RUNNING HEAD: Success for All

The National Randomized Field Trial of Success for All: Second-Year Outcomes

Geoffrey D. Borman

University of Wisconsin—Madison

Robert E. Slavin

Johns Hopkins University

Alan C.K. Cheung, Anne M. Chamberlain, Nancy A. Madden, Bette Chambers

Success for All Foundation

Contact Information:

Geoffrey D. Borman, Associate Professor (Corresponding author). University of Wisconsin—Madison, Educational Leadership and Policy Analysis, Educational Psychology, and Educational Policy Studies, 1161D Educational Sciences Building, 1025 West Johnson Street, Madison, WI 53706. 608-263-3688; 608-265-3135 (Fax); gborman@education.wisc.edu

Robert E. Slavin, Principal Research Scientist and Director of the Center for Data-Driven Reform in Education. Johns Hopkins University, 200 West Towsontown Blvd., Baltimore, MD 21204. rslavin@cddre.org

Alan C. K. Cheung, Associate Professor. Hong Kong Institute of Education, Department of Educational Policy and Administration, 10 Lo Ping Road, Taipo, New Territories, Hong Kong SAR, China.

Anne M. Chamberlain, Research Scientist. Success for All Foundation, 200 West Towsontown Blvd., Baltimore, MD 21204. achamberlain@successforall.net

Nancy A. Madden, President and CEO. Success for All Foundation, 200 West Towsontown Blvd., Baltimore, MD 21204. nmadden@successforall.net

Bette Chambers, Vice President of Development. Success for All Foundation, 200 West Towsontown Blvd., Baltimore, MD 21204. bchambers@successforall.net

Biographical Sketches:

Geoffrey D. Borman is an Associate Professor Educational Leadership and Policy Analysis, Educational Policy Studies, and Educational Psychology at the University of Wisconsin—Madison, 1161D Educational Sciences Building, 1025 West Johnson St., Madison, WI 53706; email *gdborman@education.wisc.edu*. His areas of interest include stratification, policy, research methodology, and evaluation.

Robert E. Slavin is a Principal Research Scientist and the Director of the Center for Data-Driven Reform in Education, Johns Hopkins University, 200 W. Towsontown Blvd., Baltimore, MD 21204; email *rslavin@cddre.org*. His areas of interest include cooperative learning, comprehensive school reform, evidence-based policy, English-language learners, and research methods.

Alan C.K. Cheung is an Associate Professor at the Hong Kong Institute of Education, Department of Educational Policy and Administration, 10 Lo Ping Road, Taipo, New Territories, Hong Kong SAR, China. His areas of specialization include large-scale assessment, research reviews, research methods, and private education.

Anne M. Chamberlain is a Research Scientist at the Success for All Foundation, 200 W. Towsontown Blvd., Baltimore, MD 21204; email *achamberlain@successforall.net*. Her areas of interest include program evaluation, mixed methodology, and participatory evaluation.

Nancy A. Madden is the President and CEO of the Success for All Foundation, 200 W. Towsontown Blvd., Baltimore, MD 21204; email *nmadden@successforall.net*. Her areas of interest include cooperative learning, comprehensive school reform, evidence-based policy, English-language learners, and research methods.

Bette Chambers is the Vice President of Development at the Success for All Foundation, 200 W. Towsontown Blvd., Baltimore, MD 21204; email *bchambers@successforall.net*. Her areas of interest include early childhood education, early reading, and the use of technology in education.

Abstract

This article reports literacy outcomes for a two-year longitudinal student sample and a combined longitudinal and in-mover sample, both of which were nested within 38 schools. Using a cluster randomization design, schools were randomly assigned to implement Success for All or control methods. Hierarchical linear model analyses for the longitudinal sample revealed statistically significant school-level effects of assignment to Success for All on three of four literacy outcomes measured. Effects were as large as one quarter of a standard deviation—a learning advantage relative to controls exceeding half of a school year. Impacts for the combined longitudinal and in-mover sample were smaller in magnitude and more variable. The results correspond with the Success for All program theory, which targets school-level reform through a multi-year sequencing of intensive literacy instruction.

Key Words: educational policy, experimental design, school reform

Success for All is one of the most widely implemented, widely researched, and widely critiqued educational interventions in the United States today. More than 1,200 schools, mostly high-poverty Title I schools, in 46 states are currently implementing the 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.

The program focuses on prevention and early, intensive intervention designed to detect and resolve reading problems as early as possible, before they become serious. 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. Using the program’s benchmark assessments, teachers assess each student’s reading performance at eight-week intervals and make regrouping changes based on the results. Instead of being placed in special classes or retained in grade, most students who need additional help receive one-on-one tutoring to get them back on track.

The use of cooperative learning methods also helps children develop academic skills and encourages them to engage in teambuilding activities and other tasks that deal explicitly with the development of interpersonal and social skills (Slavin, 1995). In addition, a Success for All school establishes a Solutions Team, serving to increase parents' participation in school generally, to mobilize integrated services to help Success for All families and children, and to identify and address particular problems such as irregular attendance, vision correction, or problems at home. Finally, 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. These are the main features of Success for All, as originally conceived in 1987 and as currently disseminated (see Appendix for more details concerning the Success for All program components).

Of the 33 comprehensive school reform programs reviewed in a recent meta-analysis by Borman, Hewes, Overman, and Brown (2003), Success for All was one of only three that had positive and statistically significant achievement effects across a large number of rigorous quasi-experimental studies. The evidence reviewed by Borman and colleagues from 46 separate 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 ($d = .20$). In addition, quasi-experimental evidence from a study by Borman and Hewes (2002) demonstrated that students' multi-year participation in Success for All was associated with notable gains in reading and other school-based outcomes that were sustained through the completion of eighth grade, several years after they had left the Success for All elementary schools.

Though compelling in terms of its scope and results, this prior research has several important limitations. First, the vast majority of these previous studies of Success for All have used a quasi-experimental matched comparison-group design, in which experimental and control schools were matched on pretests and other demographic characteristics and then compared each year on the subsequent posttests. In recent years, an increasing body of evidence has emerged suggesting that such comparison-group studies in social policy (e.g., employment, training, welfare-to-work, education) often produce inaccurate 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).

Further, the authors of nearly all previous studies of Success for All have employed designs that attempt to match program and control schools, but 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 the estimates and leads the researcher to over-reject null results. Finally, because the early studies by the developers, and by many of the third-party evaluators, have involved small numbers of treatment sites, much of the prior work on Success for All may correspond with that which Cronbach et al. (1980) termed the "superrealization" stage of program development. That is, with the researchers actively involved in assuring that they are studying high-quality implementations in a select number of schools, some of these earlier evaluations may 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.

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. This ongoing 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. A distinguished group of scholars, including C. Kent McGuire, Steven Raudenbush, Rebecca Maynard, Jonathan Crane, and Ronald Ferguson, was appointed to serve as an oversight committee to ensure that all study procedures were appropriate. 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. (2005). 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. In this article, we track the second-year impact estimates for the main sample from this three-year study.

The Success for All Program Theory

Beyond the applied and empirical evidence supporting Success for All, the central features of the program's theory of action—its comprehensive approach to school reform and its focus on high-quality literacy instruction—can be understood on the basis of two distinct conceptual frameworks. A review by Ramey and Ramey (1998) of the major findings from rigorous evaluations of early interventions revealed the features associated with those programs that exhibited the strongest and most persistent effects on children's outcomes. The framework, *biosocial developmental contexualism*, derived from this review predicts that early interventions focused on education and child development are not likely to succeed unless they are intensive, high-quality, and ecologically pervasive efforts. The Success for All program design exemplifies this comprehensive and intensive approach to early intervention in two important ways.

First, rather than a targeted, and potentially fragmented, remedial approach for improving the school outcomes of children placed at risk, Success for All emphasizes a school-wide focus on reform and improvement. Though components, such as one-on-one tutoring, are in place to target children who need extra help, the program theory emphasizes the school as the critical unit of intervention and espouses a comprehensive school-level reform plan to help meet the needs of all children attending high-poverty schools. Second, the learning experiences offered through Success for All are delivered to students directly, efficiently, and effectively in classrooms that are regrouped according to students' current achievement levels. The intensity of these services is reflected by characteristics such as students' daily exposure to an uninterrupted 90-minute reading period and the program's emphasis on a multi-year approach to literacy instruction and learning.

Beyond the comprehensiveness and intensity of Success for All, the core focus of the program is on literacy. The 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). The specific sequencing of literacy instruction across the grades is a defining characteristic of the Success for All instructional program that can be understood through a larger body of empirical research and theory on beginning reading.

The Success for All 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 the strong consensus among researchers that phonemic awareness is the best single predictor of reading ability, not just in the early grades (Ehri & Wilce, 1980; 1985; Perfetti, Beck, Bell, & Hughes, 1987) but throughout the school years (Calfee, Lindamood, & Lindamood, 1973; Shankweiler et al., 1995). 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.

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.

Implications and Hypotheses

The evidence and theory supporting Success for All 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. Second, given the importance of program intensity and Success for All's multi-year approach to literacy instruction, we hypothesized that the program effects estimated for the longitudinal sample of students who had experienced the full program across two 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 2 posttest.

Third, consistent with the program theory related to the sequencing of the literacy instruction, which focuses on phonemic awareness skills initially and broader reading skills later, we hypothesized that the second-year program effects would spread into other tested literacy domains beyond the Word Attack subtest.[1] Unlike the first-year impacts, which were restricted to the Word Attack subtest, we hypothesized that we would begin to find treatment effects on the Letter and Word Identification and Passage Comprehension subtests as well. 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 two-year longitudinal sample.

Method

Sample selection

The total sample of 41 schools was recruited in two phases. The initial pilot 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 control 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 the pilot 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, have failed to document 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 and that few if any control students benefited directly from school-level Success for All services.

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.

After the first year, 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. In addition, a school in Chicago refused to implement its assigned treatment in grades 3-5, but did allow continued assessment. Because this school was implementing the 3-5 treatment rather than the K-2 Success for All treatment, it does not have a highly important consequence for the current analysis of K-2 treatment effects. In future analyses of grade 3-5 program effects, this school will be included as an intent-to-treat case. The loss of the three St. Louis schools reduced the second-year analytic sample to 38 schools, 20 implementing Success for All in grades K-2 and 18 in grades 3-5 (hereafter, K-2 schools will be referred to as “experimental” and 3-5 as “control”).

The experimental and control schools included in the Year 2 analyses of outcomes are listed in Table 1. The sample is largely concentrated in the urban Midwest (Chicago, St. Louis, and Indianapolis) and 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 74% 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, 57% of the sample is African American, compared to about 40% within the typical Success for All school, and 11% of the sample is Hispanic, compared to the national average of 35%. The percent of white students, 29%, 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 2 outcomes. As the results suggest, the 20 experimental and 18 control schools were 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. (2005), 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 have made quarterly implementation visits to each school, as is customary in all implementations of the Success for All program. These 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 eight 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.

As of January 2005, all grade K-2 classes in 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 many 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 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. In no case did the trainers observe teachers in non-Success for All classes implementing Success for All components. It is possible that some ideas or procedures from Success for All did influence instruction in the non-treatment control grades, but any such influence was apparently subtle. Instructional materials and core procedures were clearly distinct from each other in the treatment and control grades.

Measures

Students in grades K-1 were pretested on the Peabody Picture Vocabulary Test and then individually posttested on the Woodcock Reading Mastery Tests—Revised (WMTR). The six schools from the first phase of recruitment were pretested in fall 2001 and posttested during the spring 2002 and the spring of 2003. The 35 schools from the main sample were pretested in fall 2002 and posttested in spring 2003 and spring 2004. The pilot and main samples were combined for the analyses. Because the metrics of the 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.

Pretests. All children were individually assessed in fall, 2001 (first phase) or fall, 2002 (second phase) on the Peabody Picture Vocabulary Test (PPVT III). The few children who were Spanish-dominant were pretested in Spanish on the Test de Vocabulario en Imagenes Peabody.

Posttests. During the spring of 2002 and 2003 (first phase) and the spring of 2003 and 2004 (second phase)—and during each subsequent spring through 2005—students in the main longitudinal cohorts (which started in K-1) were individually assessed on the four subtests of the Woodcock Reading Mastery Tests—Revised (WMTR): Letter Identification, Word Identification, Word Attack, and Passage Comprehension. The WMTR was normed on a national sample of children and the internal reliability coefficients for the four subtests used were 0.84, 0.97, 0.87, and 0.92, respectively. Children in the initial cohorts are being followed into any grade as long as they remain in the same school; retention does not change their cohort assignment. They are also being followed into special education. Children who entered Success for All or control schools after fall, 2002 are also posttested each year and included in analyses that combine the baseline cohorts and in-moving student cohorts. Children who are English language learners but are taught in English are posttested in English each year. In this analysis, we focused on the outcomes for the Year 2 posttests, which were administered to students in the pilot schools during spring 2003 and students in the main sample of schools during 2004.

Results

The prior review of baseline data for the school-level sample revealed no important differences between treatment and control schools and 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 second-year analyses of achievement outcomes, we begin by assessing whether there was 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 samples were composed of 1,672 students from the 20 Success for All treatment schools and 1,618 students from the 18 control schools. Listwise deletion of student cases with missing posttest data did not cause differential attrition rates by program condition, χ^2 (1, *N* = 5,736) = 2.02, *p* = 0.67, leaving 56% of the baseline sample of 2,966 treatment students and 58% of the 2,770 baseline controls for the preliminary analyses. The data and sample attrition occurred for three reasons. Of the students who were excluded from the analysis, 1,195 (49%), were dropped because they had moved out of the school before the Year 2 posttests were administered and, thus, had no outcome data, and 1,021 (42%) remained in the treatment and control schools but were missing either pretest, posttest, or other important demographic data. Finally, the closure of three participating schools prevented posttesting of 230 students (9%) in Year 2.

We compared the pretest scores of those treatment students who were dropped from the analyses to the pretest 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* (0.50), *p* = 0.62 (two-tailed), suggesting that the baseline achievement levels of the treatment and the control group students who were dropped from the analyses were statistically equivalent.

To address the issue of external validity, we also 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 (-6.74), p < .05 (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 = 5,736) = 3457.99, p < .001. Thus, both low-achieving and mobile students from the sample schools were underrepresented in the analyses, but this does compromise 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 largest 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 is most likely to result in downward biases of the treatment effect estimates. Second, because the primary missing data mechanism was mobility from the study schools, analysis of this longitudinal sample 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).

In addition to the analyses of impacts for the longitudinal sample, we conducted a second set of analyses of treatment effects for the combined longitudinal and in-moving student samples. These analyses included the longitudinal sample of 3,290 students, who remained at the Success for All and control schools from baseline through the second-year posttest, 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 does 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 affords a type of school-level intent-to-treat analysis of the program.

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 most appropriate method (Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). We applied Raudenbush's (1997) relatively recently 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 fully specified level 1, or within-school model nested students within schools with an indicator of the student's baseline grade level (-0.5 = kindergarten and 0.5 = first grade). The linear model for this level of the analysis was expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{GRADE})_{ij} + r_{ij},$$

which represents the spring posttest achievement for student *i* in school *j* regressed on grade level plus the level-1 residual variance, r_{ij}, that remained unexplained after accounting for the grade level of the students.

In this model, each student's grade level was centered around zero. In this way, we controlled for school-to-school differences in the proportions of kindergartners and first graders. With grade coded as -0.5 for students in kindergarten at baseline and 0.5 for students in first grade at baseline, the level 2 school-specific intercept represented the overall average school performance of students from across both grade levels. We treated the within-school grade-level gap—the difference between the posttest scores of baseline kindergarten and first grade students in school *j*—as fixed at level 2 because it was intended only as a covariate and we have no empirical or theoretical reason to model this source of between-school variability as an outcome.

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.[2] The fully specified level 2 model was written as

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

$$\beta_{1j} = \gamma_{10}$$

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}. The within-school posttest difference between baseline kindergarten and first grade students, β_{1j}, was specified as fixed, predicted only by an intercept.[3]

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 2 posttests. Across the four outcomes, the impact estimate for Success for All assignment ranged from a standardized effect of approximately d = 0.12, for Passage Comprehension, to d = 0.25 for the Word Attack subtest. Three of the four treatment effects were statistically significant, the impact on Word Attack of 0.25 at the $p < .01$ level of confidence, the impact on Letter Identification of 0.18 at the $p < .05$ level, and the treatment effect on Word Identification of 0.16 at the $p < .10$ level of confidence. In all four models, the school-level mean pretest covariate was an important predictor of the outcome and the fixed within-school posttest difference between baseline kindergarten and first grade students was between nearly half of one standard deviation and more than three quarters of one standard deviation.

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

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 four Year 2 literacy outcomes for the combined longitudinal and in-mover sample. The Success for All impact estimates across these multilevel models were relatively smaller in magnitude relative than the effects found for the longitudinal sample in Table 3. Across the four outcomes, the impact estimate for Success for All assignment ranged from a standardized effect of approximately *d* = 0.09, for Passage Comprehension, to *d* = 0.19 for Word Attack. In addition to the somewhat smaller effects, the impacts were estimated with greater uncertainty, as indicated by the larger standard errors for the treatment coefficients. The smaller impacts and greater uncertainty of the impact estimates are, in part, explained by the greater variability among students in their exposure to the Success for All treatment. In this sample, students' participation in the Success for All treatment ranged from two years to several months.

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

INSERT TABLE 4 HERE

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

Discussion

The second-year results of the randomized evaluation of Success for All show school-level impacts of assignment to the intervention that are extraordinarily consistent with key aspects of the program theory and with past meta-analytic evidence on program effects. The cluster randomized trial provided strong evidence that the Success for All comprehensive school reform strategy is capable of producing school-level effects of both statistical and practical importance. The magnitude of these effects across the four literacy outcomes are summarized in Table 5 as effect sizes and months of additional learning relative to control schools. When converted to additional months of learning, the practical effects of the program appear substantial for Word Attack and relatively large for the other literacy measures. For Word Attack, the longitudinal sample's learning advantage relative to the controls exceeds half of one nine-month school year.

Consistent with the program theory related to the sequencing of the literacy instruction, which focuses on phonemic awareness skills initially and broader reading skills later, we found that the program effects began to spread into other tested literacy domains beyond the Word Attack subtest. Unlike the first-year impacts, which were restricted to the Word Attack subtest, we found statistically significant treatment effects on the Letter and Word Identification subtests as well.

The reliability and magnitude of these effects, though, are sensitive to the amount of exposure students had to the intervention. As the program theory suggests, due to the importance of program intensity, the multi-year sequencing of literacy instruction, and the general importance of learning phonemic awareness skills early, the effects across the four reading skills tested are greater and more reliable for the sample of students from the two-year longitudinal sample. Therefore, although Success for All is a comprehensive school-wide intervention, which may theoretically advance the academic outcomes of the whole school, students seem to need longitudinal exposure to the program in order to make the largest and most reliable gains. When one considers literacy achievement as the primary outcome, the multi-year sequencing of literacy instruction and the initial skill development in the area of phonemic awareness appear to be important mechanisms that drive broader improvements in reading.

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

INSERT TABLE 5 HERE

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

Interestingly, the effect sizes reported in Table 5 are quite similar to the overall average effect size of d = 0.20 estimated by Borman et al. (2003) in their synthesis of the quasi-experimental evaluations of Success for All and Roots and Wings. This result is consistent with findings reported by Lipsey and Wilson (1993) and Heinsman and Shadish (1996), who concluded that the mean value of the findings of a large number of non-experimental studies tends to approximate that of experiments that address the same question. In the context of the current study, the meta-analytic summary of the results of 46 quasi-experimental studies of Success for All produced an effect size estimate that was essentially the same as the impact estimates from this randomized experiment.

The effects are smaller, though, than those found in the early small-scale matched comparison-group studies performed by the developers using the same individually administered measures as those used in the present study. The earlier quasi-experimental studies, such as those reported by Madden, Slavin, Karweit, Dolan, and Wasik (1993) and Slavin and Madden (2001), found effect sizes ranging from approximately d = 0.30 to d = 0.50. The difference in the impact estimates from the current study and these earlier studies of Success for All are most likely explained by two points drawn from Lipsey's (2003) examination of meta-analytic data concerning the effects of intervention programs to prevent or reduce juvenile delinquency.

First, Lipsey's (2003) analysis demonstrated that studies employing random assignment designs were associated with smaller mean effects than those using quasi-experimental designs. Second, Lipsey found that the program type, that is a research and demonstration project versus a routine practice program, was also an important moderator for understanding differences in study outcomes. Similar to the concept of the superrealization stage of program development articulated by Cronbach et al. (1980), research and demonstration projects typically involve a small-scale pilot of an intervention designed by the developers to show the effects of the program when it is operating at its best. Not surprisingly, Lipsey's (2003) meta-analytic data revealed that the research and demonstration programs were associated with larger mean effects than routine practice programs. Therefore, both the methodological design and the program type may have important implications for understanding the differences between the achievement impacts estimated by the early Success for All quasi-experiments of demonstration programs and by the current national randomized field trial of the program operating at scale.

Putting the Results in Context

Similar recent efforts to study widespread implementations of educational interventions through a randomized design have yielded few promising outcomes. For instance, the evaluation of 44 schools implementing 21st-Century Community Learning Centers (21st CCLC) programs, which help schools and districts partner with community organizations to provide after-school programs, revealed no positive impacts on academic outcomes, homework completion, student behavior, or parent involvement (James-Burdumy et al., 2005). Similarly, the experimental study of 18 Even Start family literacy programs revealed no treatment effects on literacy and other measures for the children and parents served by the programs (St. Pierre et al., 2003). Indeed, aside from the Tennessee Student/Achievement Ration (STAR) class-size reduction study (Finn & Achilles, 1999), there are few examples of substantial field trials in education that have yielded positive effects of practical and statistical significance across a large number of sites.

Given the relatively high costs associated with implementing Success for All, though, it is reasonable to ask whether the impacts are worth the investment. One answer comes from an earlier cost-effectiveness study by Borman and Hewes (2003), who drew on quasi-experimental evidence from the original implementations of the program in Baltimore. The authors found that those students who had attended Success for All elementary schools completed eighth grade at a younger age, with better reading and math achievement outcomes, fewer special education placements, and less frequent retentions in grade at a cost that was essentially the same as that allocated to educating their comparison-group counterparts. Across the longitudinal period from first through eighth grade, the more frequent retentions and special education placements for those from the comparison group wound up costing the same amount as the additional per-pupil expenses of implementing Success for All. As these outcomes suggest, the educational practices of prevention and early intervention, as modeled by Success for All, were more educationally effective, and equally expensive, relative to the traditional remedial educational practices of retention and special education. Though no formal cost-effectiveness component has yet been designed and carried out for the current study, it is likely that these findings from previous implementations of Success for All generalize to the contexts studied here.

Conclusion

The current findings of statistically significant positive achievement effects from a large-scale implementation of a randomized field trial of a routine practice program are unusual for studies in education. A study of this kind, with 38 schools serving more than 20,000 children in districts throughout the United States, provides a rigorous assessment of the impact that can be expected when a program is scaled up in a real-world policy context. The effects may 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. The research includes schools with good and poor implementations, and with school and district staffs potentially less committed to the program than is usual because they were not paying for it. In the vast majority of experimental schools, it also included a design that required implementation of the instructional program across only three grades rather than across the whole school. In the remaining six pilot schools included in the study, the control schools were provided the same supplemental funds offered the experimental schools and were permitted to engage in any reform other than Success for All. For these reasons, we believe that the impact estimates are realistic but also somewhat conservative.

Though large-scale cluster randomized trials in field settings are rare in education and though findings of widespread treatment effects are rarer still, these outcomes attest that such studies are possible and can produce unbiased estimates of program effects. Future reports will illuminate many other aspects of the implementation and program impacts. These will include studies of the quality of program implementation, effects for subgroups, and effects in the upper elementary grades.

Footnotes

[1] 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. The degree to which students are able to use their developing phonemic awareness is directly assessed using the Word Attack subtest, Woodcock Reading Mastery Tests-Revised, which is composed of test items that ask the child to decode nonsense words.

[2] We formulated other multilevel models that included the broader array of school-level covariates listed in Table 3. After including the school mean pretest covariate, though, these more complex models did not explain appreciably more between-school variance and did not improve the precision of the Success for All treatment effect estimates. For these reasons, we used the more parsimonious models presented.

[3] 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, $1 - \beta$; 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 $p < .10$, power, $1 - \beta$, equal to 0.80, and degrees of freedom equal to $J = 38$ 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 .08 to .12, and employing Bloom's (2005) minimum detectable effect multiplier, we calculated minimum detectable effects of approximately $d = .20$ to $d = .30$. That is, our design had adequate power to detect school-level treatment-control differences of at least .20 to .30 standard deviations.

References

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

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., & Hewes, G.M. (2003*).* 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., Madden, N., & Chambers, B. (2005). Success for All: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27,* 1-22.

Calfee, R.C., Lindamood, P., & Lindamood, C. (1973). Acoustic-phonetic skills and reading: Kindergarten through twelfth grade. *Journal of Educational Psychology, 64,* 293-298.

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.

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.C., & Wilce, L.S. (1980). The influence of orthography on readers' conceptualization of the phonemic structure of words. *Applied Psycholinguistics, 1*, 371-385.

Ehri, L.C., & Wilce, L.S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic? *Reading Research Quarterly 20,* 163-179.

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

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

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.

James-Burdumy, S., Dynarski, M., Moore, M., Deke, J., Mansfield, W., & Pistorino, C. (2005). *When schools stay open late: The national evaluation of the 21st Century Community Learning Centers Program: Final report*. Washington, DC: U.S.

Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available at http://www.ed.gov/ies/ncee.

Lipsey, M.W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *Annals of the American Academy of Political and Social Science, 587*, 69-81.

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.

Madden, N.A., Slavin, R.E., Karweit, N.L., Dolan, L.J., & Wasik, B.A. (1993). Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal*, *30*, 123-148.

Perfetti, C.A., Beck, I, Bell, L., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33,* 283-319.

Ramey, C.T., & Ramey, S.L. (1998). Early intervention and early experience. *American Psychologist, 53,* 109-120.

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

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

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.

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

St. Pierre, R., Ricciuti, A., Tao, F., Creps, C., Swartz, J., Lee, W., & Parsad, A., (2003). *Third national Even Start evaluation: Program impacts and implications for improvement.* Washington, DC: U.S. Department of Education, Planning & Evaluation Service.

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 members of the Oversight Committee for this study: C. Kent McGuire, Steven Raudenbush, Rebecca Maynard, Jonathan Crane, and Ronald Ferguson.

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.

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:

A Schoolwide Curriculum. Success for All schools implement research-based reading, writing, and language arts programs in all grades, K-6. The reading program in grades K-1 emphasizes language and comprehension skills, phonics, sound blending, and use of shared stories that students read to one another in pairs. The shared stories combine teacher-read material with phonetically regular student material to teach decoding and comprehension in the context of meaningful, engaging stories.

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 1-6.

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.