Educational Evaluation and Policy Analysis
June 2013, Vol. 35, No. 2, pp. 237–251
DOI: 10.3102/0162373712466940

The Impacts of Success for All on Reading Achievement in Grades 3–5: Does Intervening During the Later Elementary Grades Produce the Same Benefits as Intervening Early?

Paul Hanselman
Geoffrey D. Borman
University of Wisconsin-Madison

We evaluate the impact of Success for All literacy instruction in grades 3 through 5 using data from the same cluster randomized trial used to evaluate effects in the earlier grades (K–2). In contrast to the early benefits, there is no effect on reading achievement in the later grades, either overall or for students and schools with high or low baseline reading achievement. This suggests that the impact of Success for All—including established long-term positive effects—may depend on early exposure. As a result, educators may experience difficulty replicating the typical positive achievement impacts of the intervention when children participate in Success for All only during the later elementary grades, as is common for mobile students in program schools.

Keywords: *educational policy, Success for All, experimental evaluation, literacy instruction*

Elementary literacy education in the United States still faces great challenges. According to the National Assessment of Educational Progress, only one third (33%) of all fourth-grade students read at a proficient level in 2009; a third (33%) did not read at even a basic level; and half of Black (53%), Hispanic (52%), and free-lunch-eligible (49%) students fell short of basic proficiency (National Center for Education Statistics, 2009). These statistics are alarming because learning to read is a "key milestone for children living in a literate society" (Whitehurst & Lonigan, 2001, p.11), and elementary reading skills beget later ability, academic achievement, and adult status (i.e., Cunningham & Stanovich, 1997; Entwisle & Alexander, 1999; Kraus, 1973). These realities provide motivation for the development and rigorous evaluation of elementary reading initiatives but, thus far, the body of evidence is disproportionately thin on the effectiveness of instructional approaches for later elementary students. For instance, as of late 2011, the What Works Clearinghouse included 128 studies meeting evidence standards of 60 interventions focusing on the beginning elementary grades (kindergarten, first, and second) and just 67 rigorous studies of 39 interventions focusing on the later elementary grades (third, fourth, and fifth).[1]

Literacy instruction in the later elementary grades merits greater independent attention because it is conceptually and practically distinct from early instruction. The conceptual goals in the early grades focus on beginning reading skills, whereas later elementary goals shift to comprehension; therefore, best instructional practices may differ across contexts (Slavin, Lake, Chambers, Cheung, & Davis, 2009). A major practical challenge for later elementary literacy instruction is the diversity of students' prior educational experiences due to student mobility. Only 56% of students, on average, attend the same school from kindergarten to

third grade (Burkam, Lee, & Dwyer, 2009), and student mobility is highest in the types of schools most targeted by school reform (Kerbow, 1996). This means that a substantial population of students receives partial exposure to a school's overall curricular program and, therefore, instructional strategies must be sensitive to limited prior exposure. For literacy instruction in the later grades, this requires not only evidence of the effectiveness of specific interventions but evidence of impacts independent of earlier experiences. In other words, if a school's instructional approach depends on early exposure for success, then it will not serve the needs of mobile students who arrive in the later grades.

This article specifically addresses literacy instruction in the later elementary grades by isolating the instructional impacts of a promising intervention, Success for All, in grades 3 through 5. Success for All is an instructive case for this purpose because it is among the most mature and proven school reform models, with rigorous causal evidence demonstrating a positive impact of instruction in the early elementary grades (Borman et al., 2007) as well as observational evidence of long-term benefits of the program (Borman & Hewes, 2002). However, relatively little attention has been paid to the program's literacy instruction in the later grades, which are designed around evidence-based practices, widely implemented, and potentially contribute to the program's success. In addition, the national randomized trial of Success for All provides the unique methodological opportunity to isolate the causal effect of Success for All literacy instruction in grades 3 through 5 and to compare these impacts to those in grades K through 2. In contrast to benefits in the early grades, we find no effects of instruction in the later grades. Also, we find no differences in the program's effectiveness by students' prior achievement levels.

Success for All and Later Elementary Literacy Instruction

Originally developed in Baltimore, Success for All is a long-standing, widely implemented, and effective comprehensive school reform model for elementary literacy instruction (Borman & Hewes, 2002; Borman, Hewes, Overman, & Brown, 2003; Borman et al., 2005a, 2005b; Borman et al., 2007; Slavin, Madden, Chambers, & Haxby, 2009). However, no research specifically focuses on *Reading Wings*, the literacy program for elementary students at the second-grade level and above (as opposed to *Reading Roots*, the earlier program). *Reading Wings* is an adaptation of Cooperative Integrated Reading and Composition (CIRC) (Stevens, Madden, Slavin, & Farnish, 1987). Students are regrouped from across grade levels into reading classes based on reading level, and classroom instruction is structured around direct instruction, cooperative work in small groups, and regular individual assessments. The classroom instruction for each lesson follows a guide developed to supplement regular reading materials (basals, novels, etc.), which walk teachers through materials for direct instruction, cooperative tasks, and ultimate assessment and celebration. Throughout, the curricular focus of these lessons is comprehension and appreciation of increasingly complex text (Slavin, Madden, et al., 2009).

To date, the evidence for benefits of *Reading Wings* is indirect and inconclusive. Success for All implementation, which includes curricular supports and training for teachers, leads to a greater emphasis on reading comprehension in classroom instruction (Correnti & Rowan, 2007). This suggests that the evidence-based instructional strategies do, indeed, make their way to students. In addition, positive impacts in the later grades would be consistent with the long-term observed impacts of the overall Success for All program (Borman & Hewes, 2002). However, since some observational evidence suggests that the success of Success for All is limited to the early grades (Venezky, 1998), such putative benefits demand additional consideration.

In sum, the Success for All instructional regime presents a specific and underutilized case to assess the effectiveness of literacy instructional strategies for students in the later elementary grades. Therefore, this article takes up the following three research questions:

Research Question 1: What is the impact of the Success for All literacy instructional program in grades 3 through 5 on students' reading achievement?

Research Question 2: Are the impacts of Success for All literacy instruction in grades 3 through 5 comparable to the effects of the Success for All instructional approach in grades K through 2?

Research Question 3: Does the effect of Success for All instruction in the upper grades differ for students and schools exhibiting lower or higher initial proficiency?

The National Randomized Success for All Trial

This study draws on supplemental data from the national cluster randomized trial of Success for All conducted with more than 5,000 students beginning in 2001. The pragmatic design of that study randomized 35 participating schools to one of two conditions for a 3-year period: Success for All implementation in kindergarten through second grade (K–2) or in third through fifth grade (3–5). As previously documented in reports of the K–2 impact study, all schools received some school-level aspects of the intervention, but student exposure to the literacy instructional component in any particular grade was prescribed only by experimental status. This means that the later grade students in K–2 schools (who received no instructional exposure to the Success for All program) provide a randomized comparison group for the grade 3–5 Success for All students.

This design provides several substantial advantages in understanding the effectiveness of the Success for All instructional approach. First, the experimental evaluation provides rigorous causal evidence. To date, all assessments of Success for All in the later grades have relied on quasi-experimental comparisons that are subject to selection bias. Second, the experimental contrast isolates the effect of Success for All instruction, controlling for both school organizational resources and experience with the previous program curriculum. This allows a precise assessment of the effectiveness of the Success for All instructional regime for students in grades 3 through 5. Finally, because the results are drawn from the same study and schools, the impact estimates for grades 3 through 5 are methodologically, contextually, and historically comparable to those for grades K through 2.

It is important to highlight the general complexity of evaluating later elementary instructional strategies, because impacts may be moderated by students' prior instructional experiences. In this case, the experimental contrast identifies the instructional effect in the later elementary grades specifically for a population of students without previous exposure to Success for All. This contrast is designed to assess the independent impact of the later elementary grades program, *Reading Wings*, in order to learn about effective literacy practices in those grades. The experimental contrast does not directly test the effects of later elementary instruction in the context of complete Success for All exposure, so results may not generalize to ideal *Reading Wings* implementation (following *Reading Roots*). However, as a practical matter, student mobility ensures that a substantial population of students will experience *Reading Wings* without *Reading Roots*. In the national Success for All randomized trial itself, approximately one third (32%) of the younger cohort were "in-movers" to the school by the end of second grade (Borman et al., 2007). It is clearly important to know the independent effects of *Reading Wings*.

Method

Sample

This article uses data from 35 schools involved in the national experimental evaluation of Success for All (Borman et al., 2007).[2] We focus on two cohorts of students with exposure to Success for All in the later elementary grades. The primary cohort, students who were in third grade in 2002–2003, experienced the program over all 3 years of the study. We also report results for a secondary cohort of students who were in fourth grade in 2002–2003 and encountered the program for just 2 years in total. However, the structure of the data does not allow students to be linked across years, which precluded identifying a longitudinal sample of students with maximum program exposure throughout the study. Instead, all analyses focus on the "schoolwide" outcomes (Borman et al., 2007) based on all students present at the targeted grade level in each year. Given our focus on the school-level impacts of treatment assignment, this assessment of cross-sectional, schoolwide outcomes is appropriate.

The characteristics of the participating schools (see Table 1) suggest that the Success for All and comparison schools are roughly equivalent on all observed characteristics. In general, study schools are located in the Midwest or South and enroll large proportions of minority students and students eligible for free and reduced price lunch, a proxy measure for family income. We present further evidence that randomization resulted in reasonably matched groups in Table 2. When comparing the two groups, neither demographic characteristics nor initial reading scores revealed statistically significant differences.

The analytical sample contains 2,420 primary cohort students and 2,172 secondary cohort students in the 35 study schools in Year 1. Study attrition reduces the analytical sample for subsequent years in two ways. First, six schools left the study during subsequent years: five due to school closures and one due to refusal to allow continued data collection (Borman et al., 2007). Second, grade-specific data refusals in grades 4 and 5 further limit the analytical sample for both cohorts of interest here. Although sample attrition is regrettable in any research context, it only jeopardizes the validity of an experimental impact estimate if attrition is excessive overall or experienced differentially between treatment and control groups. We devote the first section of our results to these issues.

Measures

Throughout the study, general reading achievement was assessed with the Gates-MacGinitie Reading Test 4th edition, levels 3–5, form S (GMRT), produced by Riverside Publishing. The GMRT is a nationally normed 55-minute pencil and paper assessment designed to assess general reading achievement. The test for levels 3–5 assesses both vocabulary knowledge, by requiring students to identify word meanings with minimal context, and reading comprehension, by asking students to demonstrate understanding of published prose excerpts. Internal consistency reliabilities for the GMRT in levels 3–5 range from .95 to .96, and test–retest reliabilities range from .89 to .93.

Students were pretested at the start of the study (in the fall of 2002) and administered a posttest during the spring of each of the 3 years of the study (2003, 2004, and 2005). The testing windows for the posttest were approximately 4 weeks in length, and posttesting occurred no earlier than 8 weeks prior to the final day of the school year. For analyses presented here, the scores have been standardized to have a mean of zero and a standard deviation of one within each cohort and year. We also use grade-level equivalent thresholds in the original scale-score metric to define sub-populations of students reading above or below grade level at baseline.

Procedures

The sample of 35 schools was recruited to begin implementation in fall of 2002. As is typical with the implementation of Success for All, all 35 of the study schools established a schoolwide "solutions" team, which addressed classroom management issues and sought to increase parents' participation in school generally, to mobilize integrated services to help Success for All families and children, and to identify and solve particular problems such as irregular attendance, problems at home, and homelessness. In addition, each of the 35 Success for All schools designated a full-time program facilitator who monitored the daily operation of the program, provided assistance where needed, and coordinated the various components. Because all students across grades K through 5 could, potentially, benefit from these schoolwide supports, and because no students in the grade 3–5 sample had prior exposure to the K–2 Success for All curriculum, our impact estimates are interpreted as the unique effect of the grade 3–5 Success for All instructional approach.[3] To assure fidelity to experimental conditions within schools, Success for All monitored treatment and control classrooms during regular quarterly visits. Most important, no evidence was found that control classrooms adopted the curriculum, instruction, and learning environments that constitute the Success for All treatment. See Borman et al. (2005a, 2005b, 2007) for additional details about the Success for All experimental evaluation.

Analysis

Models to identify the treatment effect of an intervention in a cluster randomized trial (CRT)

TABLE 1

Baseline Characteristics of Schools Participating in the Success for All (SFA) Randomized Trial, Grouped by Grade 3–5 Treatment Assignment

School	District	State	Enrollment	White (%)	African American (%)	Hispanic (%)	Female (%)	ESL (%)	Special Education (%)	Free Lunch (%)
SFA										
Augustin Lara	Chicago	IL	574	2.3	1.1	96.0	50.0	56.0	7.3	96.0
Bluford	Guilford	NC	401	3.7	92.6	1.8	48.4	0.0	21.9	45.6
Bunche	Chicago	IL	396	0.0	100.0	0.0	49.5	0.0	7.0	99.0
C. F. Hard	Bessemer	AL	398	0.0	99.8	0.0	46.7	1.0	14.6	88.4
Central	Central	KS	131	95.0	0.0	2.0	47.0	0.0	6.0	51.0
Cupples	St. Louis	MO	171	0.0	100.0	0.0	56.0	0.0	2.0	98.0
Daniel Webster	Chicago	IL	636	0.0	100.0	0.0	44.0	0.0	5.0	98.3
Dewey	Chicago	IL	436	0.0	99.5	0.0	51.0	0.0	23.0	100.0
Edward E. Dunne	Chicago	IL	560	1.0	99.0	0.0	76.0	0.0	7.0	97.0
Eutaw	Greene	AL	316	1.0	99.0	0.0	48.0	0.0	5.0	90.0
Greenwood	Bessemer	AL	395	13.0	73.0	14.0	53.0	1.0	11.0	84.0
Gulfview	Hancock	MS	520	94.0	2.6	1.0	49.0	1.0	18.0	71.0
Jamestown	Guilford	NC	496	40.1	51.6	4.6	47.6	5.8	16.0	46.6
Sigel	St. Louis	MO	302	8.3	82.8	2.7	45.4	6.8	18.2	96.4
Scullin	St. Louis	MO	282	0.0	100.0	0.0	42.0	0.0	13.0	98.0
South Delta	South Delta	MS	640	5.5	93.5	1.0	47.8	0.0	5.5	100.0
Stanfield	Stanfield	AZ	757	19.4	1.0	67.7	50.3	50.0	19.0	100.0
SFA school means			436	16.7	70.3	11.2	50.1	7.2	11.7	85.8
Control										
Benjamin E. Mays	Chicago	IL	418	0.0	90.0	10.0	49.0	0.0	10.0	95.0
Bertha S. Sternberger	Guilford	NC	342	69.0	26.0	0.8	50.0	0.0	15.0	21.0
Brian Piccolo	Chicago	IL	980	0.0	78.0	21.0	48.0	11.0	13.0	97.0
Cesar Chavez	Norwalk	CA	466	4.3	2.6	89.1	50.9	69.0	4.3	89.0
Cook	St. Louis	MO	335	0.0	100.0	0.0	45.0	0.0	15.0	100.0
Earl Nash	Noxubee	MS	484	0.4	99.5	0.0	50.8	0.0	4.1	100.0
Farragut	St. Louis	MO	350	0.0	100.0	0.0	44.0	0.0	4.0	98.0
Gundlach	St. Louis	MO	234	0.0	100.0	0.0	46.0	0.0	2.9	97.1
Harriett B. Stowe	Indianapolis	IN	275	25.0	15.0	56.0	46.0	56.0	19.0	97.0
James Y. Joyner	Guilford	NC	381	44.4	47.0	4.5	51.4	5.8	16.0	44.1
Lafayette	St. Louis	MO	297	13.2	72.8	9.4	49.7	25.0	12.0	94.0
Laurel Valley	Ligonier Valley	PA	392	99.0	0.5	0.3	47.0	0.0	8.0	45.0
Linden	Linden	AL	211	0.5	98.6	1.0	46.9	0.1	10.0	91.0
Paramount Jr.	Greene	AL	417	0.0	99.0	0.0	42.7	0.0	9.0	93.0
Pleasant Garden	Guilford	NC	588	79.0	11.8	4.3	48.0	2.9	1.9	28.8
Robert H. Lawrence	Chicago	IL	643	0.0	99.0	1.0	70.0	0.0	4.7	90.0
Waveland	S. Montgomery	IN	148	98.0	0.0	0.0	54.0	0.0	13.0	26.0
Wood	Tempe	AZ	630	21.2	10.9	40.1	50.2	25.5	8.7	48.5
Control school means			422	25.2	58.4	13.2	49.4	10.9	9.5	75.3

Note. ESL = English as a second language.

TABLE 2
Comparison of Baseline Characteristics at Success for All 3–5 (SFA) Schools (N = 17) and Control Schools (N = 18)

Variable	Condition	*M*	*SD*	*t*
Enrollment	SFA	435.94	168.64	
	Control	421.72	195.65	–0.23
% female	SFA	50.10	7.42	
	Control	49.42	5.88	–0.30
% minority	SFA	83.34	31.04	
	Control	74.78	31.04	–0.75
% ESL	SFA	7.15	17.41	
	Control	10.85	20.60	0.57
% special education	SFA	11.74	6.68	
	Control	9.48	5.08	–1.13
% free lunch	SFA	85.84	19.66	
	Control	75.25	29.71	–1.24
GMRT scale score, grade 3	SFA	426.29	18.83	
	Control	430.96	21.03	0.69
At grade level, grade 3	SFA	26.71	18.18	
	Control	32.78	21.97	0.91
GMRT scale score, grade 4	SFA	453.69	18.98	
	Control	458.07	20.96	0.65
At grade level, grade 4	SFA	23.26	19.81	
	Control	27.59	20.94	0.63

Note. GMRT = Gates-MacGinitie Reading Test; ESL = English as a second language.

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

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

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

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

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

where the mean posttest intercept for school j, β_{0j}, was regressed on the baseline school-level mean GMRT score, the SFA treatment assignment indicator, plus a residual, u_{0j}. In this model, the coefficient γ_{02} provides the intention-to-treat (ITT) estimate of the impact of the Success for All program.

In addition, we conducted supplementary analyses to investigate whether Success for All impacts in grades 3 through 5 were sensitive to the prior reading ability of students within the schools.[4] In one supplemental analysis, we interact treatment status with the school average pretest score at baseline. The interaction term in

TABLE 3
Analytic Samples for Success for All 3–5 (SFA) and Control Schools, Student Attrition, and Achieved Minimum Detectable Effect (MDE) Sizes

		SFA			Control			Student Attrition				
Cohort	Year	Schools	Students	St/Sch	Schools	Students	St/Sch	Treatment	Control	$\chi^2(df = 1)$	p	MDE
1	1	17	1197	70.4	18	1223	68.5	—	—			.15
1	2	13	874	67.2	13	869	66.8	323	354	1.15	.28	.26
1	3	11	562	51.1	14	789	56.4	635	434	75.66	<.01	.33
2	1	17	1084	63.8	18	1088	60.4	—	—			.19
2	2	14	842	60.1	16	849	53.1	242	239	0.04	.84	.32

these models can be interpreted as a test of whether the treatment was more effective in initially high- versus initially low-performing schools. Our second supplemental analysis focuses on the 1st year of the study, when we are able to identify a longitudinal sample of students with both fall pretest and spring posttest observations. This allows us to repeat our main 1st year analyses for the sample of students reading at or above grade level at the beginning of the study and for those reading below grade level.

Results

Sample Attrition

We begin by assessing sample attrition throughout the 3 years of the study. We focus on the potential threat to the internal validity of the randomized trial. Most important, if attrition is systematically different between treatment and control groups, then data losses may undermine the baseline equivalence demonstrated above. Specifically, we test whether treatment and control attrition were comparable in two ways. First, corresponding to our main analyses, we test whether the level of attrition of the upper elementary students throughout the study differed between experimental groups. Second, corresponding to our second supplemental analysis, we investigate attrition during the 1st year of the study among students with valid pretest measures.

Table 3 reports the sample sizes for the GMRT outcome in both study cohorts and both experimental conditions. There was a noticeable drop in the number of cases for all groups between the 1st and 2nd years of the study—between 20% and 30% across cohort and experimental conditions—and an even more substantial drop by the 3rd year. One component of overall attrition is at the school level, which reflects both school closures and non-compliance with testing. School attrition is comparable between the Success for All and control groups. For the primary cohort, 4 of 17 treatment schools and 5 of 18 control schools were lost, whereas for the secondary cohort, 3 of 17 treatment schools and 2 of 18 control schools were lost. Neither difference is statistically significant, $\chi^2(1, N = 35) = 0.08$, $p = .77$, and $\chi^2(1, N = 35) = 0.31$, $p = .58$, respectively. The ratio of student records per school indicates that there is also some student-level attrition within schools between Year 1 and Year 2. Nonetheless, the chi-square tests presented in Table 3 show that the overall attrition in valid test scores from Year 1 to Year 2 is statistically independent of treatment status for both cohorts. In short, there is no evidence of differential attrition in Year 2.

The outcome is different for the 3rd year of the study, during which the primary student cohort is expected to be in fifth grade. First, school attrition is less balanced between treatment and control groups (6 of 17 treatment, and 4 of 18 control), although this difference is not statistically significant, $\chi^2(1, N = 35) = 0.73$, $p = .39$. Second, overall student attrition (635 vs. 434) is different at any conventional level of statistical significance, $\chi^2(1, N = 1{,}691) = 75.66$, $p < .01$. These differences signal that the Year 3 student samples may no longer be comparable and, therefore, that the Year 3 results should be viewed with particular caution.

Attrition is also comparable between experimental groups in the longitudinal 1st-year sample in both the primary (10.27% vs. 11.51%) and secondary (12.67% vs. 12.93%) cohorts.

Neither difference is statistically significant, $\chi^2(1, N = 2{,}397) = 0.95$, $p = .33$, and $\chi^2(1, N = 2{,}179) = 0.03$, $p = .86$, respectively. We also compared the baseline characteristics of students lost over the course of the 1st year. The average pretest scores of treatment and control attrition samples were 443.97 and 443.25, respectively, and not statistically different $t(876) = 0.25$, $p = .80$, suggesting that attrition does not threaten the internal validity of the treatment–control comparison, at least with respect to baseline equivalence on the pre-intervention measure of the outcome. However, students clearly did not leave the sample at random; leavers were lower performing at baseline on average (GMRT = 443.61) than students who remained in the sample (GMRT = 456.24), $t(-7.90)$, $p < .01$ (two-tailed). Therefore, results generalize to a slightly higher achieving sub-group of the entire sample.

In sum, sample attrition is a feature of the national Success for All randomized trial and is increasingly prevalent over time. Inferences will therefore generalize directly to a population of students who are somewhat higher achieving at baseline than all students in the targeted schools. However, attrition rates are similar to those experienced in the K–2 portion of the study, which ranged from 12% to 43% (Borman et al., 2005a, 2005b, 2007), and the available evidence suggests that attrition from both the treatment and control schools is comparable and therefore does not threaten the internal validity of the experimental comparisons. The one exception is the 3rd year of data for the primary cohort, for which significantly higher attrition among the treatment group may signal non-comparable analytic samples.

Program Impacts

The findings are relatively straightforward: There is no evidence of a positive or negative effect of the Success for All program on reading performance in grades 3 through 5. Specific results are presented as follows: Table 4 provides the results of the impact models, Table 5 displays possible Success for All interaction effects with baseline school-level reading scores, and Table 6 reports the results of the subgroup analyses for Year 1 (students at grade level vs. students below grade level).

Table 4 shows that the impact estimates across both cohorts in all 3 years are substantively small, statistically non-significant, and of inconsistent sign (3 positive, 2 negative). Because the outcome variable has been standardized, treatment coefficients in Table 4 represent the estimated effect sizes for Success for All. The magnitudes of the estimated effect sizes, which range from $d = -.08$ to $d = .07$, do not reach the threshold of a "small effect" (Cohen, 1988). There is, therefore, no evidence that the 3–5 Success for All schools differed from the K–2 Success for All schools in terms of upper elementary reading achievement.

In general, these null effects stand in contrast to the positive impacts of Success for All instruction in the earlier grades. Focusing on the total impacts after 3 years, we find essentially no effect in grades 3 through 5 ($d = .02$, $SE = .13$). The comparable impacts in grades K through 2 range from an effect size of $d = .21$ ($SE = .09$) in the Passage Comprehension domain to an effect size of $d = .36$ ($SE = .11$) in the Work Attack domain (Borman et al., 2007). The magnitude of the differences between the early grade K through 2 and later grade 3 through 5 impacts is substantial, ranging from almost a fifth of a standard deviation (.19) to over a third of a standard deviation (.34). However, owing to the uncertainty in both estimates, only the latter difference is statistically significant ($z = 2.0$, $p < .05$).[5] On balance, the pattern of evidence suggests that Success for All instruction in grades 3 through 5 (absent earlier exposure) is less effective, relative to business as usual, than Success for All instruction in grades K through 2.

Table 5 presents the results of models that allow the effect of Success for All to interact with the school mean pretest. There is no discernible trend in the treatment effect by pretest score. If anything, the sign of the interaction term (negative in 4 out of 5 cohort-year combinations) suggests that the program is slightly more effective for schools with initially low-performing students, but the results do not reach conventional levels of statistical significance in any case. Overall, the Success for All program is not systematically more or less effective for students in grades 3 to 5 at lower or higher performing schools. We reach the same conclusion

TABLE 4
Hierarchical Linear Models Predicting Literacy Outcomes for Both Cohorts in All 3 Years

	Cohort 1 (3rd grade)									Cohort 2 (4th grade)					
	Year 1			Year 2			Year 3			Year 1			Year 2		
	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*
Fixed effect															
Intercept	–9.19	0.61	–15.06	–8.73	1.08	–8.06	–9.37	1.38	–6.80	–11.36	0.85	–13.45	–7.14	1.40	–5.11
Pretest[a]	0.02	0.00	15.18	0.02	0.00	8.05	0.02	0.00	6.88	0.03	0.00	13.48	0.02	0.00	5.11
SFA	0.00	0.06	–0.05	0.07	0.10	0.71	0.02	0.13	0.18	0.04	0.07	0.59	–0.08	0.11	–0.66
	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
Random effect															
School mean achievement	0.01	67.9	32	0.05	116.1	23	0.08	155.4	22	0.03	103	32	0.08	199.6	27
Within-school variation	0.82			0.80			0.78			0.77			0.79		
Students	2420			1743			1351			2172			1691		
Schools	35			26			25			35			30		

SFA = Success for All 3–5 Treatment.
[a]Pretest measure reflects the cohort average pretest.

TABLE 5
Hierarchical Linear Models Predicting Literacy Outcomes and Including a Treatment by Pretest Interaction

	Cohort 1 (3rd grade)									Cohort 2 (4th grade)					
	Year 1			Year 2			Year 3			Year 1			Year 2		
	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*
Fixed effect															
Intercept	–9.38	0.82	–11.51	–9.25	1.45	–6.38	–8.74	1.81	–4.83	–12.55	1.10	–11.36	–7.67	1.97	–3.89
Pretest[a]	0.02	0.00	11.59	0.02	0.00	6.37	0.02	0.00	4.89	0.03	0.00	11.38	0.02	0.00	3.89
SFA	0.45	1.25	0.36	1.30	2.19	0.59	–1.56	2.83	–0.55	2.66	1.63	1.63	1.01	2.83	0.36
SFA*Pretest	0.00	0.00	–0.36	0.00	0.01	–0.56	0.00	0.01	0.56	–0.01	0.00	–1.60	0.00	0.01	–0.39
	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
Random effect															
School mean achievement	0.01	68.1	31	0.05	116.2	22	0.09	154.6	21	0.03	92.1	31	0.08	198.1	26
Within-school variation	0.82			0.77			0.78			0.77			0.79		
Students	2420			1743			1351			2172			1691		
Schools	35			26			25			35			30		

SFA = Success for All 3–5 Treatment.
[a]Pretest measure reflects the cohort average pretest.

TABLE 6
Hierarchical Linear Models of Treatment Effects in Year 1 for Students Initially at and Below Grade Level

	Cohort 1 (3rd grade)						Cohort 2 (4th grade)					
	At grade level			Below grade level			At grade level			Below grade level		
	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*	Estimate	*SE*	*t*
Fixed effect												
Intercept	–5.15	0.62	–8.26	–5.77	0.24	–24.43	–6.21	0.64	–9.70	–7.99	0.32	–25.32
Pretest[a]	0.01	0.00	9.80	0.01	0.00	23.12	0.01	0.00	11.46	0.02	0.00	24.58
SFA	–0.05	0.12	–0.43	–0.01	0.05	–0.10	–0.04	0.12	–0.29	–0.02	0.07	–0.24
	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
Random effect												
School mean achievement	0.08	123.4	33	0.01	68.4	33	0.08	104.3	31	0.04	155.3	33
Within-school variation	0.54			0.41			0.40			0.34		
Students	662			1474			514			1386		
Schools	35			35			33			35		

SFA = Success for All 3–5 Treatment.
[a]Pretest measure reflects the individual pretest score.

based on 1st-year samples, for which we can identify individual students who were reading at or below grade level prior to Success for All implementation (see Table 6). The effects of the program are effectively null among both groups of students in both cohorts.

Discussion

This article provides a unique experimental evaluation of the effectiveness of the Success for All reading instructional program in grades 3 through 5. In summary, we found neither a positive nor a negative effect of the instruction on student reading achievement in the upper elementary grades. We found similarly null effects on literacy outcomes for students arriving in third grade at or below grade level. Since these results are based on a sample of students and schools without prior participation in the program, they do not directly test the intended implementation of grade 3–5 instruction after the K–2 curriculum.[6] Nonetheless, these results do speak directly to practical and substantive issues relating to the Success for All program and literacy instruction in the later grades.

The most important practical implication of these results is that Success for All may not be beneficial for students who are not exposed to the program before third grade. This is significant because student mobility leads many students to have just such limited exposure to school-based reform initiatives (Kerbow, 1996). Almost half of all students in the United States move schools at least once between kindergarten and third grade, with the highest rates of mobility among poor, minority, and low-achieving students (Burkam et al., 2009; Hanushek, Kain, & Rivkin, 2004). This suggests that mobility and variable exposure are especially relevant for the schools that typically adopt the Success for All program. The national experimental trial highlights these realities, both in terms of student demographics (Table 1) and in the fact that 32% of students were "in-movers" to the school prior to third grade (Borman et al., 2007). Our results suggest that the Success for All instructional approach will not particularly benefit the students who arrive at a school in the upper elementary grades. Although Success for All instruction seems to be no worse than the alternative, educators and policymakers need to explore more effective strategies to promote stronger impacts for this important policy group.

The results also speak to the efficacy of the bundle of instructional strategies employed by Success for All in the later elementary grades. Most notable, the *Reading Wings* approach is explicitly structured around the prior CIRC program, which emphasizes cooperative learning as a pedagogical technique to promote reading comprehension skills (Slavin, 1995). It integrates cooperative classroom interactions with periods of direct instruction, student incentives, and regular classroom assessment. Moreover, Success for All's combination of prescribed curricular materials and organizational supports helps teachers to implement these reforms successfully in the classroom (Rowan, Camburn, & Barnes, 2004). In other words, the Success for All case in the upper elementary grades represents a well-resourced and coordinated attempt to implement cooperative learning pedagogy. Although we cannot identify the effectiveness of individual program components, the null results suggest that this cooperative learning approach per se may not represent an improvement over typical instruction in these settings. Our interaction results also imply that the grade 3–5 program is not differentially effective for students at and below grade level.

In light of the previous evidence of positive benefits of Success for All instruction in the early grades (Borman et al., 2007), the current results provide the most rigorous evidence to date that the benefits of Success for All depend on early exposure to the program. Although previous research finds substantial positive benefits of the program overall that persist into middle school (Borman & Hewes, 2002), Success for All devotes a majority of resources to the early grades, and one observational study suggests that positive impacts are limited to the early grades (Venezky, 1998). The current results provide more direct evidence that program benefits depend on early exposure, since instructional intervention beginning in third grade is no better or worse for later elementary students. It is possible that the current methodology understates the impacts of the full Success for All program, to the extent that schoolwide components (which affected all participants in

the study) are beneficial. However, the fact that these impacts are substantially lower than those in the K–2 portion of the study (based on exactly the same methodology) is strong evidence that the design of the Success for All program leads to different consequences for students in different grades.

One limitation of the current design is that it cannot definitively explain why Success for All success depends on early exposure. There are several plausible explanations. First, Success for All may be uniquely effective in the early grades by concentrating instructional resources there, in line with a program philosophy stressing the prevention and early remediation of literacy problems. One of the most prominent, and costly, components of the program is the supplemental one-on-one tutoring provided to struggling readers, typically 20 minutes per day (in addition to 90 minutes of regular reading instruction), which is offered only in the early grades. If this extra instructional time is a key ingredient in Success for All's overall success, then its absence in the upper elementary grades would suggest lower effectiveness there, which would be consistent with our results.

Conversely, the instructional sequencing of Success for All may explain the importance of early exposure. The Success for All reading program in kindergarten and first grade emphasizes the development of language skills and launches students into reading using phonetically regular storybooks and instruction that focuses on phonemic awareness, auditory discrimination, and sound blending. The theoretical and practical importance of this approach for the beginning reader is supported by the strong consensus among researchers that phonemic awareness is the best single predictor of reading ability, not just in the early grades (Ehri & Wilce, 1980, 1985; Perfetti, Beck, Bell, & Hughes, 1987) but throughout the school years (Calfee, Lindamood, & Lindamood, 1973; Shankweiler et al., 1995). As this awareness is the major causal factor in early reading progress (Adams, 1990), appropriate interventions targeted to develop the skill hold considerable promise for helping students develop broader reading skills in both the short and long term. Since such fundamental skills are uniquely emphasized in early Success for All instruction (Correnti & Rowan, 2007), students may not be able to develop these skills sufficiently when they experience only the upper elementary portions of the program. This scenario would be consistent with the possibility that Success for All instruction in the later grades may have a greater impact for students with previous exposure to the program.

Finally, we stress that it is difficult to draw conclusions about the effectiveness of Success for All for students in grades 3 through 5 who do receive the intended earlier exposure, which requires extrapolation beyond the experimental contrast considered here. To the extent that the program is uniquely effective in the early grades, this highlights the priority of exposing as many students as possible to it early, since the benefits may follow students who receive early exposure and then leave (Venezky, 1998). A corollary is that implementing the instructional program in the later grades may not be necessary to achieve the demonstrated long-term benefits of the program. In other words, the achievement benefits of the *Reading Wings* program may be null. Educators and policymakers should weigh this information along with other costs and benefits in specific circumstances. For instance, our results imply that the *Reading Wings* instructional program is no less effective than the likely alternative, and there may be other benefits to schoolwide implementation (aside from the fact that it is generally packaged as such), such as staff coordination, coherence of school mission, and consistent schoolwide collaboration and professional development across all grades and teachers. In addition, the marginal cost of implementing Success for All instruction in grades 3 through 5 is likely quite low (Borman & Hewes, 2002).

More generally, the Success for All case illustrates the challenges associated with providing effective instruction in the later elementary grades. Despite the previous evidence of program benefits overall, the results of the national randomized trial do not offer such evidence for the impacts of the instructional approach in the later elementary grades. Our results suggest that greater attention to the later grades will be necessary to produce similarly positive impacts there.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Award #R305C050055 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Notes

1. Throughout this article, we refer to kindergarten, first grade, and second grade as early elementary grades and grades three, four, and five as later elementary grades. This is an intuitive and pragmatic distinction.

2. There were six schools recruited in the pilot phase of the national trial that were randomized to either full Success for All implementation or no Success for All exposure conditions. These schools are not included in the present study.

3. The schools randomized to the grade 3–5 condition did not have the Reading Roots program or tutors, as these are services typically reserved for students in grades K–2. However, in typical Success for All implementations, low-achieving grade 3–5 students, including in-movers, can be targeted for supplemental tutoring or could be placed in the Reading Roots program if necessary. Without tutors or the Reading Roots program, of course, such services were not available to students in grades 3–5. In this sense, the potential impacts of the grade 3–5 intervention may be underestimated.

4. There are plausible a priori explanations for any interaction result. For instance, in line with Success for All's focus on low-achieving students, the cooperative learning component and other aspects of the program may be most beneficial to students among third to fifth graders below grade level. However, if instruction takes for granted some mastery of the content and skills stressed in the early curriculum, it may be best suited to fostering literacy for those reading at grade level. Finally, the program may be equally effective for both types of students.

5. The fact that we cannot reject the null despite substantial differences between early and late impacts indicates low power for detecting differences between the two studies. However, in post hoc power analyses (available upon request), we did confirm that the 3–5 study itself was sufficiently powered to detect effects of the magnitude of the K–2 estimates.

6. Note that the current results do apply to the impact of Success for All instruction in later grades if impacts are the same for students with and without previous exposure, including if the impacts are the same for all students. Our interaction results provide preliminary support for this possibility, since the experimental impacts did not differ by prior reading achievement, but a more complete assessment lies outside the scope of this article.

References

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

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

Borman, G. D., & Hewes, G. (2002). The long-term effects and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis*, *24*, 243–266.

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

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

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

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

Burkam, D. T., Lee, V. E., & Dwyer, J. (2009). *School mobility in the early elementary grades: Frequency and impact from nationally-representative data*. Paper prepared for the Workshop on the Impact of Mobility and Change on the Lives of Young Children, Schools, and Neighborhoods, June 29–30, National Academies, Washington, DC. Retrieved July 21, 2010, from http://www.bocyf.org/children_who_move_burkam_paper.pdf

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

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

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*, 298–338.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945.

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

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

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

Entwisle, D. R., & Alexander, K. L. (1999). Early schooling and social stratification. In R. C. Pianta & M. J. Cox (Eds.), *The transition to kindergarten* (pp. 13–38). Baltimore: Paul H. Brookes Publishing.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Disruption versus Tiebout improvement: The costs and benefits of switching schools. *Journal of Public Economics*, *88*(9–10), 1721–1746.

Kerbow, D. (1996). Patterns of student mobility and local school reform. *Journal of Education for Students Placed at Risk*, *1*, 147–169.

Kraus, P. E. (1973). *Yesterday's children*. New York: John Wiley & Sons.

National Center for Education Statistics. (2009). *The nation's report card: Reading 2009* (NCES 2010–458). Washington, DC: Institute for Education Sciences, U.S. Department of Education.

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

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

Rowan, B., Camburn, E., & Barnes, C. (2004). Benefiting from comprehensive school reform: A review of research on CSR implementation. In C. Cross (Ed.), *Putting the pieces together: Lessons from comprehensive school reform research* (pp. 1–52). Washington, DC: National Clearinghouse for Comprehensive School Reform.

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

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

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for elementary grades: A best-evidence synthesis. *Review of Educational Research*, *79*, 1391–1466.

Slavin, R. E., Madden, N. A., Chambers, M. E., & Haxby, B. (2009). *2 million children: Success for All*. Thousand Oaks, CA: Corwin Press.

Stevens, R. J., Madden, N. A., Slavin, R. E., & Farnish, A. M. (1987). Cooperative integrated reading and composition: Two field experiments. *Reading Research Quarterly*, *22*, 433–454.

Venezky, R. (1998). An alternative perspective on Success for All. In K. Wong (Ed.), *Advances in educational policy, Vol. 4* (pp. 145–165). Stamford, CT: JAI.

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

Authors

PAUL HANSELMAN is a doctoral student in the Department of Sociology at the University of Wisconsin-Madison, 1025 W. Johnson Street, Room 453, Madison, WI 53706; phanselm@ssc.wisc.edu. His research interests include the role of educational experiences in social stratification, organizational social resources in schools, and the distributional implications of educational policies.

GEOFFREY D. BORMAN is a professor of education and sociology at the University of Wisconsin-Madison. His areas of research include experimental and quasi-experimental design, educational policy, and educational inequality.

Manuscript received December 20, 2010
Revision received August 14, 2012
Accepted August 29, 2012