Running Head: SUCCESS FOR ALL

Success for All:

First-Year Results from the National Randomized Field Trial

Geoffrey D. Borman

University of Wisconsin—Madison

Robert E. Slavin

Johns Hopkins University

Alan Cheung, Anne Chamberlain, Nancy Madden, Bette Chambers

Success for All Foundation

Success for All:

First-Year Results From the National Randomized Field Trial

Abstract

This paper reports first-year achievement outcomes of a national randomized evaluation of Success for All, a comprehensive reading reform model. Forty-one schools were recruited for the study and were randomly assigned to implement Success for All or control methods. No statistically significant differences were found on pretests or demographic characteristics between experimental and control groups. Hierarchical linear model analyses revealed a statistically significant school-level effect of assignment to Success for All of nearly one quarter of a standard deviation on the individually administered Word Attack outcome, but there were no school-level differences on the three other posttest measures. These results are similar to those of earlier matched experiments and correspond with the Success for All program theory.

Success for All:

First-Year Results From the National Randomized Field Trial

Success for All is, in many ways, the prototype for the comprehensive school reform and evidence-based education policy movements. Of 33 comprehensive school reform programs reviewed in a recent meta-analysis, it was one of only three that had positive and statistically significant achievement effects across a large number of rigorous studies (Borman, Hewes, Overman, & Brown, 2003). Further, Success for All has built unprecedented capacity to disseminate and maintain quality implementations of its programs, currently in more than 1400 schools in 47 states. First at Johns Hopkins University and currently at the nonprofit Success for All Foundation, Success for All is one of very few programs that can and does add more than a hundred schools to its network each year. A staff of developers, evaluators, and support staff located throughout the U.S. has enabled the program to serve every kind of school, rural as well as urban, and to adapt to every circumstance and need. Table 1 summarizes the main elements of the program.

Insert Table 1 about here

The extent, breadth, and consistent positive effects of the many quasi-experimental evaluations of Success for All have given this program a research base that extends beyond that which has been ordinarily required in educational policy or practice. However, in the current policy context this is not enough. For the first time, Congress and other education policymakers are beginning to require evidence of effectiveness for

programs seeking certain types of funding. In 1997, this trend began with the Comprehensive School Reform Demonstration (CSRD) program, which provided grants to schools to adopt "proven, comprehensive reform models." More recently, the Reading Excellence Act (REA) and its larger successor, Reading First, required that grant funds be used to help schools adopt those programs that incorporate "scientifically based principles" of reading instruction.

Policies of this kind greatly raise the stakes for research in education. Though in practice CSRD, REA, and Reading First grants have not held the schools to high standards of evidence, the movement to base educational policies on research must ultimately depend on building a stronger research base for replicable programs. Success for All is a logical program to be held to a new and higher standard of evidence, because of the substantial foundation of research it already has built from quasi-experimental studies and because it is the most widely disseminated program funded by the Comprehensive School Reform legislation.

In order to set a higher standard of evidence for policy-relevant research in education, and to address some of the past criticisms of research on Success for All, the Success for All Foundation applied for and received funding to carry out the first *randomized* evaluation of the program. This grant was one of six awarded by the Office of Educational Research and Improvement (OERI) to study comprehensive school reform models. The other five projects are studying various combinations of comprehensive school reform models, including Success for All. The Success for All Foundation grant, though, is the only one to include a randomized experiment.

The OERI grant was originally proposed as a first-party evaluation with strong, independent review by an Oversight Committee composed of distinguished scholars not previously connected to Success for All.[1] However, in response to clarification questions from OERI, the Success for All Foundation agreed to have the data collected by a third party, and ultimately contracted with the National Opinion Research Center at the University of Chicago.

This article presents the first-year findings from this national randomized field trial. We begin by examining the internal and external validity of the experiment. Specifically, did randomization yield treatment and control schools with comparable characteristics and did any sample and data attrition over the first year jeopardize this baseline comparability? Regarding the external validity of the experiment, how representative of the population of Success for All schools is this sample of schools and did sample and data attrition have an impact on our ability to generalize the results? Finally, we examine the pretest-to-posttest outcomes for the kindergarten and first grade cohorts that were tested from fall to spring during the first year. The two-level hierarchical linear models that we formulated respond to the question: are there fall-to-spring school-level effects of treatment assignment on four measures of literacy achievement?

The Success for All Research Base

Research on Success for All has been done by many investigators in many locations. A total of 54 individual investigators have been authors of studies of the

[1] The Oversight Committee members are: Ronald Ferguson, Harvard University; Steve Raudenbush, University of Michigan; Rebecca Maynard, University of Pennsylvania; Jonathan Crane, Progressive Policy Institute; and Kent McGuire, Temple University and MDRC.

achievement effects of Success for All; 35 of these were at institutions other than Johns Hopkins University or the Success for All Foundation. The cumulative evidence from these studies shows positive effects of Success for All on a variety of measures of student achievement, as well as on assignments to special education, retentions, and other outcomes (Borman et al., 2003; Slavin & Madden, 2001).

Despite the substantial evidence supporting the effectiveness of Success for All, there are still some criticisms of the research that have focused on two fundamental issues (Pogrow, 2000; Walberg and Greenberg, 1999). First, the typical research design in previous evaluations of the program has been a quasi-experimental, untreated control group design, in which schools using Success for All have been compared over time to non-Success for All schools that are matched on demographics, prior achievement, and other factors. Though this is a sound design, all such quasi-experiments have limitations (Shadish, Cook, & Campbell, 2002). Most importantly, they leave open the possibility that selection artifacts could explain some or all of the differences observed between the Success for All and quasi-experimental control groups.

With such a design, there always may be some systematic reason that the experimental group implemented the program while the comparison group did not. Schools whose staffs expressed interest in Success for All and achieved the required 80% majority vote to adopt it may have greater motivation and interest in improving their schools than the control schools' staffs who did not seek out the program. As indicated by the 80% agreement among staff, these schools may have strong cohesion among the teachers or have better leaders. Perhaps the experimental schools have better funding or fewer demands on their resources or energies. Alternatively, perhaps the experimental

schools are experiencing greater difficulties and have a greater need for change. These potential artifacts can make it difficult to know whether it was the characteristics related to selection of Success for All or the components of the Success for All program that caused the improvements in the schools. Most studies of Success for All have been well-designed matched experiments that have minimized selection bias—for example, by designating control schools in advance and by avoiding the use of control schools that rejected the program—but selection bias cannot be ruled out without random assignment.

Second, there has been continuing criticism of those studies of Success for All that have been done by the developers themselves. Borman et al. (2003) found that studies of comprehensive school reform programs that have been authored by the developers tend to report higher estimates of the programs' effects on achievement outcomes. The magnitude of this difference is between one sixth and one-seventh of a standard deviation relative to third-party studies. As Borman and his colleagues noted, one of the most likely explanations for this difference is that when developers are more actively involved in the study of their models, they are also more likely to be actively involved in assuring that they are studying a high-quality implementation. In this respect, many of these studies may represent what Cronbach et al. (1980) termed the "superrealization" stage of program development. Before engaging in broad field trials, interventions are often studied under optimal conditions as assessments of what the program can accomplish at its best. The extent to which the developers' studies and results may generalize across broader implementations of their school reform models, though, is of some concern.

The critics of small-scale quasi–experiments also may argue that small samples of program schools and matched control schools that are, essentially, hand-picked by the developer, can introduce particularly advantageous comparisons that overstate the impact of the school reform program. In a large randomized study, the process of choosing program and control schools is not left to the developer, but is decided by, for instance, the arbitrary flip of a coin. Hand picking schools and making comparisons that "look good" for the program are not possible within a randomized design, especially when the process of randomization is decided by a neutral third party. Success for All has been studied by researchers other than the developer and the biases for Success for All do not appear to be strong. However, as for any other program, the Success for All research base could benefit from greater third-party involvement in broader field trials.

The Role of Randomized Field Trials in Education Policy

Randomized field trials are extremely rare in educational research (Borman, 2003; Boruch, de Moya, & Snyder, 2002; Cook & Payne, 2002). Of course, there are many brief, artificial lab studies that use random assignment, but field trials of sufficient size and duration to provide findings of direct relevance to educators are hardly ever done. Recently, educational research has come under attack for its dearth of randomized experiments, and has been compared unfavorably to medical and other scientific research on this basis.

However, among the small number of experiments in education that have been conducted, several have been extremely influential. For instance, the randomized longitudinal Tennessee Class Size Study (Finn & Achilles, 1999) led directly to massive class-size reduction initiatives in several states, notably California, and to the Clinton

Administration's national class size initiative. The randomized, longitudinal evaluation of the Perry Preschool (Schweinhart, Barnes, & Weikart, 1993) led to substantial expansion of the federal Head Start program and to publicly funded preschool programs in many states and localities. Finally, the promising results from the Abecedarian Project (Ramey & Campbell, 1984), which randomly assigned mothers and their infants to a highly intensive educational childcare program beginning shortly after birth, inspired the U.S. Congress to develop and scale up the Early Head Start program.

Beyond the scientific importance of random assignment, the political importance of rigorous evidence continues to grow. If federal education legislation continues the trend toward linking funding of education programs to evidence of effectiveness, the consequences could be revolutionary. If education reform can be based on rigorous research, then genuine progress in educational practice becomes possible, as in medicine, agriculture, technology, and other parts of modern economies that long ago accepted the idea that progress must be based on rigorous research and development. In place of the famous pendulum swinging, for example, from phonics to whole language or tracking to untracking, there would be widely accepted findings providing a solid basis for policy and practice.

Imagine, for example, that schools implementing Title I programs, bilingual programs, special education programs, or dropout prevention programs were encouraged or required to adopt programs and practices that had evidence of effectiveness from high-quality experimental evaluations. This transformation of the education reform landscape will not, however, take place unless there are convincing demonstrations that replicable programs can accelerate student achievement. These demonstrations must be beyond

reproach, using the most rigorous evaluation methods known. In other words, they must use random assignment, extensive measurement of implementations and outcomes, and longitudinal designs on a large enough scale to ensure both statistical power and generalizability.

The Success for All National Randomized Field Trial

This article presents the first-year findings from a longitudinal study that should add enormously to knowledge about the effectiveness of Success for All in increasing the achievement and school success of students placed at risk. First, it is the first randomized study of Success for All, virtually eliminating selection bias as an alternative explanation for any findings. Second, it is the largest study ever to compare Success for All and control schools, enabling the use of appropriate statistical methods, especially hierarchical linear modeling, with adequate power to detect true differences. Third, the sample size permits rigorous and definitive conclusions about the effects of Success for All on key subpopulations.

Taken together, the project's interrelated activities will provide data of unprecedented richness, detail, and methodological rigor to inform educators about the effects of Success for All, the reasons for those effects, and the conditions under which the effects are most likely to be obtained. In this article, we present analyses of the first-year achievement outcomes of the project and assess the strength of the foundation that has been laid for this longitudinal effort.

Method

Sample selection

Recruitment of schools for the randomized study began in November, 2000. From the outset, there were problems in providing sufficient incentives to induce school leaders to allow their schools to be assigned at random to experimental or control conditions. Initial efforts focused on reducing the cost to schools of implementing Success for All, which would ordinarily require schools to expend about $75,000 in the first year, $35,000 in the second year, and $25,000 in the third year. At first, schools willing to be assigned at random to Success for All or control conditions were offered a $20,000 discount on first-year program costs. By spring 2001, this incentive had been increased to offer schools in either assigned condition a one-time payment of $30,000 in exchange for participation in the study. Control schools could use the incentive however they wished, and were allowed to implement any innovation other than Success for All. However, this incentive did not attract an adequate number of schools. Six schools were recruited in this way by summer, 2001, and were randomly assigned to experimental and control conditions. This number was far from sufficient.

By late spring, 2002, a satisfactory (though expensive) incentive structure was in place. Schools willing to participate were assigned at random to use Success for All either in grades K-2 or in grades 3-5, at no cost. In this way, all schools would receive at least part of the program, and they did not have to contribute significant amounts of money in a time of tightened budgets. This incentive was sufficient, and we were able to recruit a total of 41 schools, including the six schools from the previous year. 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.

All schools (and their districts) had to agree 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. Schools had to agree to allow data collection for three years, and to remain in the same treatment condition for all three years. Schools that agreed to these conditions were randomly assigned by members of the Oversight Committee to experimental or control conditions.

Final sample. The final sample of 41 schools recruited for the study is shown in Table 2, along with information on demographic variables. The six pilot schools recruited in 2001-02 are implementing either the entire Success for All program or are entirely controls. That is, a pilot Success for All school is analyzed as both a K-2 Success for All school and, in future reports, as a 3-5 Success for All school. A pilot control school is analyzed as such in both categories. For this reason, the controls for the K-2 study include 17 schools implementing Success for All in 3-5 and the three pilot control schools. All children in kindergarten and first grade in fall, 2002, are considered the main longitudinal sample. Students who were in K-1 in fall, 2001 in the six pilot schools are also included.

Insert Table 2 about here

As is clear from Table 2, the treatment and control samples are reasonably well matched on baseline demographics. The sample is concentrated in the urban Midwest

(Chicago, St. Louis, and Indianapolis) and the rural and small town South, although there are several exceptions. Overall, the students in the sample are very disadvantaged, with just a few rural exceptions. About 76% of the students qualify for free lunch, 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, 60.4% of the sample is African American, compared to about 40% of Success for All students, and 10.1% of the sample is Hispanic, compared to about 35% of Success for All schools. The percent of white students, 27%, is similar to the Success for All percent white of about 25%.

The results shown in Table 3 provide direct comparisons of the baseline characteristics of the K-2 treatment schools and control schools. As the results indicate, the percentages of females, minorities, special education students, free lunch participants, and English as second language students were statistically equivalent across treatment and control schools. Likewise, *t*-tests for school enrollment and the baseline PPVT outcomes for the treatment and control schools showed no differences.

Insert Table 3 about here

Therefore, the treatment and control samples were sufficiently well-matched at baseline on key demographic characteristics and the PPVT pretest measure. Though the school sample has a higher percentage of African American students and a smaller proportion of Latino students than other Success for All schools, it is similar with respect to the percentage of free lunch participants. The sample is also comprised of schools

from diverse locales, including high-poverty urban and rural schools across 11 states. In these respects, the sample selection process and randomization procedure appear to have produced a baseline sample of schools that has good internal validity—because there are no large, statistically significant treatment/control differences—and good external validity—because the sample has demographic characteristics that resemble those for the population of Success for All schools and it includes schools from a range of regional contexts representing the national reach of the program.

Treatment Fidelity

As of January 2004, all grades K-2 classes in schools were implementing their assigned treatments. Implementation quality varied widely. Several schools took almost one year to understand and implement the program at a mechanical level and others embraced the program immediately and are doing 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.

Measures

Students in grades K-1 were pretested on the Peabody Picture Vocabulary Test and then individually posttested on the Woodcock-Johnson by testers hired, trained, and supervised by the National Opinion Research Center (NORC) at the University of Chicago. The six pilot schools were pretested in fall 2001 and posttested during spring 2002 and the 35 schools from the main sample were pretested in fall 2002 and posttested in spring 2003. The pilot and main samples were combined for the analyses.

Pretests. All children were individually assessed in fall, 2001 (pilot sample) or fall, 2002 (main sample) by testers hired, trained, and supervised by the NORC testers, 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.

Posttest Reading Tests. During the spring of 2002 (pilot sample) and spring of 2003 (main sample)—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 WMRT 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 enter Success for All or control schools after fall, 2002 are to be posttested each year and included in analyses of cohort means. Children who are English language learners but are taught in English will be posttested in English each year.

Results

In the previous description of the sample, we concluded that the analysis of the baseline data showed few 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 analyses of the first-year 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 sample sizes were composed of 2,260 students in 21grade K-2 Success for All treatment schools and 2,254 students in 20 control schools. Listwise deletion of cases with missing values did not cause differential attrition rates by program condition, χ^2 (1, N = 5924) = 1.91, p = 0.17, leaving 75% of the baseline sample of 3,002 treatment students and 74% of the 2,922 baseline controls for the preliminary analyses. The data and sample attrition occurred for two reasons. Approximately 15% of the students who were dropped from the current analyses had moved out of the school before the posttests were administered and, thus, had no outcome data. About 10% of the students who were excluded from the analyses remained in the treatment and control schools, but had missing data on one or both of the achievement measures.

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 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 (0.49), p = 0.63 (two-tailed), suggesting that the initial academic ability of the treatment and the control group students who were dropped from our analyses were 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 (-4.27), $p < .001$ (two-tailed). Because there was no statistically significant pretest difference between

those students who dropped from the Success for All treatment and control samples, though, this has no impact on the study's internal validity. Low-achieving students from the sample schools were underrepresented due to differential attrition rates by pretest and this does compromise the external validity of the study to some extent. In addition, 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), this underrepresentation of low achievers is most likely to result in downward biases of the treatment effect estimates.[2]

Hierarchical Linear Model Analyses of First-Year Treatment Effects

Cluster randomized trials (CRTs) in education generally randomize at the level of the school or classroom and collect data at the level of the student. In many respects, they are the optimal design for school-based and classroom-based interventions. They address practical problems, including the potential difficulties of randomizing individual teachers within schools or students within classrooms to alternate treatments, and they are often well aligned with the theory of how educational interventions work best: as coordinated, systemic initiatives delivered by organizational-level elements acting in concert. Though greater attention has been paid to these designs in education in recent years (Boruch, May, Turner, & Lavenberg, in press), there are still mixed opinions among methodologists regarding proper analytical strategies (Campbell, Mollison, Steen, Grimshaw, & Eccles, 2000; Peters, Richards, Bankhead, Ades, & Sterne, 2003).

[2] We performed similar analyses for the kindergarten and the first grade cohort separately. The results were similar to those for the combined sample with one exception. The attrition rate for the Success for All first grade cohort was somewhat higher than that for the first grade cohort from the control group, but this difference did not achieve a conventional level of statistical significance, χ^2 (1, N = 2923) = 3.427, p = .07.

Analysis of treatment effects at the level of the cluster that was randomized is the most appropriate analytical strategy. When the number of clusters is very small, though, this analytical strategy will not be efficient and will lack necessary statistical power. If clustering is simply ignored and the analysis is done at the level of the individual student, statistical power will be substantially increased. However, these standard tests of statistical significance, which assume that the outcome for an individual is completely unrelated (or independent) to that for any other student, are inappropriate for CRTs. This is the case because in CRTs two students randomized together within any one classroom or school are more likely to respond in a similar manner than two students randomized from different clusters. Data analysis needs to recognize the clustered nature of the data either by analyzing at the level of randomization—the school in this case—or by adjustment for clustering in student-level analyses.

A relatively recently proposed analytical strategy for the analysis of CRTs is the use of a hierarchical linear model (Raudenbush, 1997). In this case, one may simultaneously model both student and school-level sources of variability in the outcome. Providing that there is sufficient power at the highest level of aggregation, the school level, the choice between the student and the school as the proper level of analysis is moot. Specifically, we developed 2 level hierarchical models that nested students and their pretest and posttest scores within schools. The fully specified level 1, or within-school model nested students within schools with two covariates, the PPVT pretest and a dummy code indicating the student's baseline grade level (0 = kindergarten and 1 = 1st grade). The linear model for this level of the analysis is written as

$$Y_{ij} = ß_{0j} + ß_{1j}(\text{PPVT})_{ij} + ß_{2j}(\text{GRADE})_{ij} + r_{ij},$$

which represents the spring posttest achievement for student *i* in school *j* regressed on the PPVT pretest and grade level. The term r_{ij} is the level-1 residual variance that remains unexplained after accounting for the pretest and grade level for the students.

In this model, each student's PPVT score and grade level is centered around its within-school group mean and the effect of the pretest is modeled as a source of random variation at level 2 of the model. In this way, at level 2, we may estimate school-level Success for All treatment effects on two outcomes: the mean posttest achievement outcome in school *j* and the PPVT-posttest slope in school *j*. We simply allow the within-school grade-level gap—the difference between the posttest scores of kindergarten and first grade students in school *j*—to vary at random across schools and formulate no school-level model to attempt to model this source of between-school variability.

This level 2 model, thus, allows us to estimate both the overall and compensatory effects of Success for All. That is, the model answers the questions: does Success for All affect the school-level posttest mean and does it attenuate the school-specific relationship between pretest and posttest? Including the school-level aggregate PPVT pretest score to control for potential school-level compositional effects of the pretest score in school *j*, the fully specified level 2 model is written as

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

$$ß_{1j} = \gamma_{10} + \gamma_{11}(\text{MEANPPVT})_j + \gamma_{12}(\text{SFA})_j + u_{0j,}$$

$$ß_{2j} = \gamma_{20} + u_{0j}$$

where both the mean posttest intercept for school *j*, $ß_{0j}$, and the PPVT-posttest slope, $ß_{1j}$, are regressed on the school-level mean PPVT score, the SFA treatment indicator, plus a residual, u_{0j}. The within-school posttest difference between kindergarten and first grade

students, $ß_{2j}$, is specified as randomly varying, predicted only by an intercept and school-specific residual.

For each of the four achievement outcomes, we specified a series of four multilevel models. The preliminary unconditional model partitions the variation in the outcome among students and schools, and is used as a basis for comparing the fit of subsequent models, which introduce student and school-level predictors of the posttest. Model 1 includes the student-level pretest covariate as a predictor of the posttest. This model, which is unconditional at level 2, serves as a referent for the subsequent models when calculating the percentage of variance explained for both the level 2 school mean achievement intercepts and for the level 2 PPVT pretest slopes. Model 2 introduces the school-level average PPVT pretest score as a predictor of both the achievement intercepts and the within-school PPVT pretest slopes. Finally, Model 3 adds the Success for All treatment indicator as a predictor of the achievement intercepts and pretest slopes.

Letter Identification Outcome. The first series of multilevel models, shown in Table 4, assesses student and school-level effects on the Woodcock-Johnson Letter Identification posttest. The unconditional model with no student- or school-level predictors reveals the overall average value on the outcome measure, partitions the variance in the outcome into its between and within school components and tests whether there is a statistically significant amount of between-school variance to model with independent variables. For the Letter Identification outcome, the unconditional model yielded an average spring score of 437.58. The intraclass correlation coefficient was 0.10, which indicated that 10% of the variance in the Letter Identification posttest was between

schools and that there was a statistically significant, χ^2 (40, N = 41) = 493.70, amount of level 2 variability potentially explainable by school-level characteristics.

In Model 1 within Table 4, we included the student-level PPVT pretest and the grade-level indicator as predictors, thus explaining 30% of the within-school variability on the posttest. The coefficient for the pretest suggested that each 1 unit increase on the PPVT was associated with a 0.27 point increase on the posttest. The average within-school difference between kindergarten and first grade students on the posttest was 13.54 points. All three school-level effects, school mean achievement, the PPVT pretest slopes, and the K-1 achievement difference, showed statistically significant random variation across schools. In the ensuing models, we attempted to explain this random variation across schools in their PPVT pretest slopes and their mean achievement using the aggregate school-level PPVT score and, most importantly, the Success for All treatment indicator.

Model 2 in Table 4 adds the school mean PPVT as a predictor of the mean achievement and the PPVT pretest slope. In both cases, the measure was a statistically significant predictor. The effect for school mean achievement suggests that schools with higher mean PPVT pretest scores also tend to have higher Letter Identification posttest scores. Because we have used group-mean centering of the student-level PPVT covariate, the school-level compositional effect of the school mean PPVT is calculated by subtracting the estimated intercept for the within-school PPVT pretest slope (0.27) from the level 2 coefficient for the mean PPVT pretest (0.35). The difference, 0.08, suggests that the compositional effect of the school mean PPVT pretest is nonzero but not substantial. The statistically significant and negative coefficient for the mean PPVT

pretest as a predictor of the PPVT slope indicates that schools with higher mean PPVT scores have somewhat flatter PPVT-Letter Identification slopes. That is, in schools that performed better on the pretest, the positive relationship between students' pretest and posttest scores is attenuated. Taken together, these results suggest that schools achieving better performance on the PPVT pretest tend to also achieve greater excellence and equity on the Letter Identification posttest. By including the school mean PPVT pretest as a level 2 predictor, the model explained 55% of the variability in school mean achievement and 49% of the variability in the PPVT slope.

Finally, Model 3 introduces the Success for All treatment indicator as a predictor of school mean achievement and the PPVT pretest slope. This predictor shows no statistically significant first-year school-level effects of assignment to Success for All for kindergarten and first grade students. By dividing the Success for All coefficient for the mean achievement by the sample standard deviation for the Letter Identification posttest, we estimate a school-level effect size for assignment to treatment of $d = 0.02$.

Insert Table 4 about here

Word Identification Outcome. In Table 5, we model the student and school-level effects on the Woodcock-Johnson Word Identification posttest. The preliminary unconditional model indicates that 10% of the variance in the posttest was between schools and that there was a statistically significant, χ^2 (40, N = 41) = 200.49, amount of level 2 variability in school mean achievement that can be explained by school-level predictors.

In Model 1 within Table 5, we include the student-level PPVT pretest and the grade-level dummy code as predictors. These independent variables explain 53% of the within-school posttest variance. For the three school-level effects, school mean achievement, the PPVT slope, and the K-1 achievement difference, the model revealed statistically significant random variation across schools. In Model 2, we explained this school-level variation using the school-level average PPVT score as a predictor of both school mean achievement and the PPVT slope. The measure was a statistically significant predictor of school mean achievement but was not a statistically significant predictor of the PPVT pretest slope.

In Model 3 within Table 5, we introduced the Success for All treatment indicator as a predictor of school mean achievement and the PPVT slope. For the Word Identification outcome, there was no statistically significant first-year effect of school-level assignment to Success for All. Using the Success for All coefficient for the mean achievement outcome, 0.48, and dividing this value by the standard deviation for the Word Identification posttest, we estimate a school-level effect size for assignment to treatment of $d = 0.01$.

Insert Table 5 about here

Word Attack Outcome. In Table 6, we present the results from the hierarchical models predicting Woodcock-Johnson Word Attack outcomes. Again, we begin with the unconditional model, which indicates that the average posttest Word Attack score was 467.75. The intraclass correlation coefficient for this model was 0.12, indicating that

12% of the variance in the Word Attack posttest was between schools. The estimate for the school mean achievement random effect indicates a statistically significant, χ^2 (40, N = 41) = 711.81, amount of level 2 variability in school mean achievement.

Model 1 included the student-level PPVT pretest and the grade-level indicator as predictors. Together, these two student-level predictors explained 36% of the within-school posttest variance. Though the PPVT was an important predictor of within-school differences on the posttest, the relationship between PPVT and the posttest—the PPVT slope—did not randomly vary across schools. Because the PPVT slope was statistically equivalent across all schools, it was treated as fixed at level 2 of the model and no attempt was made to model the effect with school-level predictors. The two other school-level effects, school mean achievement and the K-1 achievement difference, did exhibit statistically significant random variation across schools.

In Model 2 we explained the school-level variation in school mean achievement using the school-level average PPVT score as a predictor. The aggregate PPVT pretest was a statistically significant predictor of school mean achievement. The positive relationship between the school mean PPVT and the posttest was 0.13 larger than the student-level effect of the pretest, suggesting a small but nonzero compositional effect of the pretest. The school mean PPVT pretest explained 37% of the between-school variability in school mean achievement.

Finally, Model 3 in Table 6 introduced the Success for All treatment indicator as a predictor of school mean achievement. This model for the Word Attack outcome revealed a statistically significant first-year effect of school-level assignment to Success for All. Again, by taking the Success for All coefficient of 4.89 for the mean achievement

outcome and dividing this value by the standard deviation for the Word Attack posttest, we estimated the magnitude of this school-level effect as nearly one quarter of one standard deviation, $d = 0.22$. By adding the Success for All treatment indicator, we explained 45% of the variability across schools in school mean achievement.

Insert Table 6 about here

Passage Comprehension Outcome. Results for the final outcome measure, Passage Comprehension, are shown in Table 7. The unconditional model indicated that the average spring Passage Comprehension score was 446.61. Similar to the other models, 10% of the variation in the outcome was between schools. This represented a statistically significant amount of between-school variability for the school mean achievement intercepts, χ^2 (40, $N = 41$) = 539.67.

In Model 1, we added the student-level PPVT pretest and the grade-level indicator as predictors. These two predictors accounted for 50% of the within-school variation on the posttest. There was considerable random variation across schools for both the PPVT slope and the K-1 achievement difference. Likewise, the school mean achievement intercepts exhibited statistically significant between-school random variation.

We explained the variation in school mean achievement using the school-level average PPVT score as a predictor in Model 2. The aggregate PPVT pretest was also modeled as a predictor of the PPVT pretest slope. It was a statistically significant predictor of school mean achievement but was not a statistically significant predictor of

the PPVT slope. The schools' average PPVT pretest scores explained 45% of the between-school variability in the school mean posttest outcome.

The final model, Model 3, in Table 7 included the Success for All treatment indicator as a predictor of school mean achievement and the PPVT pretest slope. This model for Passage Comprehension showed no statistically significant first-year effects of school-level assignment to Success for All for the two outcomes. The effect size calculation using the Success for All coefficient of -0.46 for the mean achievement outcome yielded the value d = -0.02.

Insert Table 7 about here

Discussion

In most respects, the results of the first-year outcomes of the Success for All national randomized trial are quite positive. First, the selection and randomization processes worked well. With considerable effort and expense, we were able to obtain the cooperation of a sufficient number of Success for All and control schools to provide an acceptable level of statistical power to detect school-level effects within a multilevel model framework. No matter how carefully drawn, a sample of 41 schools is not likely to represent the population of over 1,400 Success of All schools with great precision. However, the process does seem to have developed a sample of schools that is similar with respect to the overall poverty level of the population of Success for All schools and it is a geographically diverse sample that is spread across 11 states. Randomization produced control and K-2 treatment samples that were reasonably well matched on a

variety of baseline characteristics, including demographics and PPVT pretest scores. No statistically significant baseline school-level differences were detected.

Second, the data and sample attrition over the first year of the study had few impacts on the good external and internal validity that was achieved through the sample selection and randomization procedures. There was no differential rate of attrition from Success for All K-2 treatment schools and control schools and comparisons of those students who dropped out of the analysis revealed no treatment-control differences with respect to pretest scores. However, when comparing the pretest scores of those students who dropped from the analyses—across both Success for All and control schools—to those students who remained in the analytical sample, there are statistically significant differences. Specifically, those students who dropped from the analyses have lower PPVT pretest scores than those students who remained in the analytical sample. This compromises the external validity of the findings somewhat, because low-achieving students are underrepresented in the analysis. Given that previous quasi-experimental studies of Success for All have consistently shown stronger academic benefits for those students who performed poorly on baseline tests (Slavin & Madden, 2001), this differential attrition is most likely to depress estimates of the treatment effects.

Third, the treatment fidelity and Success for All implementation quality seem reasonably good. In some schools, the tight deadlines involved in the selection and randomization process appear to have provided insufficient time for the program to get established and flourish. These qualitative differences in implementation quality will be an important subject of future work.

Finally, the pattern of first-year treatment effects we found appears to be consistent with previous quasi-experimental work on Success for All, the Success for All program theory, and more general research and theory on the development of young children's emergent literacy skills. We found effects of both statistical and practical significance on the Word Attack posttest, but did not find such effects on Letter and Word Identification and Passage Comprehension. Though prior quasi-experimental studies of Success for All have shown effects of greater magnitudes, the same pattern has existed. That is, the strongest first-year effects of the program tend to be in the domain tested by Word Attack. In later years, the phonetic and structural analysis skills tapped by the Word Attack test help Success for All children develop and further their skills in more advanced comprehension and broad reading skills, especially those measured by the Passage Comprehension tests. The treatment effects in these other skill areas typically have become more pronounced after children's exposure to two or more years of the Success for All program.

This pattern of effects fits the theory of Success for All, which focuses on the development of "Reading Roots" in the early grades. A strong focus of the Reading Roots component of Success for All is to increase children's ability to hear sounds within words (phonemic awareness) and to use phonetic strategies to decode text. Though developing a love of reading and nurturing a child's literacy development within the context of meaningful literature are key components of developing these early skills, story-related activities and direct instruction in reading comprehension are more clearly stressed in the later grades within the Success for All "Reading Wings" component. This program theory for developing children's literacy skills is consistent with more general

theories of how young children develop as emergent readers (Snow, Burns, & Griffin, 1998). Specifically, powerful decoding strategies and phonemic awareness, as stressed by the kindergarten and first grade Success for All program, are key building blocks upon which children can develop a broader range of skills.

Implications and the Future of the Success for All National Randomized Field Trial

In future research, we will examine how well the program theory matches the outcomes that we observe for the various reading assessments. In addition, we will study the extent to which the schoolwide Success for All program impacts other important and more general school outcomes, including special education referrals, attendance, and retention rates. This first-year investigation establishes a strong foundation for this future work.

The study also responds to the many doubts that have been raised about the viability and appropriateness of randomized experiments in school settings (Cook & Payne, 2002). As a randomized field trial, rather than a relatively artificial laboratory experiment, the results of this study have strong external validity and relevance for policy and practice. Further, combined with future survey and qualitative data regarding implementation and process, this research will open the experimental "black box" and generate additional information that will be helpful to policymakers, practitioners, and scholars alike. Of course, we will continue to generate estimates of treatment effects in subsequent analyses, but additional information will help inform questions regarding how the effects were, or were not, attained. This kind of data is the type that can, ultimately, improve programs and practices. Through this information and careful description of the

study design, we also hope to continue to provide insights that will help researchers design and implement future randomized field trials on educational interventions.

As we have noted in this article, the process of randomization was not easy or inexpensive, but a prominent goal of this study included recruiting a group of schools that was agreeable to the idea of randomization. In this way, experimentation was viewed as a partnership with school personnel rather than as a process imposed upon practitioners. Schools choose to implement Success for All. This is the population of schools to which we wish to generalize: those schools who would like to adopt and implement Success for All. To foist reform on schools unwilling to implement it would not be consistent with the Success for All model or with how reform generally occurs in education. Such a study would have limited generalizability. Rather, this project ties together two central themes of educational research and policy today: the scale up, or replication, of school-based interventions and the development of high-quality evidence of their causal effects. These first-year outcomes have established that randomized field trials involving nationally replicated interventions are both possible and desirable for producing unbiased estimates of the effects of educational treatments.

References

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

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.

Boruch, R., de Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (eds.), *Evidence matters: Randomized trials in education research* (pp. 50-79). Washington, DC: Brookings.

Boruch, R., May, H., Turner, H., & Lavenberg, J (with Petrosino, A., de Moya, D., Grimshaw, J., & Foley, E.) (in press). Estimating the effects of interventions that are deployed in many places: Place-randomized trials. *American Behavioral Scientist*.

Campbell, M.K., Mollison, J., Steen, N., Grimshaw, J.M., & Eccles, M. (2000). Analysis of cluster randomized trials in primary care: A practical approach. *Family Practice, 17,* 192-196.

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

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.

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

Peters, T.J., Richards, S.H., Bankhead, C.R., Ades, A.E., & Sterne, J.A.C. (2003). Comparison of methods for analyzing cluster randomized trials: An example involving a factorial design. *International Journal of Epidemiology, 32,* 840-846.

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

Ramey, C.T., & Campbell, F.A. (1984). Preventive education for high-risk children: Cognitive consequences of the Carolina Abecedarian Project. *American Journal of Mental Deficiency, 88,* 515-523.

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

Schweinhart, L.J., Barnes, H.V., & Weikart, D.P. (1993). *Significant benefits: The High/Scope Perry Preschool Study through age 27.* Ypsilanti, MI: High/Scope Press.

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

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

Snow, C.E., Burns, M.S., & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children*. Washington, DC: Committee on the Prevention of Reading Difficulties in Young Children, National Research Council.

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

Author Note

This research was supported by grants from the Institute of Education Sciences (IES), U.S. Department of Education (No. R306 S000009) and No. R-11 7-40005). However, any opinions expressed are those of the authors and do not necessarily represent IES positions or policies.

Table 1

Major Elements of Success for All

Success for All is a schoolwide program for students in grades pre-kindergarten to six that organizes resources to attempt to ensure that every child will be successful in reading from the beginning of their time in school, that they will never begin the process of falling behind. The emphasis of the program is on prevention and early, intensive intervention designed to detect and resolve reading problems as early as possible, before they become serious. The main elements of the program are as follows.

A Schoolwide Curriculum. During reading periods, students are regrouped across age lines so that each reading class contains students all at one reading level. Use of tutors as reading teachers during reading time reduces the size of most reading classes to about 20. 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 activities built around partner reading, identification of characters, settings, problems, and problem solutions in narratives, story summarization, 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. Classroom libraries of trade books are provided for this purpose. 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.

Preschool and Kindergarten. The preschool and kindergarten programs in Success for All emphasize language development, readiness, and self-concept. Preschools and kindergartens use thematic units, language development activities, and a program called Story Telling and Retelling (STaR).

Quarterly Assessments. Students in grades 1-6 are assessed every quarter to determine whether they are making adequate progress in reading. This information is 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.

Family Support Team. A family support team works in each school to help support parents 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 vice principals.

Facilitator. A program facilitator works with teachers to help them implement the reading program, manages the eight-week assessments, assists the family support 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.

Training. Success for All provides very extensive training to help all teachers use the program effectively. New schools begin with a week-long training for the principal and facilitator, followed by a three-day workshop before school opening for all staff. Implementation visits and additional training sessions are then provided throughout the first year, and continue on a gradually diminishing basis through the second, third, and subsequent years. In addition, school facilitators provide training and followup daily to all staff.

Table 2

Schools Participating in Randomized Evaluation of Success for All, Grouped by Assignment

School	District	ST	Enrollment	% White	% African American	% Hispanic	% Female	% ESL	% Spec Ed	% Free Lunch
Northwood*	Moorseville	IN	424	94.70	0.00	0.00	49.50	0.00	4.00	26.00
Jefferson*	Midland	OH	315	98.50	2.00	0.00	49.00	0.00	16.00	29.00
Bertha S. Sternberger	Guilford	NC	381	62.00	30.00	5.00	50.00	4.00	20.00	30.00
Pleasant Garden	Guilford	NC	698	76.00	13.70	4.56	49.50	3.50	16.00	33.00
Waveland	S Montgomery	IN	152	99.00	0.00	1.00	46.00	0.00	15.00	34.00
James Y. Joyner	Guilford	NC	454	41.80	42.90	5.00	48.50	0.05	38.00	35.00
Laurel Valley	Ligonier Valley	PA	409	99.98	0.01	0.01	47.00	0.00	9.00	51.00
Wood	Tempe	AZ	642	19.60	9.60	40.10	49.70	37.50	11.00	51.70
Cesar Chavez	Norwalk	CA	589	6.00	3.00	88.00	47.00	39.00	5.00	71.00
Haven*	Savannah	GA	373	0.00	99.00	0.00	65.00	0.00	5.00	85.00
Brian Piccolo	Chicago	IL	1069	0.10	79.50	20.10	48.00	11.40	11.60	85.60
Robert H. Lawrence	Chicago	IL	635	0.00	98.60	0.02	60.00	0.00	11.30	89.00
Harriett B. Stowe	Indianapolis	IN	174	28.71	25.74	43.56	38.00	40.59	17.80	90.59
Linden	Linden	AL	291	1.00	97.00	1.00	49.00	0.00	15.90	91.00
Lafayette	St. Louis	MO	297	13.20	72.80	9.40	49.70	25.00	12.00	94.00
Benjamin E. Mays	Chicago	IL	263	0.00	95.00	5.00	49.00	0.00	10.00	95.00
Paramount Jr.	Greene	AL	493	0.00	100.00	0.00	41.00	0.00	11.00	97.00
Farragut	St. Louis	MO	350	0.00	100.00	0.00	44.00	0.00	4.00	98.00
Gundlach	St. Louis	MO	365	0.00	100.00	0.00	48.00	0.00	4.40	99.00
Cook	St. Louis	MO	335	0.00	100.00	0.00	45.00	0.00	15.00	100.00
Earl Nash	Noxubee	MS	509	1.00	98.00	1.00	51.00	0.20	6.50	100.00
Treatment school means			439	30.60	55.60	10.60	48.76	7.68	12.31	70.70
Newby*	Mooresville	IN	252	98.00	0.00	0.00	53.00	0.00	10.00	35.00
Jamestown	Guilford	NC	516	41.00	47.00	3.40	45.00	8.00	15.00	39.00
Central	Central	KS	194	87.00	1.00	5.00	46.00	0.00	10.00	40.00
Walnut Cove*	Walnut Cove	NC	355	79.00	17.00	1.00	51.00	0.01	26.00	41.00
Bluford	Guilford	NC	420	9.50	86.70	1.20	48.00	0.00	13.00	42.00
Greenwood	Bessemer	AL	375	13.00	75.00	10.00	52.00	11.00	13.00	76.00
Gulfview	Hancock	MS	590	95.00	3.00	1.00	49.00	0.00	6.70	80.00
Eutaw	Greene	AL	338	0.57	98.50	0.00	49.00	0.00	4.00	86.00
C. F. Hard	Bessemer	AL	619	0.20	99.70	0.00	46.00	0.00	6.00	90.00
Daniel Webster	Chicago	IL	671	0.00	100.00	0.00	44.00	0.00	5.00	95.00
Augustin Lara	Chicago	IL	616	0.50	0.40	98.60	50.00	54.00	8.70	95.70
Edward E. Dunne	Chicago	IL	531	0.00	100.00	0.00	48.00	0.00	3.80	96.00
Bunche	Chicago	IL	643	0.00	100.00	0.00	68.00	0.00	9.70	98.00
Cupples	St. Louis	MO	171	0.00	100.00	0.00	56.00	0.00	2.00	98.00
Dewey Elem.	Chicago	IL	616	0.00	98.00	0.00	65.00	25.00	6.30	98.00
Scullin	St. Louis	MO	282	0.00	100.00	0.00	42.00	0.00	13.00	98.00

(Table 2 continued)

M. E. Lewis*	Sparta	GA	631	1.16	97.67	0.00	50.60	0.16	17.30	99.00
Sigel Elem.	St. Louis	MO	340	8.10	85.30	1.80	45.00	18.00	16.00	100.00
South Delta	South Delta	MS	769	5.00	94.00	1.00	53.00	0.00	4.00	100.00
Stanfield	Stanfield	AZ	757	19.40	1.00	67.70	50.30	50.00	19.00	100.00
Control school means			484	22.90	65.20	9.50	50.55	8.31	10.43	80.34

Note: * Denotes the schools selected in the pilot phase of the study.

Table 3

Comparison of Baseline Characteristics for Success for All K-2 Treatment Schools and Control Schools

Variable	Condition	*N*	*M*	*SD*	*t*
PPVT	Control	20	89.25	8.40	0.42
	Treatment	21	90.38	8.78	
Enrollment	Control	20	484.00	186	0.46
	Treatment	21	439.00	206	
% Female	Control	20	50.54	6.47	0.35
	Treatment	21	48.76	5.60	
% Minority	Control	20	77.10	35.78	0.52
	Treatment	21	69.63	39.35	
% ESL	Control	20	8.31	16.49	0.90
	Treatment	21	7.68	14.35	
% Special Education	Control	20	10.43	6.14	0.39
	Treatment	21	12.31	7.64	
% Free Lunch	Control	20	80.34	25.14	0.26
	Treatment	21	70.71	29.02	

Table 4

Multilevel Models Predicting Letter Identification Outcome

	Unconditional Model			Model 1			Model 2			Model 3		
Fixed Effect	Effect	*SE*	*T*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	437.58**	0.76	575.64	437.58**	0.76	575.78	437.58**	0.53	831.63	437.58**	0.53	827.62
Mean PPVT pretest							0.35**	0.06	5.40	0.35**	0.06	5.42
SFA treatment										0.30	0.90	0.34
PPVT pretest slope												
Intercept				0.27**	0.02	13.48	0.27**	0.02	15.89	0.27**	0.02	22.41
Mean PPVT pretest							-0.01**	0.00	-4.78	-0.01**	0.00	-4.38
SFA treatment										0.00	0.04	0.14
Grade												
Intercept				13.54**	0.61	22.10	13.55**	0.60	22.41	13.55**	0.60	22.41
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	21.93	493.70	40	22.64	707.70	40	10.27	341.53	39	10.63	344.88	38
PPVT pretest slope				0.01	85.00	40	0.00	55.33	39	0.00	55.02	38
Grade				9.82	106.30	40	9.48	106.35	40	9.50	106.34	40
Within-school variation	200.69			140.02			140.02			140.03		
Variance Explained												
Within-school variation				30%			30%			30%		
School mean achievement							55%			53%		
PPVT pretest slope							49%			47%		

Note: ** $p < .001$.

Table 5

Multilevel Models Predicting Word Identification Outcome

	Unconditional Model			Model 1			Model 2			Model 3		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	406.35**	2.03	200.49	406.27**	2.01	201.31	406.33**	1.57	258.40	406.33**	1.57	258.71
Mean PPVT pretest							1.01**	0.16	6.40	1.01**	0.16	6.28
SFA treatment										0.48	3.05	0.16
PPVT pretest slope												
Intercept				0.73**	0.04	18.28	0.72**	0.04	18.30	0.73**	0.04	18.33
Mean PPVT pretest							0.00	0.00	0.44	0.00	0.00	0.50
SFA treatment										-0.02	0.08	-0.30
Grade												
Intercept				49.03**	1.68	29.13	49.14**	1.67	29.38	49.14**	1.67	29.42
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	156.89	572.43	40	163.65	1221.28	40	98.88	737.43	39	101.21	735.44	38
PPVT pretest slope				0.03	65.55	40	0.02	65.54	39	0.03	65.80	38
Grade				90.05	179.17	40	88.79	179.13	40	88.61	179.15	40
Within-school variation	1340.31			629.20			629.24			629.19		
Variance Explained												
Within-school variation				53%			53%			53%		
School mean achievement							40%			38%		
PPVT pretest slope							0%			0%		

Note: ** $p < .001$.

Table 6

Multilevel Models Predicting Word Attack Outcome

	Unconditional Model			Model 1			Model 2			Model 3		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	467.75**	1.24	377.35	467.73**	1.24	378.03	467.76**	0.98	475.35	467.76**	0.91	512.85
Mean PPVT pretest							0.56**	0.10	5.43	0.53**	0.10	5.32
SFA treatment										4.89*	1.85	2.64
PPVT pretest slope												
Intercept				0.43**	0.02	22.55	0.43**	0.02	22.56	0.43**	0.02	22.56
Grade												
Intercept				20.56**	1.21	16.97	20.56**	1.21	16.99	20.55**	1.21	16.96
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	59.61	711.81	40	61.11	1106.61	40	38.56	696.68	39	33.62	569.18	38
Grade				48.99	221.50	40	48.86	221.49	40	49.04	221.51	40
Within-school variation	420.26			270.44			270.43			270.43		
Variance Explained												
Within-school variation				36%			36%			36%		
School mean achievement							37%			45%		

Note: * $p < .05$; ** $p < .001$.

Table 7

Multilevel Models Predicting Passage Comprehension Outcome

	Unconditional Model			Model 1			Model 2			Model 3		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
School mean achievement												
Intercept	446.61**	1.22	366.01	446.58**	1.21	367.67	446.61**	0.91	492.19	446.61**	0.90	493.60
Mean PPVT pretest							0.72**	0.10	7.46	0.72**	0.10	7.38
SFA treatment										-0.46	1.66	-0.28
PPVT pretest slope												
Intercept				0.49**	0.02	21.84	0.50**	0.02	22.51	0.50**	0.02	23.36
Mean PPVT pretest							0.00	0.00	1.30	0.00	0.00	1.18
SFA treatment										0.04	0.04	0.86
Grade												
Intercept				28.49**	1.13	25.23	28.59**	1.12	25.47	28.58**	1.12	25.42
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	56.57	539.67	40	58.96	1079.71	40	32.30	559.92	39	32.89	556.71	38
PPVT pretest slope				0.01	55.65	40	0.01	53.51	39	0.00	51.53	38
Grade				41.84	198.32	40	41.27	198.23	40	41.55	198.18	40
Within-school variation	510.38			255.28			255.34			255.40		
Variance Explained												
Within-school variation				50%			50%			50%		
School mean achievement							45%			44%		
PPVT pretest slope							4%			18%		

Note: ** $p < .001$.