Effects of Success for All on Reading Achievement:

A Secondary Analysis Using Data From the Study of Instructional Improvement (SII)

Alan Cheung

The Chinese University of Hong Kong

Robert E. Slavin

Johns Hopkins University

June, 2016

Abstract

This study examined the effects of the Success for All (SFA) whole-school reform approach on student reading achievement. The data were collected for the Study of Instructional Improvement (SII) by the University of Michigan, which did not previously report the achievement outcomes in detail, but did make the data available online. Using propensity matching, we matched 27 SFA with 27 comparable schools based on several key demographic variables. The evaluation used hierarchical linear modeling (HLM) with students nested within schools. Results showed that SFA students significantly outperformed their counterparts in the matched schools on reading achievement, with an effect size of +0.26 for students in a 3-year longitudinal comparison. Effect sizes were similar for two-year cohorts (mean effect size=+0.31). Policy implications are discussed.

Key words: Success for All, Study of Instructional Improvement, secondary analysis, reading achievement

Introduction

In recent years, evidence of effectiveness from rigorous evaluations has begun to play an increasing role in educational policy and practice. Evidence-based reform has advanced in the policy arena from the ill-defined emphasis on using practices "based on scientifically-based research" enshrined in No Child Left Behind to much clearer definition of what counts as strong, moderate, or promising evidence of effectiveness that appears in the Every Student Succeeds Act (ESSA) adopted in 2015. Legislation regarding School Improvement Grants for struggling schools has established an eligible funding category in which low-performing schools may apply for funding to implement whole-school reform programs supported by moderate to strong evidence of effectiveness, defined as at least one large, rigorous experiment showing positive achievement impacts.

These policy developments are increasing the importance of large scale, especially third-party evaluations of educational programs. Rigorous, replicated experiments are gradually becoming the gold standard for impact on policy and practice.

One program central to the discussions of evidence-based reform is Success for All, a whole-school reform model focused on improving reading outcomes in high-poverty Title I elementary schools (Slavin, Madden, Chambers, & Haxby, 2009). Success for All provides reading materials, software, and extensive professional development to all teachers in Title I elementary schools. The professional development focuses on building teachers' skills in implementing cooperative learning and providing effective instruction in phonics, metacognitive strategies in reading comprehension and writing, classroom management, and other approaches. It also provides professional development to tutors to work with struggling readers, as well as family support and leadership approaches.

Success for All is composed of elements proven effective in research, and the program itself has been extensively evaluated. Evaluations include a randomized longitudinal study in 35 schools by Borman, Slavin, Cheung, Chamberlain, Madden, & Chambers (2007) and another third party randomized study in 37 schools by MDRC (Quint et al., 2015). These, and numerous first-party and third-party quasi-experiments, have clearly established the effectiveness of Success for All in increasing reading achievement in the early grades (see Slavin, Lake, Chambers, Cheung, & Davis, 2009). However, the largest study ever done to evaluate Success for All has never been fully reported. This was a study carried out by Brian Rowan, Richard Correnti, and their colleagues at the University of Michigan, from 1999 to 2004. The study, called the Study of Instructional Improvement (SII), compared Success for All and two other whole-school reform models, America's Choice and Accelerated Schools, to each other and to a control group. The full study involved 115 high-poverty schools across the U.S.

Rowan, Correnti, Miller, & Camburn (2009) and Correnti (2009) reported the procedures and findings of the SII study in terms of teachers' behaviors, derived from teacher logs (Rowan & Miller, 2007; Correnti & Rowan, 2007). However, they never reported achievement outcomes in any detail, only estimating that Success for All students gained more than students in the other three programs on TerraNova reading tests, with an effect size of about +0.40 at the end of three years, moving the average child from the 30th to the 50th percentile. Even these findings were only presented in a technical report and an AERA paper, and were never published. Insufficient detail was provided to enable the achievement findings to be confirmed according to today's standards of rigor for experimental evaluations. The studies examining teacher logs noted that teacher behavior in Success for All and America's Choice (but not Accelerated Schools) changed in directions consistent with these models' emphases (see below), but reading outcome data have not been adequately reported.

The Study of School Improvement, though completed more than a decade ago, has particular relevance to policy issues today. Anyone who follows the What Works Clearinghouse or other summaries of research on educational interventions is aware that few programs are finding consistent positive impacts on achievement outcomes in rigorous, large-scale evaluations. Many of the programs that have shown positive effects are one-to-one or one-to-small group tutoring models, not whole-school or even whole-class interventions with potential for broader impact. If evidence of effectiveness is to become increasingly important in federal, state, and local policy, it is essential to have a broad array of proven programs meeting the highest standards of rigor in their evaluations (see Slavin, 2013; Cohen & Moffitt, 2009).

The Study of School Improvement happened to have evaluated two programs that did make substantial differences in achievement and one that did not. Further, it collected and analyzed detailed teacher logs that made it possible to quantify what teachers were doing differently in the different whole-school designs. It also commissioned studies of the organizations that created each of the programs, making possible comparisons at that level (Cohen et al., 2014; Peurach, 2011).

The only one of the programs evaluated by SII that is still in widespread use today is Success for All. Success for All has had positive reading achievement outcomes in the great majority of its evaluations, averaging an effect size of +0.31 (Slavin, Madden, Chambers, & Haxby, 2009). For example, in Social Programs That Work (http://evidencebasedprograms.org), Success for All in grades K-2 is one of just two whole school or whole class programs to meet "top tier" standards (the other is career academies for high schools). The What Works Clearinghouse (2009) accepted 7 studies evaluating Success for All, one "without reservations" and six "with reservations." The weighted mean effect sizes across the seven studies were +0.25 for letter-word identification, +0.39 for word attack, +0.20 for comprehension, and +0.27 for general reading.

One reason Success for All may have had relatively consistent positive effects for reading in the primary grades is that the program provides a clear structure for teachers and extensive professional development and on-site coaching for all school staff, which lead to significant changes in daily instruction, aligned with the program's theory of action (Peurach & Glazer, 2012). Summaries of teacher log data reported by the SII research make this point repeatedly for SFA with respect to early elementary reading (and with respect to writing for America's Choice) (Rowan & Miller, 2007). These are the areas in which each program produced significantly greater gain than the other programs and the control group. For example, Correnti & Rowan (2007) reported that Success for All teachers taught reading comprehension in 65% of lessons, in comparison to 50% in comparison schools. Success for All teachers reported spending significantly more time than controls on reading comprehension and word analysis, the very areas in which program impacts were strongest. During reading comprehension instruction, SFA teachers were significantly more likely to have students discuss text with each other in cooperative groups, to focus on literal comprehension, and to check students' comprehension by eliciting brief answers from students. They were no less likely than control teachers to focus on advanced reading strategies or to have students write extended text about what they read, but they provided much more direct instruction on comprehension strategies. Further, teachers in Success for All schools showed much less variability in instruction than did teachers in other schools.

Focusing at the school level, Rowan & Miller (2007) reported that teachers in Success for All schools (and in America's Choice schools) felt they had greater levels of instructional guidance than did teachers in Accelerated Schools or control schools. They also reported that their improvement efforts were more closely monitored.

Teachers in Accelerated Schools reported higher autonomy, values-based decision making, and strength of professional community. However, these same teachers' logs did not reveal any actual

change in daily teaching behaviors, in comparison to controls. Based on their logs, teachers' behaviors in Accelerated Schools were indistinguishable from those of teachers in control schools. As a likely consequence, the Accelerated Schools did not show any greater gains in student learning in comparison to controls.

The comparisons in emphases and outcomes among the three whole-school reform models are crucially important for understanding the situation of evidence-based reform today. The SII researchers concluded from their data that in order for whole-school reform models to have significant impacts on student learning, especially in often-chaotic and stressed high-poverty schools, they must have a clear plan for reform and implement it with sufficient specificity, professional development, and classroom supports to ensure that teachers' behaviors change in directions likely to improve student outcomes.

Today, Success for All represents the main surviving example of a whole-school reform with a strong emphasis on providing high-poverty schools with specific, well-structured strategies, supportive classroom materials and software, and extensive professional development and on-site coaching. The present study reaches back in time to explore data from the SII to better understand the impacts of Success for All and implications for evidence-based reform and educational policies of today. Our hope is that the lessons of Success for All derived both from the unreported achievement outcomes and from previously reported teacher logs and institutional analyses (Correnti & Rowan, 2007; Peurach, 2011; Rowan & Miller, 2007) will help current and future designers create additional whole-school approaches as effective or more effective than Success for All to serve the many disadvantaged children in need of higher quality instruction and better learning outcomes.

In the Rowan et al. (2009) report was a URL for a web site containing all of the data from the SII study. We obtained these data in order to carry out a summative evaluation of this major third-party

evaluation in an attempt both to confirm the reported findings and to add depth to them, examining different program durations available from the data files.

The present analysis used a propensity matching strategy in which SFA schools were matched based on multiple demographic variables with schools not using SFA across the entire SII sample, regardless of which non-SFA approach was in use. Hierarchical linear modeling (HLM) was used to deal with the clustered nature of the data. The hypothesis of this secondary analysis was that after controlling for pretests, SFA students would score significantly higher than comparison students on Terra Nova reading tests, and that this difference would be largest among students who had been in the program for two to three years.

Method

Data

As noted earlier, the data used in the current study were collected for the Study of Instructional Improvement (SII), conducted by the University of Michigan in collaboration with the Consortium for Policy Research in Education. SII was a large-scale quasi-experimental study that examined the effectiveness of three widely-disseminated comprehensive school reform (CSR) programs on instruction and student achievement in high-poverty elementary schools between 1999 and 2004. As indicated in Table 1, the SII sample consisted of 115 elementary schools, including 30 Success for All (SFA), 31 America's Choice (AC), and 28 Accelerated Schools Plus (ASP). In addition to following schools that adopted these three CSR programs, the study also followed 26 control schools (Correnti, 2009). The data were made publicly available and were downloaded through the Inter-University Consortium for Policy and Social Research (ICPSR) website (http://www.icpsr.umich.edu/).

==========

TABLE 1 HERE

==========

Demographic characteristics of participating schools

Background information on participating schools in the SII study is summarized in Table 2. It is clear that the majority of the participating schools served very disadvantaged minority students from high poverty communities. Forty percent of students were from single-parent homes and 70% of them qualified for free lunch. Approximately 70% of students were ethnic minorities, mostly African Americans. In terms of academic achievement, only 30% of students met state proficiency standards in reading and mathematics.

There were some key differences among the three sets of CSR program schools and the comparison schools in terms of school characteristics. For instance, SFA and AC schools served more disadvantaged students (75%) and had a higher percentage of minority students (75%) than ASP (63%) and the comparison schools (65%). Students in SFA and ASP schools had lower reading scores at pretest (in kindergarten) than those in AC and comparison schools on the Woodcock-Johnson assessments. AC (30%), SFA (30%), and ASP (31%) schools had lower percentages of students meeting state proficiency standards in reading at pretest than comparison schools (36%). As indicated in Table 4, before matching the SFA schools and the other schools showed considerable imbalance on various covariates.

==========

TABLE 2 HERE

==========

Propensity Score Matching Method

Given the fact that participating schools in the SII study were not randomly assigned and key differences existed between the SFA schools and other schools in the sample, we decided to use

propensity matching to locate better matched schools from the other two CSR programs and the comparison group. The problem with non-randomized designs is that the treatment group and the comparison group may systematically differ from each other based on school characteristics or covariates (Fan & Nowell, 2011; Rosenbaum & Rubin, 1983). Propensity score matching was employed to control for demographic or pretest differences by excluding participating schools that could not be well matched, so that systematic error could be reduced (Rosenbaum & Rubin, 1983). As Lane and colleagues (2012) argued, non-random sampling may introduce bias when comparing treatment effects between groups given an unequal and unknown probability of group assignment. Propensity score matching is an approach to tackle this problem by using regression techniques to generate predicted scores for each school regarding the likelihood of a school to be assigned to the treatment group given theoretically relevant covariates. A matching method is applied subsequently to the schools in treatment and comparison groups by those predicted scores (i.e. propensity score) so that schools of both groups would have an equal likelihood of receiving the treatment (Guo & Fraser, 2015).

A propensity score:

$$\pi_i = P(T_i = 1|X_i)$$

where π_i is the propensity score for school *i* which is the conditional probability (P) of assigning a school to treatment group (T = 1) give a set of covariates (X) of school *i*.

The four major steps for performing propensity score analysis in this study were as follows:

1. A list of eight relevant covariates was selected.
2. The probabilities or propensity scores were calculated for each school by using logistic regression (Thoemmes, 2012).

3. A one-to-one nearest-neighbor matching method with a caliper .25 standard deviations of the propensity score (Stuart, 2010) was adopted in this study. The aim was to pair each SFA school with a non-SFA school in the sample with the nearest propensity score.
4. An examination of the balance of covariates was conducted for the newly matched sample.

With the balance introduced by propensity score matching, we expected that there would be no systematic differences between the SFA schools and matched schools, and the treatment effect could be tested for the matched sample.

SPSS custom dialog "psmatching3.03", provided by Thoemmes (2012), with the R plug-in "Matchit" package (Ho, Imai, King, & Stuart, 2007) was used to perform propensity score matching. The program generated 27 SFA schools matched with 27 non-SFA schools, which included schools using various CSR programs in the SII sample (see Table 3). According to Rubin (2001), the absolute standardized difference in the mean of propensity score between two groups should be less than .20, and ratio of propensity score variances of both groups should be close to one for the matched sample. In our matched sample, the absolute standardized difference in the mean propensity score between two groups was reduced from a pre-matching of d=.89 to a post-matching of d = .05 (Table 4). The difference in propensity scores between the two post-matched groups were not statistically significant ($t[52]$=-.20, p=.841). The variance ratio dropped from 3.55 (pre-matching) to .99 in the post-matched sample. Thus, distribution of propensity scores of both group were similar.

A multivariate test generated from the SPSS R plug-in program was used to evaluate covariate balance. A balance test developed by Hansen and Bowers (2008) which was analogous to Hotelling's T^2 statistic tested overall covariate balance. In this study, the non-significant test result suggested a balance of covariates ($\chi^2[8]$=2.255, p=.972). Standardized mean differences for each covariate between the two groups were examined and no covariate exhibited a large imbalance (i.e. |d| >.25). Given the

results of multivariate and univariate tests, we assumed that covariate balance was established in the matched schools sample. Thus, tests of the treatment effect could proceed with the matched sample.

Final Sample

The final sample consisted of 54 matched schools: 27 SFA schools[1] and 27 matched control schools (16 AC schools, 5 comparison schools, and 6 ASP schools). Key demographic variables, such as a community disadvantage index, proportion of households with assistance income, proportion of households in poverty, free lunch, etc., were used as covariates in the propensity matching method.

There were two phases in the data collection of SII. The first cohort of students began the study during the 2000–2001 academic year; the rest began the study during the 2001–2002 academic year. Both cohorts were followed from kindergarten to second grade. There were also replacement students entering the study in the beginning of each academic year. For the purpose of our analysis, we treated the beginning year of the longitudinal study as Year 1 regardless of the phase of study, the following year as Year 2, and so on. For the longitudinal sample, i.e. those who entered the program at Time Point 1, 842 kindergarten students (SFA=411) took *TerraNova* tests in Spring Y1, 568 students (SFA=246) of the same cohort took tests in Spring Y2, and 453 (SFA=181) took tests in Spring Y3. Of those who entered the study at Time Point 2 (mostly 1st graders), 292 (SFA=149) took the tests in the spring of same academic year, and 191 (SFA=104) took the tests in the spring of the following year. There were 191 students, mostly 2nd graders, who entered the program at Time Point 3 and took the tests in the spring of the same academic year.

[1] On average, the three unmatched SFA schools tended to serve more disadvantaged student populations than their 27 matched SFA counterparts. For example, the proportion of students receiving free and reduced lunch was greater in three unmatched schools (86% vs 72%). In addition, these three schools had a higher percentage of minorities (94% vs 81%) and had lower Woodcock-Johnson pretest scores (-0.21 vs -0.16).

==========

TABLES 3 AND 4 HERE

==========

==========

FIGURE 1 HERE

==========

Measures

The primary outcome of interest was reading achievement. Participants were assessed on reading outcome measures in the fall as pretests and in each spring thereafter as posttests. Two measures were used for the current study: *The Woodcock-Johnson III (WJIII) Tests of Achievement* by Riverside Publishing and the *TerraNova Tests and Assessments* by McGraw Hill. The Letter/Word Identification scale of the WJIII subtest was used as a pretest for those students who started in kindergarten in the fall of Y1. The internal reliability coefficient for the Letter/Word identification subtest used was 0.92 (Woodcock, McGrew, & Mather, 2001). The Reading/Language Arts subtest of the *TerraNova Assessment* was used for all other grades as pretests and posttests. The internal reliability coefficient for the TerraNova Reading and Language Arts section was 0.87 (CTB/McGraw-Hill, 2001). For ease of interpretation, all scores were standardized.

Analyses

Due to the nested nature of the data, we employed a hierarchical linear model (HLM) with the school as the unit of analysis (Raudenbush & Bryk, 2002). The pretests were used as the covariate. The HLM approach was the optimal design for the current study because it addressed the practical problems of accounting for the effects of attending a given school, using degrees of freedom associated with the number of schools rather than students. HLM allows us to simultaneously model both student and

school-level sources of variability in the outcome (Raudenbush & Bryk, 2002). Specifically, a 2-level hierarchical model that nested students within schools was developed to analyze the data collected. The fully specified level 1, or within-school model, nested students within schools. The linear model for this level of the analysis is written as

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

This represents the post-test achievement for student *i* in school *j* regressed on the level-1 residual variance, r_{ij}. We also included a grade indicator as a predictor in the level 1 model. We treat the within-school grade-level gap—the difference between the post-test scores of different grades in school j—as fixed at level 2 because it is intended only as a covariate.

At level 2 of the model, we estimate SFA treatment effects on the mean post-test achievement outcome in school *j*. We included a school-level covariate, the school mean pre-test score, to help reduce the unexplained variance in the outcome and to improve the power and precision of our treatment effect estimates. The fully specified level 2 model is written as

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Mean Pre-test})_j + \gamma_{02}(\text{SFA})_j + u_{0j}$$

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

where the mean post-test intercept for school *j*, $ß_{0j}$ is regressed on the school-level mean pretest score, the SFA treatment indicator, plus a residual, u_{0j}. The within-school posttest difference between grades, $\beta_{1j,}$ is specified as fixed, predicted only by an intercept.

In the previous description of the sample, we concluded that the analysis of the baseline data showed few important differences between the SFA and the matched control schools.

Results

Pretest Differences

As indicated in Table 5, after matching, the SFA schools scored non-significantly higher than matched comparison schools at pretest.

==========

TABLE 5 HERE

==========

Outcomes for the 3-year longitudinal sample

The multilevel models, shown in Table 6, assessed student and school-level effects on their posttest scores. In Year 1, the treatment students scored non-significantly higher than the controls on the posttest with an effect size of +0.12 ($p<0.25$). In Years 2 and 3, students in the treatment condition significantly outperformed the controls with effect sizes of +0.34 ($p<0.01$) and +0.26 ($p<0.05$), respectively.

==========

TABLE 6 HERE

==========

Outcomes for 2-year longitudinal samples

We also examined the effects of experimental schools that had experienced one year and two years of treatment on posttest achievements. As indicated in Table 7, the average 1-year effect was +0.16 ($p<0.08$). The 2-year effect size was +0.40 ($p<.001$).

==========

TABLE 7 HERE

==========

One-year outcomes

The sample in the 1-year outcome included a total of 1,325 kindergarteners, 1st and 2nd grade students who had been in the study for only one year. The one-year impacts were summarized in Table 8. At posttest, the SFA schools scored marginally higher than the controls, with an effect size of +0.16 ($p<0.08$).

==========

TABLE 8 HERE

==========

Discussion

The purpose of this secondary analysis of data from the Study of Instructional Improvement was to provide well-supported achievement data for the study only partially reported by Correnti (2009) and Rowan et al. (2009). Effect sizes after the three-year longitudinal comparison in the present analysis were statistically significant in a rigorous HLM analysis. Correnti (2009) estimated an effect size of about +0.40 for the cohort that received three years of Success for All, while the current analysis found an effect size of +0.26, still an educationally meaningful effect.

Two-year impacts of Success for All were similar to three-year impacts, with an effect size of +0.36 for students in the second year of the three-year longitudinal sample, +0.16 ($p<.08$) for the K-1 sample, and +0.40 ($p<.001$) for the grade 1-2 sample. The mean effect size for students who received two years of treatment was +0.31. For students who received one year of treatment, effect sizes averaged +0.16 ($p<.08$) in the HLM analysis. The substantial increase in effect sizes from one year to two or three years

of Success for All experience matches findings from Borman et al. (2007) and Quint et al. (2015), although the 3-year Borman et al. (2007) study found that effect sizes continued to increase in the third year of SFA.

The observation that it takes two years or more to make comprehensive school reforms show their full effects has also been made by Fullan (2001), Borman et al. (2003), and others.

The findings provide further confirmation of the effectiveness of Success for All in improving reading, though (as in previous studies) the program had to be provided for at least two years to show its full effect (Borman et al., 2007). Because the data come from a very large national study carried out by third-party researchers, the SII study adds confidence that Success for All can be effective at substantial scale, an increasingly important issue in a policy climate of increasing focus on evidence of effectiveness in education.

Policy Implications

The findings of the SII study, both as originally reported and as largely confirmed in the secondary analysis, have broad implications for educational policy. From reports on large-scale, mostly randomized experiments evaluating educational interventions, it is becoming apparent that most innovations do not consistently improve students' achievement outcomes. Most of the exceptions involve one-to-one or one-to-small group teaching. For whole schools and major subjects, there are few clear examples of programs that are shown to be effective in comparison to control groups on measures that fairly assess what was taught in experimental and control groups.

Success for All is one of few examples of whole-school reforms that have had such positive impacts in two large cluster randomized experiments (Borman et al., 2007; Quint et al, 2015) and in the SII study reported here, as well as in many smaller experiments (Slavin et al., 2009).

At this point in time, it is important to ask why this particular program has been so consistent in tis impacts. One potential answer is provided by the original SII study, which obtained detailed teacher logs to characterize program implementation. The teacher log data reported by Correnti & Rowan (2007) and Rowan & Miller (2007), showed a clear impact of Success for All on teachers' reported behaviors, which were in line with the SFA theory of action and emphasis. Similarly, the logs reported in the evaluation of America's Choice also documented teaching in line with the programs' theory of action and emphasis. In both cases, outcomes mirrored the programs' emphasis, with the main outcomes of Success for All seen in early elementary reading while those of America's Choice were seen in upper elementary writing. Accelerated Schools, whose teachers did not report much change in behaviors, also did not find any effects on achievement. These findings suggest the possibility that many attractive-sounding interventions, such as Accelerated Schools, may be failing to show positive effects on achievement measures because they are not achieving major changes in teachers' behaviors.

Success for All places a substantial emphasis on extensive and explicit coaching to help teachers change their daily teaching behaviors (Slavin et al., 2009). It provides all first-year schools at least 26 person-days of on-site coaching, an in-school facilitator to work with all staff, a week of training for principals and facilitators, an annual conference, and constant electronic communications and sharing of data, ideas, and feedback. To bring about profound changes in teachers' daily behaviors, it might be argued that nothing less is likely to be effective.

The Success for All Foundation (SFAF), which developed and supports the program, has an institutional culture focused on leaving as little as possible to chance. In fact, as part of the SII project, Peurach (2011) studied SFAF over more than a decade, and documented this cultural focus, as well as ongoing attempts to learn from its own network of schools to incorporate best practices in its training, coaching, and materials (see also Cohen et al., 2014; Peurach, 2011).

It is entirely possible that many whole-school reform approaches starting from very different theoretical or philosophical bases could be effective in improving student achievement. However, the experience of Success for All, especially as revealed in the Study of Instructional Improvement, suggests that educational interventions are most likely to achieve their desired outcomes if they make certain to provide the professional development and schoolwide supports necessary to bring about meaningful changes in instructional practices throughout the school. If evidence-based reform is to transform America's schools, we need many whole-school approaches with strong evidence of effectiveness from rigorous evaluations. The Study of School Improvement and the broader experience of Success for All suggests that whatever these approaches may be, ensuring quality of implementation is essential.

If federal education policies are to make substantially greater impacts on student outcomes, they must sooner or later embrace policies promoting the use of federal resources to implement proven, replicable programs (Cohen & Moffitt, 2009). However, this shift is unlikely to take place if there are too few proven programs for schools to choose. The lessons of the SII study, emphasizing the need to ensure that programs do whatever it takes to see that teachers embrace and regularly utilize innovative strategies, support the idea that systematic, school-by-school reform can amount to important changes in outcomes.

References

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. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal. 44*(3), 701-731.

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

Cohen, D. K., & Moffitt, S. L. (2009). *The ordeal of equality: Did federal regulation fix the schools?* Cambridge, MA: Harvard University Press.

Cohen, D. K., Peurach, D. J., Glazer, J. L., Gates, K. G., & Goldin, S. (2014). *Improvement by design: The promise of better schools.* Chicago: University of Chicago Press.

Correnti, R. & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal, 44*(2), 298-338.

Correnti, R. (2009, March). *Examining CSR program effects on student achievement: Causal explanation through examination of implementation rates and student mobility*. Paper presented at the 2nd annual conference of the Society for Research on Educational Effectiveness, Crystal City, VA.

CTB/McGraw-Hill. (2001). *TerraNova technical report*. Monterey, CA: Author.

Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly*, 55(1), 74-79.

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

Guo, S. & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Hansen, B., & Bowers., J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*, 219-236. doi:10.1214/08-STS254

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199-236. doi:10.1093/pan/mpl013

Lane, F. C., To, Y. M., Henson, R. K., & Shelley, K. (2012). An illustrative example of propensity score matching within education research. *Career and Technical Education Research, 37*(3), 187-212. doi: 10.5328/cter37.3.187

Peurach, D. J. (2011). *Seeing complexity in public education. Problems, possibilities, and Success for All*. New York: Oxford University Press.

Peurach, D. J., & Glazer, J. L. (2012). Reconsidering replication: New perspectives on large-scale school improvement. *Journal of Educational Change, 13*, 155-190.

Quint, J, Balu, R., DeLaurentis, M., Rappaport, S., Smith, T., & Zhu, P. (2015). *The Success for All model of school reform: Findings from the Investing in Innovation (i3) scale-up*. New York: MDRC.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*, Second Edition. Thousand Oaks: Sage Publications, Inc.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. doi:10.2307/2335942

Rowan, B., & Miller, R. J. (2007). Organizational strategies for promoting instructional change: Implementation dynamics in schools working with comprehensive school reform providers. *American Educational Research Journal*, *44*(2), 252-297.

Rowan, B., & Miller, R. J. (2009). *SII Multi-Component Survey Data Files User's Guide*. Ann Arbor, MI: Consortium for Policy Research in Education (CPRE) in D. L. Ball, D. K. Cohen, & B. Rowan. *Study of Instructional Improvement (SII): Codebook*. ICPSR26282-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Rowan, B., Correnti, R., Miller, R.J., & Camburn, E.M. (2009). *School improvement by design: Lessons from a study of comprehensive school reform programs.* Ann Arbor, MI: Consortium for Policy Research in Education.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169-188. doi:10.1023/A:1020363010465

Slavin, R. E. (2013). Overcoming the four barriers to evidence-based education. *Education Week, 32* (29), 24.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, *79*(4), 1391-1465.

Slavin, R.E., Madden, N.A., Chambers, B. & Haxby, B. (Eds.) (2009). *Two million children: Success for All*. Thousand Oaks, CA: Corwin.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1-21. doi:10.1214/09-STS313

Thoemmes, F., (2012). *Propensity score matching in SPSS*. Available at http://arxiv.org/ftp/arxiv/papers/1201/1201.6385.pdf.

What Works Clearinghouse (2009). *Success for All intervention report.* Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_sfa_081109.pc

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock Johnson III Tests of Achievement.* Rolling Meadows, IL: Riverside.

Table 1. Participating schools in the SII study

Intervention Program	**Total**
SUCCESS FOR ALL (SFA)	30
AMERICA'S CHOICE (AC)	31
ACCELERATED SCHOOLS PROJECT (ASP)	28
COMPARISON	26
Total	115

Table 2. Demographic characteristics of original sample of SII schools by CSR program (Correnti, 2009)

	SFA N=30	AC N=31	ASP N=28	Comparison N=26
Ethnic minorities	71%	79%	65%	65%
Receiving free lunch	74%	75%	62%	64%
From single parent home	46%	49%	37%	38%
Proportion households in poverty in the community	23%	19%	14%	22%
Proportion individuals without a high school diploma in the community	35%	34%	32%	34%
Percent of students meeting state proficiency standards in reading	30%	30%	31%	36%

Table 3: Result of Propensity Score Matching

	No. of schools		
Intervention Program	Non-matched	Matched	Total
SUCCESS FOR ALL	3	27	30
AMERICA'S CHOICE	15	16	31
ACCELERATED SCHOOLS	22	6	28
PROJECT			
COMPARISON	21	5	26
TOTAL	61	54	115

Table 4: Covariates used in propensity score matching

	Before Matching					After Matching				
	Other Schools (N=85)		SFA (N=30)			Matched (N=27)		SFA (N=27)		
	Mean	S.D.	Mean	S.D.	Stand Mean diff	Mean	S.D.	Mean	S.D.	Stand Mean diff
Propensity Score	0.23	0.12	0.36	0.22	0.89	0.30	0.14	0.30	0.14	0.05
Community Disadvantage Index - School Tracts	0.56	1.09	1.06	1.48	0.41	0.71	1.24	0.74	1.19	0.02
Proportion households w/ assistance income	0.13	0.09	0.19	0.14	0.57	0.15	0.11	0.16	0.11	0.05
Proportion households in poverty	0.18	0.13	0.23	0.16	0.33	0.20	0.15	0.20	0.14	0.01
Proportion individuals w/o h.s. diploma	0.33	0.13	0.35	0.18	0.14	0.32	0.15	0.33	0.17	0.10
Proportion single parent households	0.14	0.09	0.16	0.08	0.25	0.16	0.12	0.15	0.07	0.09
Proportion individuals unemployed	0.10	0.06	0.12	0.08	0.34	0.09	0.06	0.10	0.06	0.13
Inverse median income (x.001)	0.04	0.02	0.05	0.03	0.44	0.04	0.02	0.04	0.02	0.04
Percent free lunch	0.67	0.23	0.74	0.19	0.30	0.74	0.17	0.72	0.19	0.08

Table 5: Comparison of Standardized Pretest Reading Scores of Success for All (SFA) Schools and Matched Comparison Schools at the Student Level.

	Other Schools			SFA					
	Mean	S.D.	n	Mean	S.D.	n	Stand. Mean diff	t-value	sig
Kindergarten (WJ Language Arts)	-0.01	0.97	471	-0.14	1.05	434	0.13	1.90	
Kindergarten and 1st grade	-0.01	1.00	642	-0.13	1.02	608	0.12	2.12	*
Kindergarten, 1st grade and 2nd grade	-0.02	0.99	735	-0.11	1.02	722	0.10	1.82	

Note. * $p < .05$

Table 6. 3-year longitudinal results

	YEAR 1 (Kindergarten) (N= 54, n= 842)			YEAR 2 (1st Grade) (N= 49, n= 568)			YEAR 3 (2nd Grade) (N= 49, n= 453)		
Type of Measure	Effect	SE	t	Effect	SE	t	Effect	SE	t
School mean achievement									
Intercept	-0.05	0.06	-0.76	-0.12	0.08	-1.47	-0.08	0.06	-1.25
SFA	**0.12**	0.11	1.18	**0.34****	0.11	2.98	**0.26***	0.13	1.97
Mean pretest score	0.48**	0.16	3.11	0.67** *	0.17	3.90	0.48*	0.21	2.27
Random effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.08	125.97	51	0.09	98.30	46	0.11	96.67	46
Within-school variation	0.90			0.87			0.87		
Variance explained (%)									
School mean achievement	19.2			39.5			19.1		

Note. * *p* <.05, ** *p*<.01, *** *p*<.001.

Table 7. Year 1 and Year 2 results, Two Years of Treatment

	YEAR 1 (Kindergarten & 1st Grade) (N's= 54 schools, 1134 students)			YEAR 2 (1st and 2nd Grade) (N's= 49 schools, 759 students)		
Type of Measure	Effect	SE	t	Effect	SE	t
School mean achievement						
Intercept	-0.05	0.05	-0.94	-0.14*	0.07	-2.08
SFA	**0.16**	0.09	1.83	**0.40*** **	0.10	3.97
Mean pretest score	0.60***	0.13	4.78	0.78***	0.16	5.01
Grade						
Intercept	1.08***	0.09	12.66	0.56***	0.06	9.27
Random effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*
School mean achievement	0.06	139.94	51	0.07	103.78	46
Within-school variation	0.72			0.81		
Variance explained (%)						
Within-school variation	22.0			6.4		
School mean achievement	34.1			53.2		

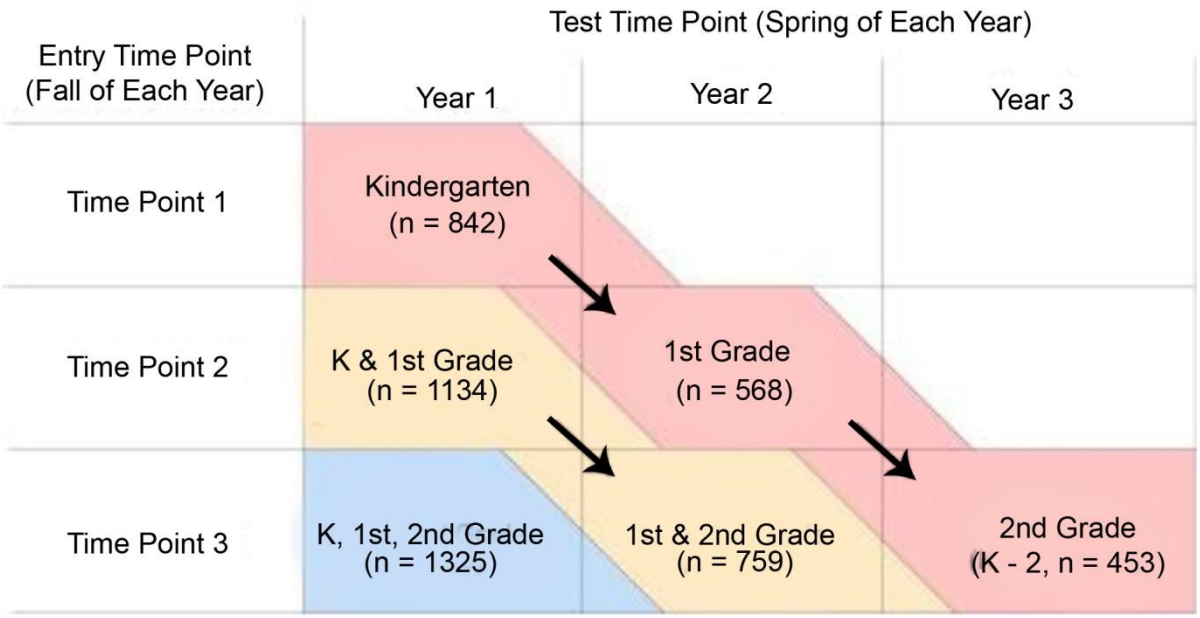
Note. * $p < .05$, *** $p < .001$.

Table 8. One- year results

	YEAR 1 (Kindergarten to 2nd Grade) (N's= 54 schools, 1325 students)		
Type of Measure	Effect	SE	t
School mean achievement			
Intercept	-0.07	0.05	-1.24
SFA	**0.16**[a]	0.09	1.78
Mean pretest score	0.59***	0.12	4.78
Grade			
Intercept	0.86***	0.04	24.05
Random effect	Estimate	χ^2	*df*
School mean achievement	0.07	217.72	51
Within-school variation	0.55		
Variance explained (%)			
Within-school variation	39.9		
School mean achievement	28.4		

Note. a p<.10; *** *p<.001*

Figure 1: Description of the sample



Note. Students retained in grade were included in the analysis.