RUNNING HEAD: Open Court Reading

The National Randomized Field Trial of Open Court Reading

Geoffrey D. Borman

N. Maritza Dowling

Carrie Schneck

University of Wisconsin—Madison

Abstract

In this article we report the achievement outcomes of a national randomized evaluation of Open Court Reading 2005, a K-6 literacy curriculum published by SRA/McGraw-Hill. Forty-nine first through fifth grade classrooms from predominantly minority and poor contexts from diverse locations across the nation participated in the study. Blocking by grade level within schools, the 49 classrooms were assigned at random to receive Open Court Reading curricular materials and professional development or not. Multilevel analyses of classroom-level effects of assignment to Open Court Reading revealed statistically significant achievement advantages for the treatment group on all three of the Comprehensive Test of Basic Skills, 5th edition, Terra Nova literacy posttests. The Open Court effect sizes were $d = 0.16$ for the Reading Composite, $d = 0.19$ for Vocabulary, and $d = 0.12$ for Reading Comprehension. These effects achieved across this diverse group of classrooms and schools from across the nation demonstrate the potential for large-scale improvement of literacy outcomes through the scale-up of Open Court Reading.

The National Randomized Field Trial of Open Court Reading

National statistics indicate that nearly 40% of American children fail to reach functional levels of literacy (U.S. Department of Education, 2005). Perhaps even more distressing, though, are the pronounced gaps in reading achievement between minority and white children and poor and more affluent children. According to the National Assessment of Educational Progress (NAEP) the achievement disparities between fourth grade African American and white, Hispanic and white, and poor and non-poor children are the equivalent of 2½ to nearly 3 years of learning (U.S. Department of Education, 2005). Only 13% of fourth grade African Americans and 16% of Hispanics scored at the proficient level on the NAEP, compared to 41% of whites, and just 16% of those eligible for free lunch scored at the proficient level, compared to 42% of non-eligible students. Indeed, the national movement to improve early elementary literacy instruction and learning has left many of America's poor and minority children behind.

The development of early reading skills plays a critical role in promoting children's overall academic success (Whitehurst & Lonigan, 2001). Children who read well tend to read more and, as a result, they acquire more knowledge in various academic domains (Cunningham & Stanovich, 1998). Differences in early-elementary reading outcomes typically become differences in high school graduation, college attendance, and ultimately adult status (Entwisle & Alexander, 1999; Kraus, 1973). That is, many of the inequalities in school and society can be traced to the first few years of formal schooling and children's initial experiences learning to read (Entwisle & Alexander, 1989; Garnier, Stein, & Jacobs, 1997; Husen, 1969; Kerckoff, 1993; Lloyd, 1978).

Nevertheless, converging evidence from two decades of research suggests that with appropriate instruction, nearly all students can become competent readers (Denton & Mathes,

2003; Lyon, Fletcher, Fuchs, & Chhabra, 2006; Mathes & Denton, 2002; Snow, Burnes, & Griffin, 1998). Indeed, current federal educational programs such as Reading First and No Child Left Behind have attempted to fund and promote the use of research-based practices aimed at improving early literacy and preventing reading difficulties. Though these initiatives have been embraced broadly at national, state, and local levels, the extant literature on core reading programs is limited in terms of rigorous efficacy studies or broader randomized field trials that examine the impact of these programs on children's reading skills.

The Open Court Reading (OCR) program, published by SRA/McGraw-Hill, has been widely used since the 1960s and offers a phonics-based K-6 curriculum that is grounded in the research-based practices cited in the National Reading Panel report (National Reading Panel, 2000). According to market research, OCR is among the top ranked reading series (Education Market Research, 2002). To date, a total of 1,847 districts and over 6,000 schools have adopted the OCR program across the United States (SRA/McGraw-Hill, 2006). As one of two language arts curricula approved by California's State Board of Education, OCR has been used in large urban districts like Los Angeles, Long Beach, and Oakland as well as in smaller districts like Alum Rock, Paramount, and North Sacramento since 2000. There have been large-scale district-wide adoptions as well in large public school systems like the Baltimore City Public Schools, which began implementing OCR district-wide in 1998, and the Charlotte-Mecklenburg Schools, which began its implementation at the start of the 2001-2002 school year. Despite its widespread dissemination, though, OCR has never been evaluated rigorously through a large-scale randomized field trial.

In this article, we describe the results of a multi-site, cluster randomized controlled trial in which 49 elementary school classrooms from grades 1 through 5 were randomly assigned to a

treatment condition, which included delivery of the OCR professional development and curricular materials, or a no-treatment control condition in which teachers continued to deliver their reading instruction as usual. The study followed teachers and students from the randomized classrooms from fall to spring over the 2005-2006 school year. The central question we addressed is, relative to the "business-as-usual" control condition, what is the effect of assignment to receive the OCR professional development and curricular materials on the spring literacy achievement outcomes of elementary-school classrooms? The intention-to-treat analyses we performed to inform this question provide evidence of the causal impacts of a relatively large-scale dissemination of the widely used OCR program.

The OCR Program

Student materials. The OCR curriculum includes grade-appropriate student textbooks, workbooks, decodable books, and anthologies. There are three components to the OCR curriculum: Preparing to Read; Reading and Responding; and Language Arts. First, the Preparing to Read activities, depending upon the grade level, build skill in phonemic awareness, sounds and letters, phonics, fluency, and word knowledge. Second, Reading and Responding focuses instruction on building background, thinking about text prior to reading, developing vocabulary, reading from the student anthology, and emphasizing reading for understanding through instruction in comprehension strategies and skills as well as through inquiry and investigation. Lastly, the Language Arts section emphasizes the writing process, spelling, grammar usage and mechanics, additional vocabulary, penmanship, and listening, speaking, and viewing. These three components are present at all grades levels, but the skills emphasis and amount of time spent on each area vary according to grade level.

Teacher materials. Teachers receive a teacher edition and diagnostic and assessment packages. The teacher editions include scripted lessons of grade-specific units that rely on direct instruction—a structured learning process and set of instructional routines that helps students learn "how to learn" while they build specific skills in reading and other subjects. The teacher's edition also provides the information necessary for teaching systematic, explicit skill instruction. Each grade level is organized around a set of literacy, science, and social studies themes. Each unit is centered on a common theme—for example, friendship—and lasts six weeks. To support the teacher, the OCR program provides a pacing plan outlined in the teacher's Lesson Planner. With district-wide adoptions of OCR, the pacing schedule may be used to help teachers across the district maintain a consistent schedule of instruction. This feature may benefit transient students, because when a student moves to a new school within a district he or she can rather easily pick up exactly where he or she left off, resulting in less learning disruption.

Professional development. In addition to classroom materials, SRA/McGraw-Hill provides initial training and professional development opportunities with SRA consultants. Using a common set of professional development materials, trained SRA consultants carry out the initial training through two to three day summer workshops. Professional development activities are designed to help the participants learn the pedagogical foundations of the program, become familiar with the materials and program design, and practice the instructional routines. Follow-up visits are scheduled by SRA consultants to provide support to teachers as they implement the curriculum. In these subsequent visits, the consultants may model lessons, observe classroom instruction, and provide feedback to teachers.

The OCR Evidence Base

There is a limited but growing body of research that has provided some support for the OCR program and for the curricular and instructional approaches that underlie the program. For instance, results from evaluations of OCR in California suggest that, in comparison to other reading curricula in the state, OCR has been associated with better reading outcomes and was found to be particularly successful with low-performing students (Skindrud & Gersten, 2006; McRae, 2002; Edsource, 2006). In contrast, though, the findings reported by Westat (2001) on the 3-year adoption of OCR in 102 Baltimore city schools suggested that the gains were approximately the same as other schools with competing reading programs. In all cases, though, these non-experimental studies, which utilize non-equivalent control group designs or one-group pretest-posttest designs, do not support strong causal conclusions regarding the achievement effects of OCR.

Although the program has not been subjected to the most rigorous evaluations, the core components of OCR do have some empirical support through research on reading more generally. For example, Foorman, Francis, Fletcher, Schatschneider, and Mehta (1998) contrasted the longitudinal achievement growth of students participating in three different approaches to reading instruction and found that children taught receiving direct instruction in letter-sound correspondences with decodable text—the OCR program—improved in reading at a faster rate and had higher word-recognition skills in grades 1 and 2 than students using embedded phonics (a researcher-developed program) or implicit phonics (whole language) reading programs. Though this evidence did appear to support OCR and explicit, systematic phonics, Foorman et al. (1998) used word recognition rather than comprehension of connected text as the primary outcome measure. In addition, relative to the OCR condition, the comparison conditions were

somewhat weak and untested in the field—in the case of embedded phonics—or conceptually weak in terms of the research literature—in the case of whole language.

Therefore, although the general theory of literacy promoted by OCR is well aligned with converging and highly influential findings from the National Reading Panel (National Reading Panel, 2000) and the National Research Council Report, *Preventing Reading Difficulties in Young Children* (Snow, Burns, & Griffin, 1998), little rigorous evidence has provided direct support for the effectiveness of the program on basic and broader literacy outcomes. Boards of education and legislators across the U.S. developed criteria for the adoption of core reading curricula in the primary grades. A number of states, including California, Texas, North Carolina, and Indiana, recommended that state funds be allocated for research-based reading materials, and an increasing number of school districts responded by implementing comprehensive reading curricula with the OCR series prominent among them. This scale-up and widespread adoption would obviously benefit from more rigorous evidence of the program's effects.

Method

We conceived of this study as a multi-site cluster randomized trial (CRT), with randomization at the level of the classroom within six school sites. This design addressed practical problems, including the potential difficulties of randomizing individual students within classrooms to alternate treatments, and it was well aligned with the theory of how the OCR intervention works, as a classroom-level program targeting the reform of teachers' curricular choices and instructional practices and supported by materials and professional development. Classrooms in the treatment group used OCR 2005, and control group classrooms continued to utilize the reading curricula previously used by their school. Researchers observed the classroom environment and teachers' methods to ensure fidelity to the treatment or control curricula. We

pretested and posttested students from the treatment and control classrooms using literacy tests from the Comprehensive Test of Basic Skills, 5th edition (CTBS/5), Terra Nova

Sample

During the spring and summer of 2005, SRA/McGraw-Hill recruited for the study a select group of schools from across the nation that had not previously used the OCR curriculum. These schools had approached SRA/McGraw-Hill as would any other school interested in adopting the OCR program. However, rather than purchasing the OCR materials and professional development from the vendor, the schools were offered the opportunity to receive free of charge approximately $3,000 in OCR materials as well as free training and support in exchange for participation in the study. Participation in the study required each school to agree to distribute the materials and training only to those classrooms and teachers randomly assigned to the treatment. In addition, schools were asked to abide by the data collection schedule, which included pretesting and posttesting along with periodic implementation checks by the developer. If interested in participation in this randomized field trial, a school official completed a nomination form. The nomination forms were stratified by region, and a random sample of six schools was selected.

The initial study sample consisted of six schools from across the nation, one from the Southeast (North Carolina), two from the South (Georgia and Florida), one from the Midwest (Indiana), one from the Southwest (Texas) and one from the West (Idaho). The schools came from diverse contexts, with two from rural areas, two from suburban areas, and two from urban locales. The minority and poverty concentration varied across the schools, ranging from 10% to 98% minority and 19% to 93% free or reduced-price lunch participation rates. A total of 57 grade 1 through 5 classrooms containing 1,099 students comprised the initial sample.

We implemented a blocked randomization plan within each of the six schools. Each grade level, 1 through 5, that was targeted by each school represented a block in the design. This design ensured that classrooms from each grade level that the schools wished to target with the resources provided by the study would, indeed, receive assistance. In 9 of 15 cases, blocking by grade level produced a matched grade-level pair of classrooms. In the other 6 cases, the block included more than two classrooms. Table 1 provides a list of all of the classrooms in the treatment and control groups, with the classrooms grouped by their respective randomized block, 1 through 15. In the one large Texas school there were 5 blocks composed of multiple classrooms: one in which there were 7 first grade classrooms,; one composed of 5 second grade classrooms; another with 5 third grade classrooms,; one with 5 4th grade classrooms; and one block comprising 4 5th grade classrooms. The school from Indiana had 5 first grade classrooms in one randomized block. The randomized block design provided a fair and ethical way to distribute limited resources in a uniform manor among the grade levels targeted by each school and ensured that the treatment and control groups would also be matched by grade level.

In addition to these practical advantages of the design, blocking can improve statistical power to detect the treatment effects by reducing unexplained variation in the outcome across classroom clusters (Raudenbush, Martinez, & Spybrook, 2007). As long as there is substantial variability across the classrooms to explain and as the within-block correlations increase, matching will substantially improve statistical power. Further, the blocking ensured that the student and classroom samples would be identical with respect to the school context and grade level. With a relatively small number of classrooms, a simple random assignment design without blocking could have resulted in chance differences across these salient dimensions and may have

compromised the face validity of the study. These practical, statistical, and face validity concerns motivated the sampling and randomized block design.

Sample attrition. Two types of sample attrition occurred in the study: First, cluster-level non-response occurred through the refusal of one Southern school from Georgia to participate in the posttesting. The school cited constraints associated with the local testing program and other data collection demands that were too burdensome. This resulted in the attrition of two grade 2 and two grade 3 classrooms from both treatment and control conditions, or a total of 4 treatment and 4 control classrooms and 161 children. Second, from among the 49 remaining classrooms across the five schools participating in the posttesting, student-level data attrition occurred due to absences of students from school and their unavailability for posttesting. A total of 28 control students and 29 treatment students were not available for the Reading posttest, 29 control and 33 treatment students missed the Vocabulary posttest, and 30 control and 33 treatment children had missing data for the Reading Composite outcome. Considering both forms of attrition together, 106 control and 112 OCR students had missing data for the Reading outcome, 107 controls and 116 OCR students had a missing Vocabulary posttest, and 108 control and 116 treatment students were missing the Reading Composite posttest score.

Therefore, the final analytical sample was composed of 5 schools from which 49 grade 1 through 5 classrooms and 938 students participated. The sample and data attrition claimed approximately 20% of the control students and 18% of the treatment students and 14%, or 8 of 57, of the participating classrooms. Listwise deletion of student cases with missing posttest data did not cause differential attrition rates by program condition, $\chi^2(1, N = 1{,}099) = 0.02$, $p = 0.99$. Similarly, when considering only the 49 classrooms from the five schools participating in the posttesting, the proportion of students with missing data for the outcome variables did not differ

by treatment status, χ^2 (1 N = 1,099) = 0.07, p = 0.97, with attrition claiming approximately 7% of control students and 6% of treatment students.

While conceding that there were some limitations due to cluster-level and student-level attrition, it is the case that the overall rates of attrition were relatively low, claiming only 14% of the 57 baseline classrooms and less than 20% of the 1,099 baseline students from within the 57 classrooms. Further, there is no conflict in this experiment between random assignment of treatment and missing at random. As noted by Rubin (1976) and Little and Rubin (1987), the missing data process is *ignorable* if, conditional on treatment and fully-observed covariates, the data are *missing at random* (MAR).

Measures

Students in grades 1-5 were pretested and posttested on the Comprehensive Test of Basic Skills, 5th edition (CTBS/5), Terra Nova Reading Comprehension and Vocabulary tests by independent trained testers hired and supervised by the researchers. The CTBS/5 Terra Nova pretests were administered to all students during the week of October 17 through 21 of 2005 and the posttests were administered during the week of May 22 through 26 of 2006. To preserve the anonymity of individual students, we did not attempt to match individual students from pretest to posttest. Instead, we used classroom-level pretest means as covariates in our analytical models predicting classroom-level of effects of assignment. We modeled student-level posttest scores as the outcome measures in our multilevel models. This approach capitalized on the power of the cluster-level covariate to improve the precision of the classroom-level treatment impact estimate (Bloom, 2006), but captured student-level outcomes on the posttests while maintaining the anonymity of the students.

According to its most recent Technical Report, the CTBS/5, Terra Nova represents a national curriculum in the sense that content is systematically sampled from across curricula all over the country (CTB/McGraw-Hill, 2001). In this respect, it was an appropriate and relevant measure for the diverse classrooms, which were sampled from across the United States. The CTBS/5, Terra Nova Reading comprehension subtest requires students to read passages of text and respond to questions about the text that measure information regarding students' comprehension of it. The majority of reading passages were taken from published work and the passages are linked together by coherent themes. The Vocabulary subtest assesses both in-depth mastery of word meanings and the students' acquisition of transferable skills or strategies for the ongoing process of word mastery. Passage and paragraph-related questions serve to reinforce the concept that words are more than isolated sets of symbols to memorize and that students must have a good grasp of text meaning to accurately discriminate among plausible answer choices.

Finally, the Reading Composite measure is formed as a simple average of each student's Vocabulary and Reading Comprehension outcomes and expresses the overall outcome for students in the reading domain assessed by the CTBS/5, Terra Nova. We present Reading Composite scores as an indicator of outcomes and program effects in the overall reading domain and also provide the Reading Comprehension and Vocabulary outcomes to document the magnitudes of program impacts in these specific skill areas.

Treatment fidelity. Follow-up visits by SRA consultants to provide support to teachers as they implemented the curriculum also served as implementation checks for the study. These visits established the fidelity of each classroom and teacher to the OCR program and provided consultants an opportunity to work with teachers in setting goals towards improving implementation. During the visits, the consultants modeled lessons, observed classroom

instruction, and provided feedback to teachers, as they would with any other implementation of the program. These procedures, followed in all OCR implementations, were used in the study classrooms to attempt to obtain a high level of fidelity of implementation.

As of January 2006, all treatment classrooms were implementing OCR. There was some variability in implementation, which may be the subject of future analyses. For instance, though the program calls for daily delivery of 2.5 hours of OCR literacy instruction for grades 1 through 3 and 2 hours for grades 4 through 6, some teachers devoted only 90 minutes per day to OCR instruction. The late recruitment of several schools and their classrooms also inhibited quality implementation in some circumstances.

In the control classrooms, teachers were reminded by the consultants to continue using their usual materials and approaches, and not to use anything from OCR. During the implementation visits, the consultants also observed control classrooms. Specifically, these observations focused on whether the materials, instruction, and behaviors in the control classrooms resembled those of the OCR classrooms. In no case did the trainers observe teachers in control classes using OCR materials and implementing OCR practices. It is possible that some ideas or procedures from OCR did influence instruction in the control classrooms, but no overt influences were observed. Instructional materials and core procedures were clearly distinct from each other in the treatment and control classrooms.

Results

The final analytical sample, which included 27 treatment classrooms and 22 control classrooms, is shown in Table 1. As Table 1 illustrates, the blocked randomization design produced samples of treatment and control classrooms with similar overall characteristics. Treatment classrooms were, on average, composed of 71% minority students and 77% of the

Table 1

Profile of Classrooms Participating in the Randomized Evaluation of Open Court Reading, Grouped by Experimental Condition.

Randomized Block	Open Court (*n* = 27)							Control (*n* = 22)						
	State	Grade level	Class Size	% Free Lunch	% ESL	% Special Education	% Minority	State	Grade level	Class Size	% Free Lunch	% ESL	% Special Education	% Minority
1	ID	1	19	78	44	17	50	ID	1	24	72	33	17	56
2	ID	2	22	70	22	0	35	ID	2	21	70	26	0	48
3	ID	3	22	58	8	0	8	ID	3	22	72	0	4	68
4	ID	4	21	53	16	5	32	ID	4	19	71	33	0	48
5	FL	1	18	48	0	17	39	FL	1	21	28	4	16	28
6	FL	2	28	19	0	35	19	FL	2	16	68	0	16	74
7	FL	3	21	40	0	10	35	FL	3	18	33	0	24	17
8	NC	2	15	100	7	7	100	NC	2	13	80	13	7	87
9	NC	5	16	71	0	29	88	NC	5	18	61	6	6	89
10	TX	1	15	100	6	6	100	TX	1	15	100	63	0	100
	TX	1	15	100	0	0	100	TX	1	15	100	0	0	100
	TX	1	15	100	0	13	100	TX	1	17	100	0	6	100
	TX	1	16	100	0	12	100							
11	TX	2	17	100	0	11	100	TX	2	18	100	0	0	100
	TX	2	19	100	0	0	100	TX	2	18	100	0	0	100
	TX	2	20	100	100	0	100							
12	TX	3	18	100	0	17	94	TX	3	17	100	0	5	95
	TX	3	17	100	0	12	94	TX	3	18	100	0	0	100
	TX	3	15	100	84	0	100							
13	TX	4	18	100	38	5	100	TX	4	19	100	0	14	100
	TX	4	21	100	0	9	100	TX	4	19	100	0	14	100
	TX	4	21	100	0	10	100							
14	TX	5	20	100	0	19	95	TX	5	18	100	0	10	95
	TX	5	15	100	0	4	100	TX	5	19	100	4	0	100
15	IN	1	23	9	0	0	4	IN	1	23	8	0	4	13
	IN	1	22	22	9	0	4	IN	1	24	13	9	0	13
	IN	1	22	0	4	4	9							
Means			19.48	77	13	9	71			18.73	76	9	6	74

students received free or reduced-price lunches. Similarly, the composition of control classrooms was 74% minority and 76% free or reduced-price lunch. The OCR classrooms had slightly higher percentages of students identified as special education participants and English-language learners and had a somewhat higher average class size. Thus, although there were some exceptions—most notably, the school and classrooms from Indiana—the study classrooms were from schools and communities with rather high concentrations of minority and poor children and families.

The results shown in Table 2 provide direct comparisons of the baseline characteristics of the treatment classrooms and control classrooms. As the results indicate, the percentages of minorities, special education students, free lunch participants, and English as second language students were statistically equivalent across treatment and control classrooms. Likewise, *t*-tests for the three baseline CTBS/5, Terra Nova outcomes, Reading Comprehension, Vocabulary, and Reading Composite, for the treatment and control schools showed no statistically significant differences.

Therefore, the treatment and control samples were sufficiently well-matched at baseline on key demographic characteristics and the CTBS/5, Terra Nova pretest measures. The sample is also comprised of schools from diverse locales, including high-poverty urban and rural schools across five states. In these respects, the sample selection process and randomization procedure appear to have produced a baseline sample of classrooms that has good internal validity—because there are no large, statistically significant treatment/control differences—and good external validity—because the sample has demographic characteristics that include schools and classrooms from across a range of regional contexts representing the national reach of the program.

Table 2

Comparison of Baseline Characteristics for Open Court Treatment Classrooms and Control Classrooms

Variable	Condition	*N*	*M*	*SD*	95% CI for Difference		*t*
					Lower bound	Upper bound	
Reading Comprehension	Open Court	27	591.11	41.72	-21.34	25.12	0.16
	Control	22	589.22	38.25			
Reading Vocabulary	Open Court	27	562.95	53.83	-28.84	31.49	0.09
	Control	22	561.63	50.12			
Reading Composite	Open Court	27	577.31	47.01	-24.79	27.86	0.12
	Control	22	575.78	43.71			
% Minority	Open Court	27	0.71	0.38	-0.24	0.17	-0.33
	Control	22	0.74	0.32			
% ESL	Open Court	27	0.13	0.26	-0.09	0.17	0.61
	Control	22	0.09	0.16			
% Special Education	Open Court	27	0.09	0.09	-0.02	0.07	1.05
	Control	22	0.06	0.07			
% Free Lunch	Open Court	27	0.77	0.33	-0.18	0.19	0.03
	Control	22	0.76	0.30			

Table 3 summarizes the student-level posttest means and standard deviations by grade level, and across all grade levels, for both treatment and control. The effect sizes, *d*, presented in the far right column represent the differences between treatment and control divided, or standardized by, the pooled student-level posttest standard deviations. First, the table shows, as also suggested by Table 1, that the majority of students and classrooms involved in the study were from the early grades, 1 through 3. Second, of course, the vertically equated scale scores for the three literacy measures increase in value from grade 1, where they range from approximately 552 to 587, through grade 5, where they range from about 636 to 663. Third, the overall outcomes favored the OCR classrooms and students, with effect sizes across all grades of $d = 0.19$ for the Reading Composite, $d = 0.19$ for Vocabulary, and $d = 0.17$ for the Reading Comprehension subtest. Finally, the impacts consistently favored the treatment classrooms and students across all grade levels, with the one exception of the grade 4 classrooms, where the effect sizes were $d = -0.08$ for the Reading Composite, $d = -0.01$ for Vocabulary, and $d = -0.16$ for Reading Comprehension.

Hierarchical Linear Model Analyses of the OCR Treatment Effects

This cluster randomized trial (CRT) involved randomization at the level of the classroom and collection of the outcome data at the level of the student. With such a design, estimation of treatment effects at the level of the cluster that was randomized is the appropriate method (Bloom, 2005; Donner & Klar, 2000; Raudenbush, 1997). We applied Raudenbush's (1997) relatively recently proposed analytical strategy for the analysis of CRTs: the use of a hierarchical linear model. In this formulation, we simultaneously accounted for both student and classroom-level sources of variability in the outcomes by specifying a multilevel statistical model that estimated the classroom-level effect of random assignment.

Table 3

Posttest Outcomes for the Open Court and Control Samples by Grade Level

Grade	Posttest	Open Court				Control				*d*
		N		CTBS Scale Score		*N*		CTBS Scale Score		
		Classes	Students	*M*	*SD*	Classes	Students	*M*	*SD*	
1	Reading Composite	9	165	575.79	37.19	7	139	567.81	42.26	0.20
	Vocabulary	9	165	563.59	46.01	7	139	551.72	49.93	0.25
	Reading Comprehension	9	165	587.46	36.15	7	139	583.35	42.30	0.11
2	Reading Composite	6	121	610.01	37.50	5	86	599.97	35.10	0.27
	Vocabulary	6	121	596.74	43.41	5	86	590.41	42.10	0.15
	Reading Comprehension	6	121	622.74	38.51	5	86	608.99	39.18	0.35
3	Reading Composite	5	93	642.44	45.35	4	75	623.63	35.42	0.46
	Vocabulary	5	93	633.18	48.64	4	76	616.46	40.25	0.37
	Reading Comprehension	5	93	651.17	47.14	4	75	630.23	35.44	0.49
4	Reading Composite	4	77	637.23	39.33	3	56	640.39	44.14	-0.08
	Vocabulary	4	77	631.08	41.98	3	56	631.34	50.65	-0.01
	Reading Comprehension	4	81	641.42	41.79	3	57	647.96	42.76	-0.16
5	Reading Composite	3	51	661.65	33.19	3	54	644.80	26.30	0.56
	Vocabulary	3	51	660.18	40.25	3	54	636.17	28.46	0.69
	Reading Comprehension	3	51	662.53	33.04	3	55	650.84	38.18	0.33
All Grades	Reading Composite	27	507	614.15	49.36	22	410	604.82	48.55	0.19
	Vocabulary	27	507	604.23	55.85	22	411	593.73	55.49	0.19
	Reading Comprehension	27	511	623.46	48.11	22	412	615.18	47.92	0.17

The method of accounting for variability attributable to the randomization blocks could employ a random-effects approach or a fixed-effects approach. In the random-effects conception, we model the school- and grade-specific blocks as level 3 clusters in the multilevel analysis, with the OCR treatment effect specified as randomly varying across blocks. In this sense, any heterogeneity in the treatment impact across randomization blocks is modeled as a random effect. The fixed-effects approach, described by Schochet (2005), is often more realistic in evaluations of education interventions. The current study, like most education evaluations, included a relatively small number of purposively-selected sites. In many such evaluations, it is untenable to assume that the study sites are truly representative of a broader, well-defined population of sites. Furthermore, as Schochet pointed out, inflating the standard errors to incorporate between-block effects will slant the study in favor of finding internally valid impact estimates that are not statistically significant, thereby providing less information to policymakers on potentially promising interventions.

In the current study, though, we stratified nominated sites by region and selected a random sample of six schools to represent the broader population of similar sites. With the treatment units (classrooms, in this case) then randomly assigned to a research condition within grade levels across sites, the most appropriate design is one in which the grade- and site-specific effects are treated as random. As Schochet (2005) indicated, the random-effects approach is typically employed in large-scale studies of well-established interventions that require externally valid impact estimates (and where the burden of evidence of program effectiveness is set high). In random-effects designs, study results can be generalized more broadly than in the fixed-effects designs. In addition, modeling the OCR impact as a random effect across the grades and sites helps us explore directly the extent to which the treatment effect randomly varies across contexts

or whether it is relatively stable and homogeneous across grades and schools. However, these benefits involve costs in terms of precision, because the variance formulas must be inflated to account for between-site random effects rather than fixed effects. Intuitively, in repeated sampling, different sets of sites would be selected for the evaluation, which could produce variability in the impact findings. Therefore, the variance expressions must account for the extent to which mean student outcomes may vary across sites.

Specification of the Multilevel Model. The fully specified level 1, or within-classroom model nested students within classrooms within randomization blocks. The linear model for this level of the analysis is written as

$$Y_{ijk} = \pi_{0jk} + e_{ijk},$$

which represents the spring posttest achievement for student *i* in classroom *j* and block *k* predicted by the classroom mean achievement intercept plus the student-specific level-1 residual variance, r_{ij}. At level 2 of the model, we estimated OCR treatment effects on the mean posttest achievement outcome in classroom *j*. As suggested by the work of Bloom, Bos, and Lee (1999) and Raudenbush (1997), we included a classroom-level covariate, the applicable classroom mean CTBS/5, Terra Nova pretest score (i.e., Reading Composite, Vocabulary, or Reading Comprehension), 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

$$\pi_{0jk} = \beta_{00} + \beta_{01}(\text{MEANCTBS})_{jk} + \beta_{02}(\text{OCR})_{jk} + r_{jk},$$

where the mean posttest intercept for classroom *j* within block *k*, π_{0jk}, was regressed on the block-level mean, β_{00}, the classroom mean CTBS/5, Terra Nova score, the classroom-level OCR treatment indicator, plus a random classroom effect, u_{0jk}, which is the deviation of classroom *jk*'s

mean from the block mean. In this formulation, we treated the intercept and the OCR treatment impact as random effects and treated the classroom mean CTBS/5, Terra Nova pretest score as a fixed covariate effect. Of course, the parameter of central interest in the analysis is located here at level 2 of the model as the classroom-level effect of assignment to the OCR treatment.

At level 3, we represented heterogeneity across randomization blocks for the two random outcomes, the classroom mean posttest achievement intercept and the OCR treatment effect, while the fixed classroom-level mean pretest covariate effect was predicted only by a level 3 intercept. The specification for these 3 outcomes modeled at level 3 is written as

$$\beta_{00} = \gamma_{000} + u_{000k};$$

$$\beta_{01} = \gamma_{010};$$

$$\beta_{02} = \gamma_{020} + u_{020k},$$

where the block mean, β_{00}, was predicted by the grand mean, γ_{000}, plus a random block-level effect, u_{000}, which represents the deviation of the block *k* mean from the grand mean. The mean pretest covariate effect was predicted at level 3 only by the block-level intercept, γ_{010}. Finally, the level 3 model of the most substantive importance indicated the extent to which the OCR treatment effect, β_{02}, exhibited residual block-to-block variation as captured by the error term u_{020k}.[1]

Preliminary Variance Decomposition. We began our multilevel analyses by specifying unconditional models with no predictors at any of the three levels represented. This preliminary analysis partitioned the variance in the three literacy outcomes across students, classrooms, and blocks. The outcomes of this analysis revealed that between 55% and 62% of the variance in the

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

literacy outcomes was between students within classrooms, 6% to nearly 7% of the variance was located between classrooms, and 32% to 39% of the variability in the outcomes was across randomization blocks. This analysis provided evidence that there was considerable block-to-block variation and that the three-level model, with random effects for blocks, was appropriate.

Much of the variability across the grade- and site- specific blocks, though, was due to simple grade-to-grade differences on the literacy scale score outcomes. For example, the results in Table 3 illustrate that the average posttest outcomes from the lowest grade level, grade 1, to the highest level, grade 5, differed by nearly 100 scale score points. These differences captured by the vertical scale scores were attributable primarily to normative age-related variation in students' outcomes across grade levels rather than to substantive differences across schools and classrooms in terms of value-added learning outcomes. That is, higher scale scores are expected for children from the later than earlier grades regardless of the qualities of the school or the classroom instruction.

Therefore, we specified a second model in which the mean classroom pretest was introduced as a level 2 covariate. The mean classroom pretest covariate was grand mean centered and treated as fixed and, thus, the level 3 block-level posttest mean achievement intercept is interpretable as the adjusted outcome after accounting for differences across blocks on the pretest. After this adjustment for the pretest, the remaining differences in the posttest outcomes may be interpreted more clearly as the residual, or value-added, effects attributable to classrooms and blocks. Adjusting for the pretest accounted for 99% of the variability across blocks in the outcome. The adjustment accounted for relatively less classroom-to-classroom variability, with 52% explained for Vocabulary, 78% explained for Reading Comprehension, and 83% variance explained for the Reading Composite outcome. Thus, after taking into account the

initial classroom-level pretest variability, the block-to-block differences for the adjusted posttest outcomes were considerably smaller in comparison to the preliminary unconditional model.

Fully Specified Model of OCR Treatment Effects. The results for the fully specified multilevel models predicting the three literacy outcomes are presented in Table 4. The estimated coefficients for the average classroom mean, the average classroom OCR impact, and the average classroom pretest covariate effect are tabulated for the Reading Composite posttest in the first set of columns, the Vocabulary posttest in the second set of columns, and the Reading Comprehension posttest in the final set of columns at the far right of the table. With the OCR treatment indicator coded 0.5 for treatment and -0.5 for control and the classroom-mean pretest covariate grand-mean centered and fixed, the intercept is interpretable as the average classroom mean after adjusting for the pretest covariate. This adjusted classroom mean achievement intercept was equal to approximately 611 scale score points for the Reading Composite, 600 scale score points for Vocabulary, and 620 scale score points for Reading Comprehension. The coefficients reported for the average classroom pretest covariate suggest that a 1 scale score point advantage on the pretest was associated with an advantage of between 0.68 and 0.74 scale score points on the literacy posttests.

In all cases, the treatment indicator revealed a statistically significant classroom-level effect of assignment to OCR. As indicated by the coefficients reported in Table 4 for the average classroom OCR impact, these treatment effects were 7.95 scale score points for the Reading Composite, 10.79 scale score points for Vocabulary, and 5.86 scale score points for Reading Comprehension. We also expressed these estimated impacts as effect sizes by dividing each coefficient by the respective student-level control group standard deviation for all grades, which can be found at the bottom of Table 3. The resulting effect sizes were $d = 0.16$ for the Reading

Table 4

Multilevel Models Predicting Student, Classroom-Level, and Block-Level Literacy Outcomes for the Experimental Sample

	Literacy Outcomes								
	Reading Composite			Reading Vocabulary			Reading Comprehension		
Fixed Effect	Effect	*SE*	*t*	Effect	*SE*	*t*	Effect	*SE*	*t*
Average classroom mean	610.82	1.74	351.08***	600.47	2.18	275.09***	619.99	1.69	366.90***
Average classroom OCR impact	7.95	1.83	4.34*	10.79	3.69	3.25*	5.86	2.80	2.80*
Average classroom pretest covariate effect	0.72	0.03	22.86***	0.68	0.03	22.14***	0.74	0.04	16.75***
Random Effect	Estimate	χ^2	*df*	Estimate	χ^2	*df*	Estimate	χ^2	*df*
Students (Level 1)	1327.93			1798.77			1420.19		
Classrooms (Level 2)	3.23	49.83	18	24.32	54.56	18	18.37	59.11	18
Blocks (Level 3)									
Intercept (u_{000})	21.82	29.45	14	38.96	26.59	14	12.18	22.26	14
OCR impact (u_{020})	7.10	12.05	14	97.71	21.99	14	0.82	10.16	14

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

Composite, $d = 0.19$ for Vocabulary, and $d = 0.12$ for Reading Comprehension. Interestingly, in no instances in the fully specified models was there statistically significant random variation at level 3 for the OCR impact. In other words, after adjusting for the pretest, there was no evidence that the treatment effect varied across the grade- and school-specific blocks.

Discussion

The results of the OCR national randomized field trial are quite noteworthy for a variety of reasons. First, the selection and randomization processes worked well. Randomization produced control and treatment samples that were well matched on a variety of baseline characteristics, including demographics and CTBS/5, Terra Nova pretest scores. No statistically significant baseline classroom-level differences were detected. With some effort and expense, we were able to obtain the cooperation of a sufficient number of OCR schools and classroom teachers to provide an acceptable level of statistical power to detect classroom-level effects within a multilevel model framework. No matter how carefully drawn, a sample of six schools and 49 classrooms is not likely to represent the overall population of OCR schools and classrooms with great precision. However, the process does seem to have developed a sample that is regionally diverse and primarily targeted toward the communities most likely to be in need of effective research-based literacy programs—those with high minority and poverty concentrations. Further, through the application of a random effects modeling strategy, we are able to generalize our findings across grade levels targeted by OCR and across schools and classrooms representing diverse locales across the nation.

Second, the data and sample attrition over the first year of the study had few impacts on the good external and internal validity that was achieved through the sample selection and randomization procedures. There was no differential rate of data attrition from the OCR and

control conditions and overall attrition rates remained under 20% at the student level and 15% at the classroom level. Third, the treatment fidelity and OCR implementation quality seem reasonably good. In some schools and classrooms, the tight deadlines involved in the selection and randomization process may have artificially curtailed the quality of OCR implementations. These qualitative differences in implementation quality may be an important subject of future work. Our measurement of the implementation of classroom practices in OCR and control schools will allow us to not only describe implementation variability across classrooms, but to also estimate the causal effects of compliance—through application of statistical models that estimate Complier Average Causal Effects (CACE)—with the OCR components on achievement outcomes.

Fourth, though the effects of OCR are of statistical significance, the interpretation of their practical significance is also enlightening and important. Overall, students from OCR classrooms scored from 12% to 19% of one standard deviation higher on the reading assessments than controls not served by OCR. Using a metric devised by Cohen (1988), U_3, the largest effect size of $d = .19$ for the Vocabulary domain tells us that the average student from an OCR classroom outperformed nearly 58% of his or her control-group counterparts. How should we interpret the magnitude of this effect?

Cooper (1981) has suggested a comprehensive approach to effect size interpretation that utilizes multiple criteria and benchmarks for understanding the magnitude of the effect. For instance, how do the OCR effects compare with the important national achievement gaps in reading? Using data from the Early Childhood Longitudinal Study—Kindergarten Cohort (ECLS-K), we calculated the reading achievement gaps separating African American and white students and poor and non-poor students at the end of the first grade. According to these

nationally representative data, the black-white achievement gap was equivalent to 0.70 *SD*s and the difference between the outcomes of poor and non-poor students was half of one standard deviation. After exposure to OCR, students from the treatment schools held advantages over their counterparts from the control condition that equaled from nearly one fifth to nearly two fifths the magnitude of these gaps.

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

The treatment impact found for a relatively recent and high-profile evaluation, the Tennessee Student-Teacher Achievement Ratio (STAR) study, provides yet another important criterion to which we may compare the OCR effects. This intervention also was targeted toward children in the elementary grades, from kindergarten through third grade. Like the current study, it also applied an experimental design, which included random assignment of children and teachers to small classes of 13-17 students, conventional classes of 22-26, or conventional

classes with a teacher's aide. Also similar to the study reported here, the STAR study involved implementation of an educational intervention at scale, involving 79 schools across the state of Tennessee. Though there were no effects for those students whose classrooms were served by a teacher's aide, Nye, Hedges, and Konstantopoulos (1999) found advantages of $d = .11$ to $d = .22$ favoring the children assigned to receive the class-size reduction over those in conventional classes. Thus, the effect sizes for OCR are essentially equivalent to the impact for class size reductions found through the STAR randomized trial.

Finally, along with other recent efforts, including those of Borman et al. (2005), Cook et al. (1999), and Porter et al. (2005,) these outcomes have helped establish that cluster randomized field trials involving nationally replicated school-based interventions are both possible and desirable for producing unbiased estimates of the effects of educational treatments. As a randomized field trial, rather than a relatively artificial laboratory experiment, the results of this study have strong external validity and relevance for policy and practice related to the scale up of educational interventions. Applying the multilevel analyses, which modeled the school site and grade level blocks as random effects, we found that there is no statistically significant heterogeneity for the OCR treatment effects across schools and grades. The outcomes from these analyses not only provide evidence of the promising one-year effects of OCR on students' reading outcomes, but they also suggest that these effects may be replicated across varying contexts with rather consistent and positive results.

References

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

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

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

Cook, T.D., Habib, F.N., Phillips, M., Settersten, R.A., Shagle, S.C., & Degirmencioglu, S.M. (1999). Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36,* 543-597.

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

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

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

Denton, C.A., & Mathes, P.G. (2003). Intervention for struggling readers: Possibilities and challenges. In B.R. Foorman (Ed.), Preventing and remediating reading difficulties: Bringing science to scale (pp. 229–251). Timonium, MD: York Press.

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

Edsource (2006). *Elementary school curriculum program and API.* Mountain View, CA: Author.

Education Market Research (2002). *Elementary reading market.* Rockaway Park, NY: Author.

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

Foorman, B.R., Francis, D.J., Fletcher, J.M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading disabilities in at-risk children. Journal of Educational Psychology, 90, 37–55.

Garnier, H., Stein, J., & Jacobs, J. (1997). The process of dropping out of high school: A 19-year perspective. *American Educational Research Journal, 34,* 395-419.

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

Husen, T. (1969). *Talent, opportunity, and career.* Stockholm: Almqvist and Wiksell.

Kerckoff, A.C. (1993). *Diverging pathways: Social structure and career deflections.* New York: Cambridge Univeristy Press.

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

Lloyd, D.N. (1978). Prediction of school failure from third-grade data. *Educational and Psychological Measurement, 38,* 1193-1200.

Lyon, G.R., Fletcher, J.M., Fuchs, L.S., & Chhabra, V. (2006). Learning disabilities. In E. Mash and R. Barkley (Eds.), *Treatment of Childhood Disorders* (3d ed.) (pp. 512-591). New York: Guilford.

Mathes, P.G., & Denton, C.A. (2002). The prevention and identification of reading disability. *Seminars in Pediatric Neurology, 9,* 185–191.

McMaster, K.L., Fuchs, D., Fuchs, L.S., & Compton, D.L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71*, 445–463.

McRae, D. J. (2002). *Test score gains for Open Court schools in California: Results from three cohorts of schools* [On-line]. Available: http://www.sraonline.com/index.php/home/ocrr/1115

National Reading Panel (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

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

Porter, A. C., Blank, R. K., Smithson, J. L., & Osthoff, E. (2005). Place-based randomized trials to test the effects on instructional practices of a mathematics/science professional development program for teachers. *The Annals of the American Academy of Political and Social Science*, *599*, 147-175.

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

Raudenbush, S.W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29,* 5-29.

Schochet, P.Z. (2005). *Statistical power for random assignment evaluations of education programs.* Washington, DC: Mathematica Policy Research.

Skindrud, K., & Gersten, R. (2005). An evaluation of two contrasting approaches for improving reading achievement in a large urban district. *The Elementary School Journal, 106,* 389–408.

Snow, C.E., Burns, M.S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

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

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69,* 391–409.

Westat (2001). *Report on the final evaluation of the city-state partnership: New Baltimore City Board of School Commissioners and the Maryland State Department of Education.* Rockville, MD: Westat.

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