). Yet, there are far too few educational programs that have been subjected to such rigorous experiments.

This article reports the final outcomes of a three-year randomized experiment evaluating a comprehensive school-wide approach to early literacy instruction, Success for All, that is designed to help all children, regardless of their ethnicity or socioeconomic status, achieve success in reading. Specifically, we contrast the Year 3 outcomes on three measures of literacy achievement for schools and students randomly assigned to the Success for All three-year developmental literacy treatment to those for schools and students assigned to a no-treatment control condition. We assess both the potential cumulative effects of the program on school-level achievement outcomes and explore the longitudinal outcomes of students who remained in the Success for All and control schools across the three years of the study.

In this introduction, we begin with an overview of the Success for All program and the strengths and limitations of the research base supporting it. Next, we discuss the current study and the outcomes from the first two years of this three-year project. We then present the program theory, its implications, and the hypotheses that help frame our work.

The Success for All Program

More than 1,200 mostly high-poverty Title I schools in 46 states are currently implementing the Success for All comprehensive reform program with external assistance provided by the not-for-profit Success for All Foundation. The intervention is purchased as a comprehensive package, which includes materials, training, ongoing professional development, and a well-specified "blueprint" for delivering and sustaining the model. Schools that elect to adopt Success for All implement a whole-school program for students in grades pre-K to five that organizes resources to attempt to ensure that every child will reach the third grade on time with adequate basic skills and will continue to build on those skills throughout the later elementary grades.

Success for All is a school-wide intervention that focuses on prevention and early, intensive intervention designed to detect and resolve reading problems as early as possible, before they become serious. The kindergarten program is a full-day, thematically based program with a focus on language and literacy development. In grades 1-5, students in Success for All schools spend most of their day in traditional, age-grouped classes, but are regrouped across grades for reading lessons targeted to specific performance levels. Daily lesson plans guide teachers to use instructional practices that have been found effective in rigorous research. Among these are cooperative learning (Johnson & Johnson, 1999; Slavin, 1995), metacognitive comprehension strategies (Pressley & Woloshyn, 1995), effective classroom management methods such as rapid pace and active involvement of all students (Evertson, Emmer,& Worsham, 2000), and embedded multimedia (Chambers, Cheung, Madden, Slavin, & Gifford, 2006; Chambers et al., 2006). Using the program's benchmark assessments, teachers formally assess each student's reading performance quarterly and make regrouping changes and changes to

classroom instruction based on the results. Informal, observation-based assessments are also used daily. Instead of being placed in special classes or retained in grade, most students who need additional help receive one-to-one tutoring to get them back on track.

A Success for All school also establishes a schoolwide "solutions" team, which addresses classroom management issues and seeks to increase parents' participation in school generally, to mobilize integrated services to help Success for All families and children, and to identify and solve particular problems such as irregular attendance, problems at home, and homelessness. In addition, each Success for All school designates a full-time program facilitator who oversees the daily operation of the program, provides assistance where needed, and coordinates the various components. Finally, ongoing support for implementation of the program starts with 3 days of intensive training at the beginning of the first school year. Followup services over the first year of implementation consist of 16 days of on-site support provided by Success for All program staff as well as quarterly monitoring of student progress data. After the first year, approximately 15 days of additional training are provided each year (see Appendix for more details concerning the Success for All program components).

The Strengths and Limitations of the Success for All Research Base

Recent reviews of school reform programs have suggested that Success for All is supported by a relatively strong research base (Borman, Hewes, Overman, & Brown, (2003; CSRQ, 2005). The evidence reviewed by Borman and colleagues from 46 quasi-experimental comparison-group evaluations of Success for All and its sister program, Roots and Wings, from across the United States revealed an overall achievement effect of

one fifth of one standard deviation (d = .20). Though compelling in terms of its scope and results, this prior research has three central limitations.

First, all prior studies of Success for All have used a quasi-experimental matched comparison-group design to compare the achievement outcomes of Success for All schools to similar comaprison schools that were matched on pretests and other demographic characteristics. This design, referred to by Cook and Campbell (1978) as a non-equivalent control group design with pretest and posttest, can provide an interpretable test of a causal hypothesis. But, with the non-equivalency of the comparison group, threats to internal validity are far more likely relative to a randomized design. Recent empirical evidence suggests that such comparison-group studies in social policy (e.g., employment, training, welfare-to-work, education) often produce biased estimates of an intervention's effects, because of unobservable differences between the intervention and comparison groups that differentially affect their outcomes (Glazerman, Levy, & Myers, 2002).

Second, although nearly all previous studies of Success for All have employed designs that attempt to match program and control *schools*, they have specified the *student* as the unit of analysis in statistical comparisons of program and control outcomes. Though this unit-of-analysis problem does not necessarily bias the impact estimates, it does underestimate the standard errors of these estimates and leads researchers to make Type I errors. Finally, earlier studies of Success for All have involved small numbers of treatment sites and may be most accurately interpreted as efficacy trials. That is, with the researchers actively involved in assuring that they are studying high-quality implementations in a select number of schools, most of these earlier evaluations seem to

represent assessments of what Success for All can accomplish at its best. The extent to which these results may generalize across broader implementations, though, is of some concern.

The Current Study

In 2000, the Success for All Foundation received a grant from the U.S. Department of Education to carry out a three-year study that was intended to address these limitations of the prior research base. The study reported here was designed as a cluster randomized trial (CRT), with random assignment of a relatively large sample of 41 high-poverty schools from across 11 states. The design primarily compared baseline kindergarten and first grade students nested within schools that were randomized into a grade K-2 Success for All treatment condition to kindergarten and first grade students whose schools were randomized into a grade 3-5 Success for All treatment condition. Thus, the kindergarten and first grade students within the former schools received the Success for All intervention—and served as the treatment cases—and the kindergarten and first grade students within the latter schools continued with their current reading programs—and served as the controls.

An analysis of the first-year achievement data for the main kindergarten and first-grade sample was carried out by Borman et al. (2005a). Using hierarchical linear modeling (HLM) techniques, with students nested within schools, Borman and colleagues reported school-level treatment effects of assignment to Success for All on four reading measures. They found statistically significant positive effects on the Woodcock Word Attack scale, but no effects on three other reading measures. The effect size for Word Attack was $d = 0.22$, which represents more than 2 months of additional learning gains.

The second-year analyses, reported by Borman et al. (2005b), focused on the literacy outcomes for two distinct student samples nested within the study schools. The first set of analyses was for the two-year longitudinal student sample, composed of students who remained enrolled at the treatment and control schools over the full two years of the study. These analyses revealed statistically significant school-level effects of assignment to Success for All on three of the four literacy outcomes measured, with effects as large as one quarter of a standard deviation—or a learning advantage relative to controls exceeding half of a school year—on the Word Attack outcome.

The second student sample included the longitudinal group of 3,290 students and 890 additional students who had moved into the experimental and control schools after the baseline assessments. Though the in-moving students did not benefit from the full Success for All intervention, this combined longitudinal and in-mover sample did comprise the complete enrollments of the targeted grade levels in the treatment and control schools at the time of the Year 2 posttest. In this way, the sample afforded a type of school-level intent-to-treat analysis of the program. Relative to the results for the longitudinal sample, the impact estimates for the combined longitudinal and in-mover sample were somewhat smaller in magnitude and more variable.

In this article, we estimate the Year 3 school-level effects of treatment assignment for those who received the full "dose" of Success for All; namely, the students from the three-year longitudinal student sample. Also, we analyze school-level intent-to-treat (ITT) effects, which are based on the sample of all students enrolled at the study schools at the time of the Year 3 posttest, regardless of the amount of exposure to the treatment that the students actually experienced. Respectively, these analyses address two research

questions, namely: (1) does Success for All produce achievement effects for schools and students targeted by, and exposed to, the model's three-year developmental literacy treatment; and (2) does the comprehensive package of instructional and organizational change produce broader school-wide effects for all students with variable exposure to the treatment?

The Success for All Program Theory: Pursuing School-level and Student-level Effects

Two distinct lines of research and two central features of the program's theory of action inform our work on the cumulative effects of whole-school reform and the longitudinal effects of early literacy instruction. First, due to the comprehensive and well-specified approach to reform, the significant and ongoing professional development across multiple years, and the focus on faculty support and buy-in from the outset—typically, a vote of at least 80% of teachers in favor of program adoption is required—the Success for All developers expect .school-wide change to occur rather quickly and for it to be maintained over time (Slavin & Madden, 2001; Slavin, 2004). Indeed, literature on the implementation of whole-school reforms has suggested that the longevity of a reform effort is commonly seen as an indicator of success (Cuban, 1992; Hargreaves & Fink, 2000). However, the evidence regarding sustained improvements over time for the most typically sought after outcome of school reform, student achievement, is thin. The qualitative and quantitative outcomes that do exist, though, seem to be in general agreement.

Fullan (2001) suggested that implementation of school reform occurs developmentally over time. Significant change in the form of implementing specific innovations can be expected to take a minimum of two or three years. As the reform

process unfolds, Fullan contended that successful schools typically experience "implementation dips" as they move forward. The implementation dip is literally a dip in performance and confidence as one encounters an innovation that requires new skills and new understandings.

Similarly, the meta-analysis of the comprehensive school reform evaluation literature by Borman et al. (2005) suggested that achievement effects of 29 widely used reform models were relatively strong during the first year of implementation. During the second, third, and fourth years of implementation, though, the effects declined somewhat, providing further evidence of the implementation dip noted by Fullan (2001). After the fifth year of implementation, the effects of school reform began to increase substantially. Schools that had implemented reform models for five years showed achievement advantages that were nearly twice those found across the overall sample of schools, and after seven years of implementation, the effects were more than two and half times the magnitude of the overall impact of $d = .15$. Though research relating implementation of school reforms and achievement outcomes is limited--and not without some important qualifications--it suggests that reform efforts take some time to produce school-wide achievement effects and that many schools may experience performance lags during the early years of implementing innovations.

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). However, others, including Bodilly (1996; 1998) and Nunnery (1998), contend that 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. Well-implemented reforms also 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 (Datnow & Stringfield, 2000). Although there have been no systematic analyses across a wide range of whole-school reform models, it would seem that models like Success for All that have clear components addressing each of these issues would tend to result in more reliable implementations and stronger sustained effects than models without such components.

Beyond these school-level components that may influence the quality of implementation and the longevity of reform efforts, the Success for All model has a core and fundamental focus on literacy. The specific sequencing of literacy instruction across the grades is a defining characteristic of the Success for All instructional program. The reading program in kindergarten and first grade emphasizes the development of language skills and launches students into reading using phonetically regular storybooks and instruction that focuses on phonemic awareness, auditory discrimination, and sound blending. The theoretical and practical importance of this approach for the beginning reader is supported by a fairly strong consensus within the research literature that phonemic awareness is the best single predictor of future reading ability (National Reading Panel, 2000). As this awareness is the major causal factor in early reading progress (Adams, 1990), appropriate interventions targeted to develop the skill hold

considerable promise for helping students develop broader reading skills in both the short and long term (Ehri, Nunes, Stahl, & Willows, 2001; Ehri et al., 2001).

During the second through fifth grade levels, students in Success for All schools use school- or district-provided reading materials, either basals or trade books, in a structured set of interactive opportunities to read, discuss, and write. The program offered from second through fifth grade emphasizes cooperative learning activities built around partner reading; identification of characters, settings, and problem solutions in narratives; story summarization; writing; and direct instruction in reading comprehension skills. Through these activities, and building on the early phonemic awareness developed in grades K-1, students in Success for All schools learn a broader set of literacy skills emphasizing comprehension and writing.

Hypotheses

The evidence and theory concerning the cumulative effects of school reform and the development of students' early literacy skills, in general, and the Success for All model, in particular, suggest several important implications for the current study. First, Success for All is best understood as a comprehensive school-level intervention. Accordingly, we designed the study as a cluster randomized trial, with 41 schools randomized to a treatment or control condition, and we specified school-level analyses of the treatment effects of Success for All within a multilevel framework nesting students within the school-level clusters. Specifically, the research design and analysis helps us answer questions related to the hypothesized effects of school-based random assignment to the Success for All program on early elementary literacy outcomes.

Second, given the importance of program intensity and Success for All's multi-year approach to literacy instruction, we hypothesized that the program effects for the longitudinal sample of students who had experienced the full program across three years would be larger in magnitude than those effects found for the sample of all students, which included both the longitudinal sample and the group of students who had moved into the schools between the time of the pretest and Year 3 posttest.

Third, consistent with the program theory related to the sequencing of literacy instruction, which focuses on phonemic awareness skills initially and broader reading skills later, we hypothesized that the third-year program effects would spread into all tested literacy domains. Unlike the first-year impacts, which were restricted to the Word Attack subtest, and the second-year outcomes, which showed no effects on the Passage Comprehension outcome, we hypothesized that we find treatment effects across all literacy domains. Again, though, due to the importance of program intensity, the multi-year sequencing of literacy instruction, and the general importance of learning phonemic awareness skills early, we assumed that the effects across the tests of broader reading skills would be most pronounced for students from the three-year longitudinal sample.

Method

Sample selection

The total study sample of 41 schools was recruited in two phases. The initial efforts focused on reducing the cost to schools of implementing Success for All, which would ordinarily require schools to spend about $75,000 in the first year, $35,000 in the second year, and $25,000 in the third year. During the spring and summer of 2001, a one-time payment of $30,000 was offered to all schools in exchange for participation in

the study. Those schools randomly assigned to the control condition could use the incentive however they wished, and were allowed to purchase and implement any innovation other than Success for All. The schools randomized into the Success for All condition began implementing the program in grades K-5 during the fall of 2001 and applied the incentive to the first-year costs of the program. During this initial phase, only six schools were attracted by this incentive, with 3 randomly assigned to the experimental condition and 3 to the control condition. This sample was far from sufficient.

A second cohort of 35 schools was recruited to begin implementation in fall of 2002. In this cohort, all participating schools received the Success for All program at no cost, but 18 received it in grades K-2 and 17 in grades 3-5, determined at random. Grades K-2 in the schools assigned to the 3-5 condition served as the controls for the schools assigned to the K-2 condition, and vice versa. As discussed by Borman et al. (2005), this design, which included both treatment and control conditions within each school, had advantages and disadvantages.

The design proved to provide a sufficient incentive for the successful recruitment of schools, and it produced valid counterfactuals for the experimental groups that represented what would have happened had the experiment not taken place. The limitation of the design, though, was that the instructional program in the treatment grades might influence instruction in the non-treatment grades. Observations of Success for All treatment fidelity, though, did not reveal significant contamination of this kind, but to the extent it may have taken place, it would have depressed the magnitude of the treatment impacts. In addition, having the two treatments in the same school may have reduced the estimated effectiveness of school-level aspects of Success for All, such as

family support, because both control students and treatment students could have come forward to take advantage of these services. Though these limitations of the design would result in underestimation, rather than overestimation, of the treatment effects, the treatment fidelity observations have suggested that materials and instructional procedures in the Success for All and non-Success for All grades were distinct from each other in all but a few isolated cases and that few if any control students benefited directly from school-level Success for All services.

One additional compromise related to the design is applicable to this final year of data collection. During Year 1 and Year 2 of data collection, the main study focused on outcomes for two cohorts of students nested within the study schools: a baseline kindergarten cohort; and a baseline first grade group. For the Year 2 analyses reported by Borman et al. (2005b), the progress of baseline kindergartners was tracked through the spring of first grade and the progress of the baseline first graders was tracked through the spring of second grade. During Year 3 of the study, though, the majority of baseline first grade cohort students moved into grade 3. The teachers at this grade level in K-2 control schools had used the Success for All model across all three years of the study. Therefore, a viable no-treatment control condition no longer existed for the baseline first-grade students. As a result, the Year 3 analyses reported here focused on only the baseline kindergarten cohort, which progressed through the spring of second grade during this final year of the study.

During both phases of the study, the random assignment was carried out after schools had gone through the initial Success for All buy-in and adoption process, which all schools go through when applying to implement Success for All. After the schools

had hosted an awareness presentation by an authorized Success for All program representative and after 80% of the school staff had voted affirmatively by secret ballot to move forward with the Success for All program adoption, they were eligible for the study. As a final requirement, all schools agreed to allow for individual and group testing of their children, to allow observers and interviewers access to the school, and to make available (in coded form, to maintain confidentiality) routinely collected data on students, such as attendance, disciplinary referrals, special education placements, retentions, and so on. The schools were required to agree to allow data collection for three years, and to remain in the same treatment condition for all three years of the study. The schools that went through this initial process and that agreed to these conditions were randomly assigned by the members of the Oversight Committee to experimental or control conditions.[1]

After the first year of the study, three schools in St. Louis, which were selected during the second phase of recruitment, were closed due to insufficient enrollments. These included one school implementing Success for All in grades K-2 and two implementing in grades 3-5 (hereafter, K-2 schools will be referred to as "experimental" and 3-5 as "control"). In 2004-2005, three more schools, two Success for All and one control, dropped out of the study. One treatment and one control school from St. Louis closed. Also, a treatment school from Arizona dropped the Success for All model due to

[1] Over the course of the three-year study, the Oversight Committee met regularly to help ensure that all procedures were appropriate and to provide important feedback to the research team. The members of the committee included: C. Kent McGuire (Temple University), Steven Raudenbush (University of Chicago), Rebecca Maynard (University of Pennsylvania), Jonathan Crane (Progressive Policy Institute), and Ronald Ferguson (Harvard University).

local political problems and refused to participate in the data collection. The loss of these six schools reduced the third-year analytic sample to 35, 18 experimental and 17 control.

The experimental and control schools included in the Year 3 analyses of outcomes are listed in Table 1. The sample is largely concentrated in urban Midwest locations, such as Chicago and Indianapolis, and in the rural and small town South, though there are some exceptions. The schools are situated in communities with high poverty concentrations, with just a few rural exceptions. Approximately 72% of the students participate in the federal free lunch program, which is similar to the 80% free lunch participation rate for the nationwide population of Success for All schools. The sample is more African American and less Hispanic than Success for All schools nationally. Overall, 56% of the sample is African American, compared to about 40% of Success for All students nationally, and 10% of the sample is Hispanic, compared to the national average of 35%. The percent of white students, 30%, is similar to the Success for All percent white of about 25%.

==========

INSERT TABLE 1 HERE

==========

Table 2 compares the baseline characteristics of the experimental and control schools included in the analyses of Year 3 outcomes. As the results suggest, the 18 experimental and 17 control schools were reasonably well matched on demographics, and there were no statistically significant school-level aggregate pretest differences on the Peabody Picture Vocabulary Test. As demonstrated in Borman et al. (2005a), the

original sample of 21 treatment and 20 control schools was also well matched, with no statistically significant differences on demographics or pretest scores.

==========

INSERT TABLE 2 HERE

==========

Treatment Fidelity

Trainers from the Success for All Foundation made quarterly implementation visits to each school, as is customary in all implementations of the Success for All program. These visits assessed the extent to which the Success for All program components were in place and identified other potential obstacles, including staff turnover and student attendance, that could potentially compromise implementation quality. The visits established each school's fidelity to the Success for All model and provided trainers an opportunity to work with school staff in setting goals towards improving implementation. Many efforts were made to ensure fidelity of the experimental treatment. As is the case in all implementations, teachers in Success for All schools received three days of training and then about 16 days of on-site follow-up during the first implementation year. Success for All Foundation trainers visited classrooms, met with groups of teachers, looked at data on children's progress, and gave feedback to school staff on implementation quality and outcomes. These procedures, followed in all Success for All schools, were used in the study schools to attempt to obtain a high level of fidelity of implementation.

At the time of the Year 3 followup in the spring of 2005, all grade K-2 classes in all schools were implementing their assigned treatments. There was some variability in

implementation quality, which will be the subject of future analyses. For instance, several schools took almost one year to understand and implement the program at a mechanical level and others embraced the program immediately and have done an excellent job. The difficulties in recruiting schools and the last minute recruitment of many of them significantly inhibited quality implementation in some schools, as Success for All schools would have typically done much planning before school opening that many of the study schools (especially in Chicago, St. Louis, and Guilford County, NC) did not have time to do. In general, Success for All classroom instruction was of reasonable quality in most schools, but few schools implemented the tutoring or solutions team aspects of the program adequately, and most had part-time rather than full-time facilitators.

In the non-Success for All grades, teachers were repeatedly reminded to continue using their usual materials and approaches, and not to use anything from Success for All. During implementation visits, trainers also observed classrooms from control grades. Specifically, these observations focused on whether the environment, instruction, and behaviors in the control classrooms resembled the characteristics of the Success for All classrooms. Trainers observed teachers at two control sites participating in strategies used by Success for All, such as the “zero-noise signal” and cooperative learning. Also, at three sites, teachers in the control condition were seen with Success for All materials in their classrooms. In these cases, the importance of the discrete conditions was reiterated for teachers, and the materials were returned to the treatment classrooms. On subsequent visits, Success for All materials were not seen outside of the treatment condition.

Though these few instances were of some concern, contamination of the control condition was minimal. The Success for All program calls for implementation of a coordinated set of practices and materials, and these examples of select Success for All strategies being applied to control classrooms do not necessarily suggest strong examples of treatment cross-over. In addition, Success for All does not claim proprietorship of individual strategies and the variants of some of the modle's procedures, especially cooperative learning, are applied across many classrooms not implementing Success for All.

There was a wide variety of literacy programming within the control condition. Most control groups had a block of time dedicated to either "literacy" or "language arts." These blocks varied in length from 30 minutes to 2 hours. Among sites with longer dedicated blocks, the time was sometimes broken into two to three sessions throughout the day. There were a few control sites that did not have specific time slots for literacy, but instead encouraged the teaching of literacy strategies throughout the day. Within the more structured blocks, control conditions in two schools reported using cross-grade regrouping. Various materials were used across the control condition, including those produced by Scott Foresman, DC Heath, Scholastic, Open Court, and McGraw Hill. All of the control conditions from the Chicago schools used the Houghton Mifflin basal series.

Testing Procedures and Measures

The students from the kindergarten cohort were pretested on the Peabody Picture Vocabulary Test (PPVT III) and then individually posttested on the Woodcock Reading Mastery Tests—Revised (WMTR). The testing windows for the spring posttests were

approximately 4 weeks in length. That is, each of the yearly posttests was completed across all schools within a 4-week time span. The posttesting occurred at the schools no earlier than 8 weeks prior to the final school day. The WMTR posttests were administered to each child during one sitting and averaged approximately 30 minutes.

The six schools from the first phase of recruitment were pretested in fall 2001 and posttested during each subsequent spring. The 35 schools from the main sample were pretested in fall 2002 and posttested each subsequent spring. The pilot and main samples were combined for the analyses. In this analysis, we focused on the outcomes for the Year 3 posttests, which were administered to students in the pilot schools during spring, 2004 and students in the main sample of schools during spring, 2005. Because the metrics of the PPVT III and WMTR tests varied, and to aid in interpretation of the impact estimates, we standardized the pretest and the posttests to a mean of 0 and standard deviation of 1.

Children in the kindergarten cohort were followed into any grade as long as they remained in the same school; retention did not change their cohort assignment. They were also followed into special education. Children who entered Success for All or control schools after fall, 2002 were also posttested each year and included in analyses that combine the baseline cohorts and in-moving student cohorts. Children who were English language learners but were taught in English were posttested in English each year.

The students were individually assessed by trained testers who were unaware of students' experimental or control assignments. Testers recruited for the study were primarily graduate students. All testers had extensive experience with children and had

some prior experience conducting standardized testing. Prior to each spring testing period, the testers participated in a two-day training session led by the researchers. The testers completed a written test and participated in a practice session of at least half of one day with children who were not in the study. The practice sessions were observed and critiqued by members of the research team. Testers returned for additional practice until the researchers were confident that they fully understood the methods for administering the instruments.

Pretests. All children were individually assessed in fall, 2001 (first phase) or fall, 2002 (second phase) on the PPVT III. This assessment served as the pretest measure for all of the reported analyses.

Posttests. During the spring of 2002, 2003, and 2004 (first phase) and the spring of 2003, 2004, and 2005 (second phase), students in the kindergarten longitudinal cohort were individually assessed with the WMTR. During Year 1 and Year 2, four subtests of the WMTR were administered: Letter Identification, Word Identification, Word Attack, and Passage Comprehension. During this final year of data collection, though, the Letter Identification subtest was not administered, because it does not test content that is typically taught in second grade classrooms.

Each of the three subtests of the WMRT required the child to complete distinct tasks that are designed to evaluate specific literacy skills. First, the Word Identification subtest requires the subject to identify and then pronounce words in print. Each word is presented in isolation, and is meant to be pronounced fluently. Second, the degree to which students are able to use their developing phonemic awareness is directly assessed using the Word Attack subtest, which is composed of test items that ask the child to

decode nonsense words. The decoding of non-words is considered the most appropriate measure of phonological recoding (Hoover & Gough, 1990; Siegel, 1993; Wood & Felton, 1994). It provides an indication of the capacity to transfer the auditory skill of phonological awareness to the task of decoding print. Finally, the Passage Comprehesion subtest assesses a child's ability to read and comprehend the meaning of increasingly long passages. In cloze format, students are asked to supply a missing word indicated by an underscored blank space in the passage. The WMTR is nationally normed and has internal reliability coefficients for the Word Identification, Word Attack, and Passage Comprehension subtests of 0.97, 0.87, and 0.92, respectively.

Results

The prior review of baseline data for the school-level sample revealed no important differences between treatment and control schools, and demonstrated that the sample of schools was geographically diverse and generally representative of the population of Success for All schools. In discussing the results of our third-year analyses of achievement outcomes, we begin by assessing whether there were differential data and sample attrition between treatment and control schools, or systematic attrition from the analytical sample that may have changed its characteristics relative to those for the baseline sample.

The final analytical sample was composed of 1,085 students in 18 grade K-2 Success for All treatment schools and 1,023 students in 17 control schools. The three-year longitudinal sample included a total of 1,445 students and the in-mover sample consisted of 663 students. Of these students, 71 remained at the treatment schools over the three years of the study, and were part of the three-year longitudinal sample, but had

missing Year 3 posttest data. So that these 71 students could be included in the analyses, we imputed each student's respective school-level posttest mean. A comparison of the PPVT pretest scores for those with imputed posttests and those with complete Year 3 posttest data revealed no statistically significant difference, t (1.67), p = .09.

Likewise, 37 in-moving students who remained at the study schools had missing Year 3 posttest data, which we imputed using the school-level posttest mean for in-movers. Comparison of the Year 2 Word Attack scores for those Year 2 in-movers with imputed Year 3 posttests and those with complete Year 3 posttest data revealed no statistically significant difference, t (0.13), p = 0.89. Similarly, we compared the Year 1 Word Attack scores for those students who moved into the study schools during Year 1 and who had imputed Year 3 posttests to those Year 1 in-movers with complete Year 3 posttest data and found no statistically significant difference, t (1.14), p = 0.26.[2]

Listwise deletion of the remaining student cases with missing posttest data did not cause differential attrition rates by program condition, χ^2 (1, N = 3357) = 0.07, p = 0.94, leaving 63% of the total sample of 1,725 treatment students and 63% of the 1,632 controls for the preliminary analyses. The data and sample attrition occurred for two reasons. Of the students who were excluded from the analysis, 1,156 (92%), were dropped because they had moved out of the school before the posttests were administered and, thus, had no outcome data, and 93 students (8%) were dropped due to the closure of three participating schools in year 2 and the dropout of three additional schools in Year 3.

To further investigate the internal validity of the study, we compared the pretest scores of those treatment students who were dropped from the analyses to the pretest

[2] A similar comparison between the prior achievement scores of the 13 Year 3 in-movers with imputed posttests to those 195 Year 3 in-movers with complete Year 3 achievement data was not possible, because no prior achievement data were available for these students.

scores of the control students who were dropped from the analyses. No statistically significant difference was found between the treatment and the control students, t (-1.50), p = 0.14 (two-tailed), suggesting that the initial academic ability of the treatment and the control group students who were dropped from our analyses was similar.

To address the issue of external validity, we compared those students who were retained in the analysis to students who were not retained. Those students who were retained had higher pretest scores than those who were not retained, t (-5.84), $p < .01$ (two-tailed). Also, not surprisingly, mobile students who had left the Success for All and control schools were overrepresented among those with missing data χ^2 (1, N = 2108) = 5.33, $p < .05$. Thus, both low-achieving and mobile students from the sample schools were underrepresented in the analyses. This somewhat compromises the external validity of the study in two ways. First, because past quasi-experimental evidence has consistently shown that Success for All tends to have the most profound educational effects on students who are struggling academically (Slavin & Madden, 2001), the omission of low-achieving students with missing posttest data who remained in the Success for All schools may result in downward biases of the treatment effect estimates. Second, because the primary missing data mechanism was mobility from the study schools, this limits generalization to non-mobile students who remained in the baseline treatment and control schools.

While conceding these limitations, there is no conflict in this experiment between random assignment of treatment and missing at random. That is, among the complete data observations, those assigned to control have similar covariate distributions to those assigned to treatment. As noted by Rubin (1976) and Little and Rubin (1987), the

missing data process is *ignorable* if, conditional on treatment and fully observed covariates, the data are *missing at random* (MAR).

Hierarchical Linear Model Analyses of Year 2 Treatment Effects

This cluster randomized trial (CRT) involved randomization at the level of the school and collection of outcome data at the level of the student. With such a design, estimation of treatment effects at the level of the cluster that was randomized is the appropriate method (Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). We applied Raudenbush's (1997) proposed analytical strategy for the analysis of CRTs: the use of a hierarchical linear model. In this formulation, we simultaneously accounted for both student and school-level sources of variability in the outcomes by specifying a 2 level hierarchical model that estimated the school-level effect of random assignment. Our level 1, or within-school model, nested students within schools with their Year 3 posttest achievement predicted by a school-level mean achievement intercept and an error term

$$Y_{ij} = \beta_{0j} + r_{ij},$$

which represents the spring posttest achievement for student *i* in school *j* regressed on a school-level intercept plus the student-specific level-1 residual variance component, r_{ij}.

At level 2 of the model, we estimated the cluster-level impact of Success for All treatment assignment on the mean posttest achievement outcome in school *j*. As suggested by the work of Bloom, Bos, and Lee (1999) and Raudenbush (1997), we included a school-level covariate, the school mean PPVT pretest score, to help reduce the

unexplained variance in the outcome and to improve the power and precision of our treatment effect estimates.[3] The fully specified level 2 model was written as

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{MEANPPVT})_j + \gamma_{02}(\text{SFA})_j + u_{0j},$$

where the mean posttest intercept for school j, β_{0j} was regressed on the school-level mean PPVT score, the SFA treatment indicator, plus a residual, $u_{0j.}$.[4]

Outcomes for the Longitudinal Sample. The multilevel models, shown in Table 3, assessed student and school-level effects on the four literacy outcomes as measured by the Woodcock-Johnson Year 3 posttests. Across the three outcomes, the impact estimate for Success for All assignment ranged from a standardized effect of approximately d = 0.21, for Passage Comprehension, to d = 0.33 for the Word Attack subtest. All three of the treatment effects were statistically significant, with the impact on Word Attack of 0.33 at the $p < .01$ level of confidence, the impact on Word Identification of 0.22 at the $p < .05$ level, and the treatment effect on Passage Comprehension of 0.21 at the $p < .05$ level of confidence. In all three models, the school-level mean pretest covariate was an

[3] We formulated other multilevel models that included a student-level pretest score and the school-level covariates percent minority and percent free lunch, which are listed in Table 3. After including the school mean pretest covariate, though, the modeling of these additional copvaritaes did not explain considerably more between-school variance and did not appreciably improve the precision of the Success for All treatment effect estimates. For these reasons, we used the more parsimonious models presented.

[4] The statistical precision of the design can be expressed in terms of a minimum detectable effect, or the smallest treatment effect that can be detected with confidence. As Bloom (2005) noted, this parameter, which is a multiple of the impact estimator's standard error, depends on: whether a one- or two-tailed test of statistical significance is used; the α level of statistical significance to which the result of the significance test will be compared; the desired statistical power, 1 - β; and the number of degrees of freedom of the test, which equals the number of clusters, J, minus 2 (assuming a two-group experimental design and no covariates).

The minimum detectable effect for our design is calculated for a two-tailed t-test, α level of $p < .05$, power, 1 - β, equal to 0.80, and degrees of freedom equal to J = 35 schools minus 3 (a two-group experimental design with the school mean PPVT pretest covariate). Referring to Tables 3 and 4 for the Success for All impact estimators' standard errors, which range from .09 to .11, and employing Bloom's (2005) minimum detectable effect multiplier, we calculated minimum detectable effects of approximately d = .26 to d = .32. That is, our design had adequate power to detect school-level treatment-control differences of at least .26 to .32 standard deviations.

important predictor of the outcome, with higher initial PPVT pretest scores predicting higher Year 3 posttest scores.

==========

INSERT TABLE 3 HERE

==========

Outcomes for the Combined Longitudinal and In-mover Sample. In Table 4, the multilevel models estimate student and school-level effects on the three Year 3 literacy outcomes for the combined longitudinal and in-mover sample. The Success for All impact estimates across these multilevel models were very similar in magnitude relative to the effects found for the longitudinal sample in Table 3. Across the three outcomes, the impact estimate for Success for All assignment ranged from a standardized effect of approximately $d = 0.21$, for Passage Comprehension, to $d = 0.36$ for Word Attack. Again, all three treatment effects were statistically significant. The impact on Word Attack of 0.36 was statistically significant at the $p < .01$ level of confidence, the impact on Word Identification of 0.24 at the $p < .05$ level, and the treatment effect on Passage Comprehension of 0.21 at the $p < .05$ level of confidence.

==========

INSERT TABLE 4 HERE

==========

Discussion

Practical and Theoretical Interpretations of the Outcomes

As depicted in Table 5, after three years of implementation, the evidence suggests that Success for All schools are capable of producing broad effects across the literacy

domain for both children who are exposed to the model over each of the first three years of their academic careers and for all children enrolled in the schools, regardless of the number of years of exposure to the reform. Though these broad effects were not realized during the first or second year of implementation, this result corresponded with the Success for All theory of action regarding literacy instruction and learning.

During the first year of implementation, the phonetically regular storybooks and instruction employed in kindergarten within Success for All schools produced strong initial advantages on the Word Attack subtest, which measures students' phonemic awareness. Consistent with empirical and theoretical work in early reading, which has provided strong evidence suggesting that phonemic awareness is the best single predictor of reading ability in the early grades and beyond, the early treatment effect in this domain appeared to be an important factor associated with the development of students' broader reading skills across the ensuing years of the study during first and second grade. By the third year of the study, at the end of second grade, children's initial advantages in phonemic awareness held and additional advantages emerged across the other literacy domains tested.

In addition, the improvements in school-wide effects across the first three years of implementation suggest that the program is sufficiently comprehensive to impact all children attending Success for All schools, regardless of the number of years they were exposed to the intervention. Like the advantages for the longitudinal cohort, though, theses effects emerged over time, spreading across the literacy domain with each ensuing year of implementation. It should be noted, though, that the emergence of these school-wide effects over time is largely explained by the developmental progress of the students

who experienced the program across all three school years. Even by Year 3 of the study, the majority of students, 69%, had remained in the Success for All and control schools over the full longitudinal period. But, it is also possible that these school-wide improvements, found for both those students who remained in the schools across all three years of the study and those children who moved into the schools over the three years, suggest organizational learning and development. That is, the treatment may become more efficacious as teachers and staff at the Success for All schools become more familiar with the procedures demanded by the program and as the quality of implementation has time to improve.

Though the results largely correspond with prevailing theory and evidence from reading research, the steady and rather quick progress of school reform, as seen in the results presented in Table 5, is somewhat different from the research evidence from the school reform literature. The process and effects of sustained efforts to transform schools, instruction, and learning are not well understood (Cuban, 1992; Kirst & Meister, 1985; Tyack & Cuban, 1995). The evidence that does exist suggests that ambitious educational change takes time and that schools may face performance setbacks in the early years as practitioners struggle to develop the new skills and new understandings demanded by the reform (Fullan, 2001). The results from this study, though, suggest a different trajectory for the outcomes of a comprehensive school reform initiative.

Indeed, the meta-analysis of outcomes from implementations of 29 whole-school reform models by Borman et al. (2003) suggested that effects of the magnitude found here after 3 years--between $d = 0.21$ and $d = 0.36$–are more typical of those found for implementations of 5 to 7 years, which ranged from between $d = 0.25$ and $d = 0.39$.

Similarly, Fullan (2001) contended that significant change can be expected to take a minimum of two or three years, and that schools often experience "implementation dips" during the early years of reform efforts. The well-specified nature of the Success for All model, the significant and ongoing professional development and implementation support, and the faculty support and buy-in that the model demands from the outset are likely to be important design features of the model that contributed to the relatively quick improvements in schools' reading outcomes.

Thus far, this discussion has considered the school-level components of Success for All and the nature of the reading instruction that the model specifies as independent supports for reform, but it is also important to consider how these school-level and instructional elements interact to promote school improvement. Scholars who have advanced theory and research on the organizational context of teaching and learning, including Gamoran, Seada, and Marrett (2000), Bidwell (2001), and Rowan (1995), have noted various conceptions of how the school as an organization may facilitate instructional improvements. For instance, one theoretical tradition assumes that the technical work of teaching is vague and imprecise and is essentially decoupled from administration, which deals with legitimacy and resource-providing exchanges with external actors (Weick, 1976). In a loosely coupled system, decisions occurring in one segment of the organization do not reverberate in clearly patterned ways in other segments. Therefore, what happens in one classroom may have little impact on another, and decisions made be the principal have modest effects on what the students actually experience and learn. Clearer educational standards, stronger forms of accountability, and the pursuit of systemic reform are just some of the examples of how educational

policymakers have attempted to improve the coupling and control of schools and the reform of instruction.

A second tradition, with a stronger focus on teachers' work, defines teaching as complex and dynamic and suggests that teachers usually work in the absence of well-specified methods and clear standards (Rowan, 1995). When teaching is understood as a nonroutine activity, support for instructional change requires organic organizational structures characterized by faculties that pursue instructional reform through decentralized, small, and informal problem-solving social systems. Rather than relying on tightening the coupling, this organizational strategy eschews bureaucratic controls and focuses on expanding teacher commitment for reform through collaborative and participative management strategies.

The Success for All model for organizational and instructional change seems to address both issues of control and commitment. First, the model attempts to address the loosely coupled and bureaucratic organization of schools and the complex and dynamic nature of teaching by making the technical work of literacy instruction clearer and more precise. Coupling is tightened and the nonroutine work of teachers becomes more predictable and focused. At the same time, though, Success for All provides a "common language" that is spoken by all teachers and administrators and offers the potential to develop a professional culture that is built around a commonly understood mission. Rather than a sole focus on bureaucratic or organic forms of organization, the leadership for school improvement is provided by a combination of clear externally provided technologies for improving reading instruction and more organic forms of interactions with school-based facilitators and other colleagues, who provide additional supports for

sustaining the model and adapting it to fit local circumstances. This combination of external support and school-level leadership for reform, combined with its strong attention to the technical core of instruction, appear to be central ways in which Success for All puts into practice central theories of organizational capacity for school reform.

==========

INSERT TABLE 5 HERE

==========

Interpreting the Magnitude of the Effects

Overall, students from Success for All schools scored from approximately one fifth to one third of a standard deviation higher on the reading assessments than controls not served by Success for All. Using a metric devised by Cohen (1988), U_3, the largest effect size of d = .36 for the Word Attack domain tells us that the average student from a Success for All school outperformed about 64% of his or her control-group counterparts. How should we interpret the magnitude of this effect?

Cooper (1981) has suggested a comprehensive approach to effect size interpretation that utilizes multiple criteria and benchmarks for understanding the magnitude of the effect. First, how do the Success for All effects compare with the important national achievement gaps in reading? Using data from the Early Childhood Longitudinal Study—Kindergarten Cohort (ECLS-K), we calculated the reading achievement gaps separating African American and white students and poor and non-poor students at the end of the first grade. According to these nationally representative data, the black-white achievement gap was equivalent to 0.70 *SD*s and the difference between the outcomes of poor and non-poor students was half of one standard deviation.

After exposure to Success for All, students from the treatment schools held advantages over their counterparts from the control condition that equaled from over half to nearly three quarters the magnitude of these gaps.

Second, and more specifically, how similar are the treatment impacts from the current study compared to other efforts to help close the achievement gap and improve the outcomes of students attending high-poverty schools with substantial minority student enrollments? General evidence regarding the overall effects that we should expect from school-wide reform efforts was provided by an analysis of NAEP reading data by Hedges and Konstantoupolos (2002). After statistically controlling for measurable student background characteristics, the authors concluded that a standardized mean difference of $d = 0.65$ separated the achievement outcomes of schools at the 10^{th} and 90^{th} percentile of the NAEP reading achievement distribution. In other words, moving a school from the bottom 10% of schools in the U.S. to the top 10% of all schools in the nation would require a treatment effect equivalent to nearly two thirds of one standard deviation.

Another obvious comparison is the overall effect of similar comprehensive school reform programs. These programs were the subject of the recent meta-analysis by Borman et al. (2003), who concluded that the overall effects of the 29 most widely deployed comprehensive school reform models were between $d = .09$ and $d = .15$. The effects of traditional federal Title I programs, which have historically funded targeted remedial interventions, such as pullout programs, and schoolwide programs designed to assist at-risk students, provide another benchmark. The achievement effects of Title I were reviewed by Borman and D'Agostino (1996), who synthesized the results from all federal evaluations conducted between 1965 and 1994. Though Borman and D'Agostino

applied a correction, these Title I evaluations, which almost exclusively utilized a non-experimental one-group pre-post design, may overestimate the true Title I effect. Across the 29 years of federal evaluations, the overall average effect size associated with Title I was $d = .11$.

The treatment impact found for a relatively recent and high-profile evaluation, the Tennessee Student-Teacher Achievement Ratio (STAR) study, provides yet another important criterion to which we may compare the Success for All effects. This intervention also was targeted toward children in the early elementary grades, from kindergarten through third grade. Like the current study, it also applied an experimental design, which included random assignment of children and teachers to small classes of 13-17 students, conventional classes of 22-26, or conventional classes with a teacher's aide. Also similar to the study reported here, the STAR study involved implementation of an educational intervention at scale, involving 79 schools across the state of Tennessee. Though there were no effects for those students whose classrooms were served by a teacher's aide, Nye, Hedges, and Konstantopoulos (1999) found advantages of $d = .11$ to $d = .22$ favoring the children assigned to receive the class-size reduction over those in conventional classes.

Beyond the statistical significance of the Success for All effects, these comparisons to other meaningful criteria suggest that the impacts are of practical importance and appear to be greater in magnitude than the effects of other interventions that have been designed to serve similar purposes and student and school populations. Are these benefits worth the costs associated with implementing the program? The three-year costs of all non-personnel expenditures, which include items such as training and

materials, are approximately $135,000 in the typical Success for All school. This figure, though, does not include additional costs that may be associated with the personnel demanded by the program, including tutors and facilitators.

The Succeess for All developers have argued that schools with concentrations of poor children generally are able to garner sufficient resources to implement the model 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 cover the program's costs by simply trading in their largely remedial approaches of the past, most often represented by federal and state Title I programs, for Success for All. As Odden & Archibald (2000) argued, this method of "resource reallocation" can make implementations of programs like Success for All essentially "costless."

There are, indeed, clear challenges in determining the relative costs and benefits of school reform models (Levin, 2002), but if one assumes that implementations in high-poverty schools generally have few additional costs, the benefits we have found are obviously well worth these modest investments. There is some research evidence to suggest that even if one does not assume that Success for All implementations are "costless" and if one were to also take into account the potential additional personnel expenses of the model, it is still capable of yielding cost-benefit ratios that equal or exceed those found for other noted educational interventions, including the Tennessee STAR class-size reduction effort (Borman & Hewes, 2003). Though this evidence is suggestive, much more cost-effectiveness research is needed for Success for All and for a broader array of educational interventions in general.

Conclusion

Using the Success for All model, the reform was replicated across 18 schools serving approximately 10,000 children in districts throughout the United States. The findings of statistically significant positive achievement effects from this large-scale implementation of a randomized field trial of a routine practice program are unusual for studies in education. This study is unlike other renowned randomized trials that also demonstrated the efficacy of early educational interventions, including the evaluation of 58 children from the Perry Preschool program in Ypsilanti, Michigan (Schweinhart, et al., 2005) and the study of 57 children attending the Abecedarian early childhood program in one site in North Carolina (Campbell & Ramey, 1994). The effects noted in this study are not based on a model implementation operating in one location as a demonstration of the optimal impact of an educational program. Instead, the results should be interpreted as those that are likely to be obtained in broad-based implementations of Success for All, with all the attendant problems of start-up and of maintaining quality at scale. In this sense, this multi-site field trial provides experimental evidence of the widespread impact that can be expected when the Success for All intervention is scaled up in a real-world policy context.

References

Adams, M.J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Bidwell, C.E.. (2001). Analyzing schools as organizations: Long-term permanence and short-term change. *Sociology of Education, 74* (Extra Issue), 100-114

Bloom, H.S. (Ed.) (2005). *Learning more from social experiments: Evolving analytic approaches.* New York: Russell Sage Foundation.

Bloom, H.S., Bos, J.M., & Lee, S-W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review, 23,* 445-469.

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

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.

Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A.M., Madden, N.A., & Chambers, B. (2005a). Success for All: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27,* 1-22.

Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2005b). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42,* 673-696.

Campbell, F. A. & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development, 65,* 684-698.

Chambers, B., Cheung, A., Madden, N., Slavin, R. E., & Gifford, R. (2006). Achievement effects of embedded multimedia in a Success for All reading program. *Journal of Educational Psychology*, *98* (1), 232-237.

Chambers, B., Slavin, R.E., Madden, N.A., Abrami, P.C., Tucker, B.,J., Cheung, A., & Gifford, R. (2006). *Technology infusion in Success for All: Reading outcomes for first graders.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Comprehensive School Reform Quality Center (2005). *CSRQ Center report on elementary school comprehensive school reform models*. Washington, DC: American Institutes for Research.

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

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.

Cuban, L. (1992). What happens to reforms that last? The case of the junior high school. *American Educational Research Journal, 29*, 227-251.

Cunningham, A.E., & Stanovich, K.E. (1998). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33,* 934-945.

Donner, A., & Klar, N. (2000). Design and analysis of group randomization trials in health research. London: Arnold.

Ehri, L., Nunes, S., Willows, D., Schuster, B., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36,* 250-287.

Ehri. L., Nunes, S., Stahl, S., & Willows, D. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71(3),* 393-447.

Evertson, C.M., Emmer, E.T., & Worsham, M.E. (2000). *Classroom management for elementary teachers (5th ed.).* Boston: Allyn & Bacon.

Entwisle, D.R., & Alexander, K.L. (1989). Early schooling as a "critical period" phenomenon. In K. Namboodiri R.G. Corwin (Eds.) *Sociology of education and socialization* (pp. 27–55). Greenwich, CT: JAI Press.

Fullan, M. (2001). *The new meaning of educational change* (3rd ed.). New York: Teachers College Press.

Gamoran, A, Secada, W.G., & Marrett, C.A. (2000). The organizational context of teaching and learning: Changing theoretical perspectives. In M.T. Hallinan (Ed.)

Handbook of the sociology of education (pp. 37-63). New York: Kluwer Academic/Plenum Publishers.

Glazerman, S., Levy, D.M., & Myers, D. (2002). *Nonexperimental replications of social experiments: A Systematic Review.* Princeton, NJ: Mathematica Policy Research, Inc.

Hargreaves, A.& Fink, D. (2000). Three dimensions of educational reform. *Educational Leadership, 57*(7), 30-34.

Hedges, L.V. & Konstantopoulos, S. (2002, April). How large an effect should we expect from school reform programs? Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

Heinsman, T.H., & Shadish, W.R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*, 154-169.

Hoover, W.A., & Gough, P.B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2,* 127-160.

Johnson, D.W., & Johnson, R.T. (1999). *Learning together and alone: Cooperative, competitive, and individualistic learning.* Boston: Allyn and Bacon.

Kirst, M., & Meister, G. (1985). Turbulence in American secondary schools. What reforms last? *Curriculum Inquiry, 15*, 169-186.

Levin, H.M. (2002). *The cost effectiveness of whole school reforms. ERIC Clearinghouse on Urban Education, Urban Diversity Series 114.* 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.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

Mosteller, F., & Boruch, R. (Eds.) (2002). *Evidence matters: Randomized trials in educational research*. Washington, DC: Brookings.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Rockville, MD: National Institute of Child Health and Human Development.

Nye, B., Hedges, L.V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment," *Educational Evaluation and Policy Analysis, 21,* 127-142.

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

Pressley, M., & Woloshyn, V. (1995). *Cognitive strategy instruction that really improves children's academic performance* (2nd ed.). Cambridge, MA: Brookline.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173-185.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 94* (2), 240-257.

Rosenshine, B., & Stevens, R.J. (1986). Teaching functions. In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed.) (pp. 376-391). New York: MacMillan.

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.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Schweinhart, L.J., Montie, J., Xiang, Z., Barnett, W.S., Belfield, C.R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40.* (Monographs of the High/Scope Educational Research Foundation, 14). Ypsilanti, MI: High/Scope Press.

Shankweiler, D. P., Crain, S., Katz, L., Fowler, A. E., Liberman, A.M., Brady, S. Thornton, R., Lundquist, E., Dreyer, L., Fletcher, J., Stuebing, K.K., Shaywitz, S.E., & Shaywitz, B.A. (1995). Cognitive profiles of reading-disabled children: Comparison of language skills in phonology, morphology, and syntax. *Psychological Science, 6*(3), 149-156.

Shavelson, R.J., & Towne, L. (Eds.) (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Siegel, L.S. (1993). The development of reading. *Advances in Child Development and Behaviour, 24,* 63-97.

Slavin, R.E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Boston: Allyn & Bacon.

Slavin, R.E. (2003). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Slavin, R. E. (2004). Built to last: Long term maintenance of Success for All. *Remedial and Special Education, 25* (1), 61-67.

Slavin, R.E., & Madden, N.A. (Eds.) (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.M., & Smith, L.M. (1994). 'Whenever and wherever we choose' The replication of 'Success for All.' *Phi Delta Kappan, 75*, 639-647.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia*. Cambridge, MA: Harvard University Press.

U.S. Department of Education (2005). *The nation's report card; Reading 2005* (NCES 2006-451). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Available online at: http://nces.ed.gov/nationsreportcard/pdf/main2005/2006451.pdf

Whitehurst, G.J., & Lonigan, C.J. (2001). Emergent literacy: Development from prereaders to readers. In S.B. Neuman & D.K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 11-29). New York: The Guilford Press.

Wood, F.B. & Felton, R.H. (1990). Separate linguistic and attentional factors in the development of reading. *Topics in Language Disorders, 14*(4), 42-57.

Author Note

We thank Steven Ross, Alan Sterbinsky, Daniel Duran, Michael Reynolds, Shoba Shagle, Margarita Calderon, Dewi Smith, and Dana Andrews for their assistance with data collection and analysis. We also thank the distinguished group of scholars who served as members of the Oversight Committee for this study. The members of the committee included: C. Kent McGuire (Temple University), Steven Raudenbush (University of Chicago), Rebecca Maynard (University of Pennsylvania), Jonathan Crane (Progressive Policy Institute), and Ronald Ferguson (Harvard University).

This research was supported by grants from the Institute of Education Sciences, U.S. Department of Education (R305P030016, R117D40005, and R305A040082). However, any opinions expressed are those of the authors, and do not necessarily represent IES positions or policies.

Table 1

Schools Participating in the Success for All Randomized Trial, Grouped by Assignment.

School	District	ST	Enrollment	% White	% African American	% Hispanic	% Female	% ESL	% Special education	% Free lunch
Haven[a]	Savannah	GA	340	1.0	95.7	0.0	48.3	0.0	10.3	83.8
Jefferson[a]	Midland	OH	548	99.6	0.0	0.0	44.3	0.0	18.3	30.9
Northwood[a]	Moorseville	IN	426	98.3	0.0	0.5	50.5	0.2	2.5	17.0
Benjamin E. Mays	Chicago	IL	418	0.0	90.0	10.0	49.0	0.0	10.0	95.0
Bertha S. Sternberger	Guilford	NC	342	69.0	26.0	0.8	50.0	0.0	15.0	21.0
Brian Piccolo	Chicago	IL	980	0.0	78.0	21.0	48.0	11.0	13.0	97.0
Cesar Chavez	Norwalk	CA	466	4.3	2.6	89.1	50.9	69.0	4.3	89.0
Earl Nash	Noxubee	MS	484	0.41	99.5	0.0	50.8	0.0	4.13	100
Gundlach	St. Louis	MO	234	0.0	100	0.0	46.0	0.0	2.9	97.1
Harriett B. Stowe	Indianapolis	IN	275	25.0	15.0	56.0	46.0	56.0	19.0	97.0
James Y. Joyner	Guilford	NC	381	44.4	47.0	4.5	51.4	5.8	16.0	44.1
Laurel Valley	Ligonier Valley	PA	392	99.0	0.5	0.25	47.0	0.0	8.0	45.0
Linden	Linden	AL	211	0.5	98.6	0.95	46.9	0.05	10.0	91.0
Paramount Jr.	Greene	AL	417	0.0	99.0	0.0	42.7	0.0	9.0	93.0
Pleasant Garden	Guilford	NC	588	79.0	11.8	4.3	48.0	2.9	1.9	28.8
Robert H. Lawrence	Chicago	IL	643	0.0	99.0	1.0	70.0	0.0	4.7	90.0
Waveland	S Montgomery	IN	148	98.0	0.0	0.0	54.0	0.0	13.0	26.0
Wood	Tempe	AZ	630	21.2	10.9	40.1	50.2	25.5	8.7	48.5
SFA school means			440	35.5	48.5	12.7	49.7	9.47	9.5	66.3
M. E. Lewis[a]	Sparta	GA	547	1.1	97.8	0.0	48.1	0.0	1.0	93.4
Newby[a]	Mooresville	IN	318	97.0	0.0	1.0	49.0	1.0	11.0	28.0
Walnut Cove[a]	Walnut Cove	NC	334	78.0	17.0	0.0	51.0	0.0	11.0	31.0
Augustin Lara	Chicago	IL	574	2.3	1.1	96.0	50.0	56.0	7.3	96.0
Bluford	Guilford	NC	401	3.7	92.6	1.8	48.4	0.0	21.9	45.6
Bunche	Chicago	IL	396	0.0	100.0	0.0	49.5	0.0	7.0	99.0
C. F. Hard	Bessemer	AL	398	0.0	99.8	0.0	46.7	1.0	14.6	88.4
Central	Central	KS	131	95.0	0.0	2.0	47.0	0.0	6.0	51.0
Daniel Webster	Chicago	IL	636	0.0	100	0.0	44.0	0.0	5.0	98.3
Dewey Elem.	Chicago	IL	436	0.0	99.5	0.0	51.0	0.0	23.0	100
Edward E. Dunne	Chicago	IL	560	1.0	99.0	0.0	76.0	0.0	7.0	97.0
Eutaw	Greene	AL	316	1.0	99.0	0.0	48.0	0.0	5.0	90.0
Greenwood	Bessemer	AL	395	13.0	73.0	14.0	53.0	1.0	11.0	84.0
Gulfview	Hancock	MS	520	94.0	2.6	1.0	49.0	1.0	18.0	71.0
Jamestown	Guilford	NC	496	40.1	51.6	4.6	47.6	5.8	16.0	46.6
Sigel Elem.	St. Louis	MO	302	8.3	82.8	2.7	45.4	6.8	18.2	96.4
South Delta	South Delta	MS	640	5.5	93.5	1.0	47.8	0.0	5.5	100.0
Control school means			435	25.9	65.3	7.3	50.1	4.8	11.1	77.4

Note: ESL = English as a second language.
[a]Selected in the initial phase of the study, during 2001.

Table 2

Comparison of Baseline Characteristics of Success for All (SFA) Schools (*N* = 18) and Control Schools (*N* = 17).

Variable	Condition	*N*	*M*	*SD*	*t*
PPVT	SFA	18	92.21	10.00	
	Control	17	90.29	9.53	0.58
Enrollment	SFA	18	440	194	
	Control	17	435	136	0.09
% Female	SFA	18	49.67	5.77	
	Control	17	50.09	7.01	-0.19
% Minority	SFA	18	61.23	42.54	
	Control	17	72.55	39.65	-0.81
% ESL	SFA	18	9.47	20.43	
	Control	17	4.80	13.52	0.43
% Special education	SFA	18	9.48	5.42	
	Control	17	11.09	6.49	-0.80
% Free lunch	SFA	18	66.34	32.11	
	Control	17	77.39	26.08	-1.11

Note: PPVT = Peabody Picture Vocabulary Test; ESL = English as a second language.

Table 3

Multilevel Models Predicting Student and School-Level Literacy Outcomes for the Longitudinal Sample.

	Literacy Outcomes								
	Word Attack			Word Identification			Passage Comprehension		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement									
Intercept	-0.01	0.05	-0.20	-0.03	0.05	-0.66	-0.03	0.04	-0.74
Mean PPVT pretest	0.23	0.06	3.82**	0.25	0.05	5.52**	0.33	0.04	7.85**
SFA assignment	0.33	0.11	3.03**	0.22	0.10	2.24*	0.21	0.09	2.37*
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.08	155.29**	32	0.06	117.51**	32	0.04	95.09**	32
Within-school variation	0.84			0.87			0.85		

Note: PPVT = Peabody Picture Vocabulary Test; SFA = Success for All.
* $p < .05$; ** $p < .01$.

Table 4

Multilevel Models Predicting Student and School-Level Literacy Outcomes for the Combined Longitudinal and In-mover Sample.

	Literacy Outcomes								
	Word Attack			Word Identification			Passage Comprehension		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement									
Intercept	-0.02	0.05	-0.32	-0.03	0.05	-0.50	-0.02	0.04	-0.53
Mean PPVT pretest	0.24	0.06	4.00**	0.29	0.05	5.98**	0.33	0.04	7.96**
SFA assignment	0.36	0.11	3.27**	0.24	0.11	2.29*	0.21	0.09	2.36*
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.09	248.18**	32	0.09	231.71**	32	0.06	159.67**	32
Within-school variation	0.82			0.83			0.84		

Note: PPVT = Peabody Picture Vocabulary Test; SFA = Success for All.
* $p < .05$; ** $p < .01$.

Table 5

Longitudinal and School-wide Success for All Effect Sizes by Year of Implementation

Outcome	Year 1 (Grade K)	Year 2 (Grade 1)	Year 3 (Grade 2)
Longitudinal Outcomes	*n* = 2,083	*n* = 1,606	*n* = 1,445
Letter Identification	-0.09	0.14	
Word Identification	0.08	0.21	0.22
Word Attack	0.31	0.30	0.33
Passage Comprehension	-0.12	0.12	0.21
School-wide Outcomes	*n* = 2,409	*n* = 2,195	*n* = 2,108
Letter Identification	-0.09	0.16	
Word Identification	0.09	0.19	0.24
Word Attack	0.32	0.29	0.36
Passage Comprehension	-0.10	0.12	0.21

Note: Effect sizes reported for "Longitudinal Outcomes" are derived from the yearly samples of hildren from the baseline kindergarten cohort who remained at the Success for All and control schools from the Year 1 fall baseline through the spring posttest for the applicable year. Effect sizes noted under "School-wide Outcomes" are derived from the yearly samples of students from the "Longitudinal Outcomes" samples plus all students who moved into the Success for All and control schools at any time between the Year 1 fall baseline and the spring posttest for the applicable year.

Appendix: Major Elements of Success for All

Success for All is a schoolwide program for students in grades pre-K to six which organizes resources to attempt to ensure that virtually every student will acquire adequate basic skills and build on this basis throughout the elementary grades, that no student will be allowed to "fall between the cracks." The main elements of the program are as follows:

Schoolwide Instructional Processes. Instruction employs cooperative learning, which maintains student engagement and motivation, to teach metacognitive strategies. The cycle of instruction includes direct instruction, guided peer practice, assessment, and feedback on progress to students. Features of the direct instruction include high time-on-task, brisk pacing, and systematic routines.

***Schoolwide Curriculum*. S**chools implement research-based reading, writing, and language arts programs in all grades, K-6.

The SFA kindergarten is a full-day program where children learn language and literacy, math, science, and social studies concepts through 16 two-week thematic units.

The reading component in grades K-1 contains a systematic phonemic awareness and phonics program that includes mnemonic picture cards and embedded video segments that support phonics and vocabulary development. It uses phonetically regular shared stories that students read to one another in pairs.

In grades 2-6, students use novels or basals but not workbooks. This program emphasizes cooperative learning and partner reading activities, comprehension strategies such as summarization and clarification built around narrative and expository texts, writing, and direct instruction in reading comprehension skills. At all levels, students are required to read books of their own choice for twenty minutes at home each evening. Cooperative learning programs in writing/language arts are used in grades K-6.

***Tutors*.** In grades 1-3, specially trained certified teachers and paraprofessionals work one-to-one with any students who are failing to keep up with their classmates in reading. Tutorial instruction is closely coordinated with regular classroom instruction. It takes place 20 minutes daily during times other than reading periods.

***Quarterly Assessments and Regrouping*.** Students in grades 1-6 are assessed every quarter to determine whether they are making adequate progress in reading. This information is used to regroup students for instruction across grade lines, so that each reading class contains students of different ages who are all reading at the same level. Assessment information is also used to suggest alternate teaching strategies in the regular classroom, changes in reading group placement, provision of tutoring services, or other means of meeting students' needs.

***Solutions Team*.** A Solutions Team works in each school to help support families in ensuring the success of their children, focusing on parent education, parent involvement, attendance, and student behavior. This team is composed of existing or additional staff such as parent liaisons, social workers, counselors, and assistant principals.

***Facilitator*.** A program facilitator works with teachers as an on-site coach to help them implement the reading program, manages the quarterly assessments, assists the Solutions Team, makes sure that all staff are communicating with each other, and helps the staff as a whole make certain that every child is making adequate progress.