Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Findings From a Multiyear Scale-Up Effectiveness Trial of Open Court Reading

Michael Vaden-Kiernan, Geoffrey Borman, Sarah Caverly, Nance Bell, Kate Sullivan, Veronica Ruiz de Castilla, Grace Fleming, Debra Rodriguez, Chad Henry, Tracy Long & Debra Hughes Jones

To cite this article: Michael Vaden-Kiernan, Geoffrey Borman, Sarah Caverly, Nance Bell, Kate Sullivan, Veronica Ruiz de Castilla, Grace Fleming, Debra Rodriguez, Chad Henry, Tracy Long & Debra Hughes Jones (2018) Findings From a Multiyear Scale-Up Effectiveness Trial of Open Court Reading, Journal of Research on Educational Effectiveness, 11:1, 109-132, DOI: 10.1080/19345747.2017.1342886

To link to this article: https://doi.org/10.1080/19345747.2017.1342886

Accepted author version posted online: 30 Jun 2017.
Published online: 26 Jul 2017.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2018, VOL. 11, NO. 1, 109–132
https://doi.org/10.1080/19345747.2017.1342886

Findings From a Multiyear Scale-Up Effectiveness Trial of Open Court Reading

Michael Vaden-Kiernan[a], Geoffrey Borman[b], Sarah Caverly[a], Nance Bell[a], Kate Sullivan[a], Veronica Ruiz de Castilla[a], Grace Fleming[a], Debra Rodriguez[a], Chad Henry[a], Tracy Long[a], and Debra Hughes Jones[c]

ABSTRACT

This multiyear scale-up effectiveness study of Open Court Reading (OCR) involved approximately 4,500 students and more than 1,000 teachers per year in Grades K–5 from 49 elementary schools in seven districts across the country. Using a school-level cluster randomized trial design, we assessed the implementation and effectiveness of Open Court Reading over two years. Implementation study results demonstrated adequate to high levels of fidelity across the treatment schools. Intent-to-treat analyses revealed no statistically significant main effects on students' reading performance in Year 1 and a small negative effect ($d = - .09$) in Year 2. There were positive impacts for particular subgroups, including kindergarten ($d = .12$) and Hispanic ($d = .10$) students in the first year. However, there were negative impacts for first grade ($d = -.13$), females ($d = -.11$), students who were not eligible for free or reduced-price lunch ($d = -.19$), and non-English language learners ($d = -.10$) in the second year of the study. Thus, relative to the "business-as-usual" reading curricula, no positive overall impacts of OCR and mixed impacts for student subgroups were found.

KEYWORDS
Open Court Reading (OCR)
reading performance
reading curriculum
cluster randomized trial

This study addresses the effectiveness of a nationally used core reading program that reflects the research-based best practices recommended by the National Reading Panel (2000). This and other similar programs are increasingly used to prevent reading difficulties and ensure that all children are reading at or above grade level by the end of third grade. 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, Burns, & Griffin, 1998). Yet national data trends over time indicate that 65% of fourth-grade students and 64% of eighth-grade students fail to reach proficient-level reading scores (National Center for Education Statistics, 2013). Initiatives over the past two decades have emphasized the critical role of early reading instruction in preventing reading difficulties, recognizing that students who do not learn to read well by third grade are less likely to build vocabulary and interact with a wide variety of

CONTACT Michael Vaden-Kiernan ✉ mvaden-kiernan@air.org American Institutes for Research, 4700 Mueller Boulevard, Austin, TX 78723, USA.
[a]American Institutes for Research, Austin, Texas, USA
[b]University of Wisconsin–Madison, Madison, Wisconsin, USA
[c]Independent Consultant, Austin, Texas, USA

texts (Good, Simmons, & Kame'enui, 2001). Such failure can have a long-term impact on children's self-confidence, motivation to learn, performance in school, and success in life (Harris & Sipay, 1990; Juel, 1988; Stanovich, 1986, 2000), and reading difficulties are the most common reason for referral to special education services (Donovan & Cross, 2002). Despite these concerns, only a handful of replicable beginning reading programs reviewed by the What Works Clearinghouse (WWC) (2007) have more than one study demonstrating potentially positive effects on students' literacy outcomes: Success for All, Voyager, Reading Recovery, and Ladders to Literacy.[1]

The Open Court Reading (OCR) program, published by SRA/McGraw-Hill Education (MHE) and widely used since the 1960s, offers a phonics-based K–6 curriculum that shows promise for preventing reading difficulties. According to market research, OCR is among the top reading series (Education Market Research, 2002). When the study began, a total of 2,917 districts and more than 8,600 schools have adopted the OCR program across all 50 states and Washington, DC (J. Demery, personal communication, June 16, 2008). Findings from independent, nonexperimental evaluations suggest that, compared with other reading curricula, OCR is associated with statistically significantly better reading outcomes and may be particularly effective with low-performing students (McRae, 2008; Skindrud & Gersten, 2006; Williams, Kirst, & Haertel, 2005). In addition, a cluster randomized efficacy trial documented the impact of the OCR program on reading achievement in Grades 1 through 5 in five schools across the country. The results revealed one-year classroom-level impacts of treatment assignment of approximately one fifth of a standard deviation (.16 for Reading Composite, .19 Vocabulary, and .12 Comprehension) on the Comprehensive Test of Basic Skills (Borman, Dowling, & Schneck, 2008). This was the only prior study of OCR that received the highest WWC rating of *Meets WWC Group Design Standards Without Reservations*; the WWC (2014) determined that OCR has potentially positive effects on general reading achievement and comprehension outcomes. Despite the program's widespread use and promising research findings, OCR has not been evaluated rigorously on a large scale as part of an objective, third-party evaluation.

Overview of the Intervention

The OCR program is widely used and incorporates the instructional practices recommended by the National Reading Panel (2000) related to phonemic awareness, phonics, fluency, vocabulary, and text comprehension. The OCR curriculum includes student materials, teacher manuals, diagnostic and assessment tools, and test preparation practice guides. The program includes a one- or two-day summer workshop at the beginning of each school year to train teachers on program implementation and ongoing support by OCR reading consultants throughout the school year. In all grades (kindergarten through Grade 5), the instructional format is a three-part lesson with specific instruction in phonemic awareness and phonics, vocabulary and comprehension skills, and writing skills. Both informal and formal assessments are used to monitor student progress and inform subsequent instruction.

[1] These programs were rated as having a "medium-to-large" amount of evidence, which requires at least two studies that meet the WWC evidence screen with two schools and a total sample size of at least 350 students or 14 classrooms across the studies. "Potentially positive effects" is evidence of a positive effect in a domain with no overriding contrary evidence.

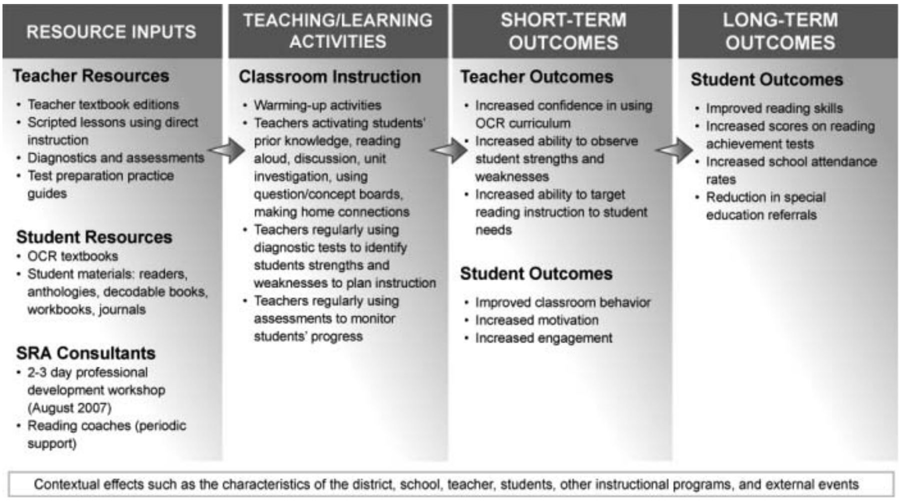
RESOURCE INPUTS
Teacher Resources
• Teacher textbook editions
• Scripted lessons using direct instruction
• Diagnostics and assessments
• Test preparation practice guides
Student Resources
• OCR textbooks
• Student materials: readers, anthologies, decodable books, workbooks, journals
SRA Consultants
• 2-3 day professional development workshop (August 2007)
• Reading coaches (periodic support)
TEACHING/LEARNING ACTIVITIES
Classroom Instruction
• Warming-up activities
• Teachers activating students' prior knowledge, reading aloud, discussion, unit investigation, using question/concept boards, making home connections
• Teachers regularly using diagnostic tests to identify students strengths and weaknesses to plan instruction
• Teachers regularly using assessments to monitor students' progress
SHORT-TERM OUTCOMES
Teacher Outcomes
• Increased confidence in using OCR curriculum
• Increased ability to observe student strengths and weaknesses
• Increased ability to target reading instruction to student needs
Student Outcomes
• Improved classroom behavior
• Increased motivation
• Increased engagement
LONG-TERM OUTCOMES
Student Outcomes
• Improved reading skills
• Increased scores on reading achievement tests
• Increased school attendance rates
• Reduction in special education referrals
Contextual effects such as the characteristics of the district, school, teacher, students, other instructional programs, and external events

Figure 1. Open Court Reading program logic model.

Logic Model for the Intervention

The logic model depicted in Figure 1 shows the theorized linkages among OCR professional development, curriculum materials, instruction, and proposed short- and long-term outcomes. The model provides a conceptual framework for the evaluation approach and informs the research design, core constructs to be measured, and timeline for the study. Figure 1 summarizes the primary constructs that are relevant for the evaluation of the OCR program and the pathways through which students' reading achievement may be influenced.

Key components of OCR (listed under Resources and Inputs in Figure 1) include teacher and student curricula and materials, as well as teacher professional development through the summer workshops and follow-up trainings. As the logic model indicates, the OCR curriculum is mediated by the way in which teachers implement it in their classrooms (see the Teaching/Learning Activities column in Figure 1). If OCR is implemented as prescribed, then the expected short-term outcomes include changes in teacher practices and student behavior and motivation. The combination of teacher and student short-term outcomes leads to improved reading skills as well as other long-term outcomes.

Overview of the Study

In this study, an independent research team evaluated the effectiveness of the OCR program at scale across a large national sample of elementary schools, with diverse school populations and conditions, and with no more support than schools would have access to if they had selected OCR as their early reading curriculum apart from participation in a research project. In this sense, the results of this study contribute to an understanding of whether OCR is effective in promoting reading proficiency in the elementary grades when implemented "at scale" with typical "real-world" levels of support.

In this article, we address two general goals:

1. To determine the extent to which the program was delivered as intended, with "fidelity of implementation".

2. To determine whether the program produced substantial impacts on student reading and how the effects of the program may have varied across subgroups.

The study design involved a multisite cluster randomized controlled trial in which 49 eligible schools within seven districts who agreed to participate in the study were randomly assigned to schoolwide training and delivery of the OCR curriculum (treatment group) or to delivery of the standard reading instruction for the school ("business-as-usual" control group). Data from teachers and students in two grade cohorts (kindergarten and Grade 3, and Grades 1 and 4) were gathered over two school years. We report findings from both the first and second years of implementation.

Research Questions

More specifically, we respond to the following research questions:

- *Fidelity of implementation*: To what extent was the intervention delivered as the curriculum developers indicated it should be implemented (specific fidelity)? Was there variation in implementation fidelity of OCR among the participating classrooms or teachers, schools, and districts? In what ways were OCR students' experiences similar to or different from those of students in the "business-as-usual" control condition (general fidelity)?
- *Overall impacts*: Does school-level assignment to the OCR curriculum intervention produce impacts on reading achievement relative to assignment to the "business-as-usual" control condition?
- *Impacts by subgroups*: Are there differential treatment impacts for OCR across critical student subgroups?

Method

The evaluation of the OCR program involved two key elements: a multisite cluster randomized trial and an implementation study. The cluster randomized trial includes 49 elementary schools, which were randomized within seven district-level blocks, to receive training and delivery of the OCR curriculum (treatment group) or to implement the standard reading curriculum for the school (control group). We recruited the districts over a three-year time frame. The research team randomly assigned the schools within each district to treatment ($n = 25$) or control ($n = 24$). We followed a Grade K–1 and a Grade 3–4 cohort across two years, assessing approximately 50 students from a minimum of two classrooms that were randomly selected in each of the designated grade levels in each school for each cohort in the fall and spring of each year (i.e., kindergarten and Grade 3 in Year 1, and Grades 1 and 4 in Year 2). The school-level random assignment design was deemed most appropriate, given that the intervention is a schoolwide program involving all grades and teachers in the school.

The OCR implementation study was a critical component to the scale-up study of the effectiveness of OCR. Its purpose was to assess the fidelity of implementation of the intervention in terms of the professional development and classroom implementation of the curriculum in treatment schools to learn more about why, how, and under what conditions the intervention was effective or ineffective. In addition, the implementation study included information about instructional practices in control classrooms to determine possible contamination, as well as to provide descriptive information about the comparative experience of teachers and students in the control condition. The implementation study includes four primary components to assess fidelity of implementation in the treatment and control condition classroom: dosage or the amount of time using the reading program, program adherence to the developer's curriculum guidelines, quality of program and instructional delivery, and student responsiveness or engagement. In addition to these measures of implementation within classrooms, we evaluated fidelity of the training and professional development components, and we examined various contextual factors at the teacher, school, and district levels that could affect program implementation.

Sampling

The study sample consisted of 49 schools within seven districts located across the United States. Participating districts represent the West, Midwest, and South and are located in urban, town, and rural geographic areas. Districts were considered eligible for participation according to the following criteria:

1. The district had not implemented and/or purchased Open Court Reading during the past three to five years.
2. The district had at least four elementary schools (kindergarten through Grade 5) with at least 44 students enrolled at each grade level. Eligible districts were recruited across a three-year time frame beginning in spring 2010.

Districts and schools committed to participating in the study for two years. During the first year of implementation, all 49 schools actively participated in the study and there was no attrition. In Year 2, one treatment school and one control school were consolidated into a new school. These two schools were considered attritors from the study and are not included in the Year 2 analyses.

The implementation study was composed of two overlapping samples at the teacher level. All teachers who taught reading (Grades K–5) in each school were targeted for training, were included in the implementation study, and administered the surveys. In addition, teachers of classrooms randomly selected for student assessments in the impact study were involved in the study of implementation involving completing classroom observations and interviews (treatment teachers only). Within each school, two teachers from each of the target grades (kindergarten and Grade 3 in Year 1, Grades 1 and 4 in Year 2) were asked to participate.

Intervention Plan

As part of the study selection criteria, schools committed to the study for two years. MHE provided each school in the treatment group with the OCR curriculum, including student and teacher materials as well as other resources, such as assessments and books for students. OCR intervention activities also included professional development training and follow-up supports throughout the school year. Teachers implementing the program in treatment schools were asked to follow lesson plans provided in the teacher manuals and regularly use OCR diagnostics and assessments to target instruction to students. Teachers in the control group schools were responsible for implementing the same reading curricula used prior to the study (instruction as usual). Teachers in the control schools reported using a variety of commonly implemented core reading programs (e.g., Scott Foresman Reading Street, Houghton Mifflin Reading, and Good Habits Great Readers).

Study Incentives

As part of an incentive package to each participating school, all teachers received free OCR program materials and professional development supports throughout the year. Control schools from five districts received a core math program (Everyday Mathematics[2]) and related professional development at no charge, whereas two districts received a cash incentive of $5,000 per school in each year of the study. In this sense, the treatment and control schools received a similar amount of resources, with treatment schools receiving additional supports for their literacy programs and control schools receiving additional supports for their core math or other educational programs. Additionally, the study provided incentives to participants. Teachers received $15 for completing a survey in the fall and spring of each year. In addition, teachers received up to $45 for completing an interview and for allowing researchers to observe their classrooms up to three times per year. A staff person from each school volunteered as a school liaison and received a $500 stipend per year for coordinating research activities at each school.

Data Collection

The data collection procedures were developed to reduce participant burden and strengthen the response rates and overall quality of the data collected to ensure rigorous and reliable results. All team members who assisted with data collection participated in extensive training and received field manuals prior to entering the field to ensure high-quality and reliable data. The current study utilized site coordinators (research team members), who interacted with a designated school liaison within each school. Site coordinators scheduled data collection visits and supported the collection of student demographics and attendance data. School visits were coordinated well in advance and through the school liaison. School liaisons provided the schools with materials to prepare them for the visits and collaborated with site coordinators to schedule interviews, observations, and assessments.

Data collection related to student assessments as well as teacher surveys and interviews was arranged to occur within a condensed time frame across all participating schools (typically two weeks per district and four to eight weeks for all schools) to reduce burden on program staff. Within each grade, two classrooms were randomly selected to complete the assessment. If the selected classrooms had fewer than 44 students, then additional students from other classrooms were randomly selected from class rosters and assessed when possible. Students were excluded from assessments if their Individualized Education Program restricted testing, they were absent, or parents opted for them not to participate.[3]

Teachers in the implementation study sample (from selected classrooms above) were videotaped during three reading or English language arts (ELA) lessons over the course of the school year. The videos were scheduled to capture reading or ELA instruction once during each season: fall (October–November), winter (December–January), and early spring (February–March). Site coordinators and school liaisons worked closely to schedule the video recordings to capture all reading or ELA instruction within a given day, avoid periods

[2] Everyday Math is a math program, so it has a very different theory of action than OCR in terms of outcomes and should not have impacted the OCR schools since it was only provided to the control schools.

[3] Based on our IRB, we were able to utilize passive parental consent. All parents within a school received an informational letter and consent form within the first two weeks of school in both treatment and counterfactual schools. Parents were able to return the signed form at any point to opt their child out of the evaluation. Over the study, we had less than 2% of the parents opt out of the study.

of school or state testing, and avoid school events. In cases where it was discovered that teachers' reading or ELA instruction had not been taped or there were technical problems, teachers were asked to reschedule and tape instruction as soon as possible.

Measures

The data collected for the study primarily relied on student assessments, teacher surveys and interviews, classroom observations, and archival records.

Group Reading Assessment and Diagnostic Evaluation (GRADE)

The primary outcome of the study was the assessment of reading outcomes using the Group Reading Assessment and Diagnostic Evaluation (GRADE; American Guidance Services, 2001). The GRADE is a standardized, norm-referenced, research-based reading assessment that can be administered to groups and takes 45–90 minutes. It is meant to be a diagnostic tool to assess the reading skills individuals have and which skills need to be taught. The GRADE is organized into 11 levels of tasks or subtests that are designed to assess reading skill development from prekindergarten through early college. The GRADE contains a separate level for each year of school from prekindergarten through sixth grade and has two equivalent forms to facilitate progress monitoring from fall to spring of each year. The GRADE measures four components of reading: reading readiness, vocabulary, reading comprehension, and oral language. We use the GRADE total test score, which is a composite of the four components, based on norms established for fall and spring for each grade level.

The GRADE has strong evidence of reliability and validity for outcomes related to literacy and reading. Technical information compiled by AGS Publishing (2001) has indicated the GRADE has a high degree of internal consistency for total, composite, and subtest scores for Grades K–5 (alphas between .95 and .99). Alternate form reliability was high (.81–.94) and test–retest reliability was also high (.80). Concurrent validity studies on Grades 1–6 indicate moderate to strong correlations between GRADE and the Iowa Test of Basic Skills (.69–.90) and strong correlations with the Gates-MacGinitie Reading Tests (.86–.90).

Teacher Surveys

Teachers of Kindergarten through Grade 5, special education, and gifted and talented who provided reading instruction in treatment and control schools were asked to complete the teacher survey in the spring. The survey was adapted from the Study of Instructional Improvement (http://www.sii.soe.umich.edu) and included teacher reports of demographic information and details about their tenure, training, and education. In addition, the surveys assessed teachers' perceptions of the professional development they received supporting their reading programs as well as the amount and types of reading instruction in their classrooms.

Teacher Interviews

Teachers in treatment schools who were included in the implementation study sample were interviewed at the end of the school year. One-on-one interviews were conducted in the teacher's school by trained interviewers on the research staff. The purpose of the semistructured interview protocol was to collect specific data from the teachers' perspectives about their experiences implementing the OCR program. The interviews were video-recorded and subsequently transcribed for analysis using NVivo qualitative software.

Classroom Observation Instrument

Videos of teachers' reading or ELA instruction captured during the course of the year were coded using an instrument developed by the research team and shared with MHE trainers and content (reading) experts. Using recommendations from Hallgren (2012), we relied on Cohen's kappa as the primary indicator of inter-rater reliability between independent and master coders, and a minimum inter-rater reliability threshold of .70 was established and maintained throughout the coding process (average 95% agreement across all items and coders). The instrument consisted of four parts: OCR materials, quality of delivery, program coverage, and a measure of student–teacher interactions. The OCR materials were a measure of the presence of essential and defining components of OCR in the classroom and consisted of six grade-specific items for kindergarten and first grade and four grade-specific items for third and fourth grade. Tallies of materials were averaged across rounds of classroom observation data. Program coverage was coded as a count of how many instructional strands of the OCR program were observed during a taped classroom session. The quality of delivery items rate teachers' OCR delivery as defined by the developer guidelines for pacing, focusing, and managing program delivery. It consists of nine items scored on a 0 (not implementing) to 4 (high implementation as intended) scale, which were averaged across items and across observations. The last section of the classroom observation measure was a 13-item scale focusing on the interaction between teachers and students during the classroom session observed (Gamse, Jacob, Horst, Boulay, & Unlu, 2008). Exploratory factor analysis (ESEM, Mplus Version 7; Muthén & Muthén, 2012) identified two subscales: student engagement (five items) and teacher effectiveness (six items). The two items that double loaded were removed, resulting in 11 items. The scores from each round were averaged to obtain teachers' overall scores for each factor.

Archival Data

Additional archival data were collected on a monthly and annual basis. Student data collected included gender, date of birth, and ethnicity, as well as special education, English language learner (ELL), and free or reduced-price lunch (FRPL) status. Teacher attendance data were collected and used as an indicator of teacher stability. Archival data were also collected to document professional development and training opportunities provided to treatment teachers and schools by the developer. These data were collected through teacher sign-in sheets at training sessions and trainer reports describing the content and attendance of ongoing professional development opportunities.

Analyses

Fidelity of Implementation

Analyses of the fidelity of implementation of the professional development components of the intervention relied on a mixed methods approach to describe the frequency and level of professional development delivered and received from teacher surveys, interviews, and archival data. The analyses addressing the classroom fidelity of implementation investigated whether there was a latent fidelity of implementation construct underlying the four components measured in

Table 1. Fidelity of implementation indicators.

Domain	Construct	Indicator	Source	Measurement
Structure	Dosage	Minutes of reading	Survey	0–59, 60–89, 90–119, 120–149, 150+ minutes
		Stability	Archival	Monthly attendance aggregated to measure presence throughout school year as yes/no categorical variable
	Adherence	Materials	Classroom observation	Percentage of materials observed in classroom; averaged across observations
		Strand coverage	Classroom observation	Count of strands covered (1 through 3); averaged across observations
Process	Quality of delivery	Program delivery	Classroom observation	Nine items measuring pacing, focusing, and management within each of 3 strands; scored on a 0 (not implementing) to 4 scale (strongly on model); averaged within and across observations
		Teacher effectiveness	Classroom observation	Five items on a 1 (Not at all true) to 4 (Very true) scale; Averaged within and across observations
	Student engagement	Student engagement	Classroom observation	Six items on a 1 (Not at all true) to 4 (Very true) scale; averaged within and across observations

the study: dosage, adherence, quality of delivery, and student responsiveness. Structural and process fidelity of implementation components were measured in both treatment and control classrooms using seven indicators. A full description of the indicators and their sources is available in Table 1. Modeling the latent fidelity construct required a technique that could handle different types (e.g., nominal, ordinal, and interval) of variables. We employed Latent Class Analysis (LCA; Hagenaars & McCutcheon, 2002; Lanza, Flaherty, & Collins, 2003) to model the latent fidelity construct across treatment and control teachers/classrooms. LCA can not only handle multiformat manifest variables, but also it empirically identifies subgroups of teachers to find commonalities or patterns in teachers' scores across structural and process fidelity components. A similar approach has been used in other studies, including one conducted by Cooper and Lanza (2014) to identify subgroups of Head Start students based on demographic characteristics. In this study, the LCA revealed patterns in teachers' general fidelity and grouped teachers with similar instructional practices accordingly.

Using Mplus (Muthén & Muthén, 2012), a series of models was run to allow teachers to be grouped into increasingly more classes until adequate statistical and substantive fit was achieved. The Bayesian Information Criterion (Schwartz, 1978) and the Lo-Mendell-Rubin test (Lo, Mendell, & Rubin, 2001) provided statistical estimates of model fit while heterogeneity across classes was estimated through entropy. Classes were added in a stepwise fashion until a minimum Bayesian Information Criterion value was achieved, Likelihood Ratio Test remained statistically significant at the $p < .05$ level, and entropy approached one (Muthén & Asparouhov, 2008; Muthén & Muthén, 2012). The classes were also evaluated for substantive fit to ensure that each additional class grouped teachers in a meaningful and interpretable manner.

Experimental Impacts

Intent-to-treat (ITT) analyses were conducted to address the two major impact research questions and their associated hypotheses: overall program impacts and program impacts for particular subgroups. The first research question involved testing the ITT effects of the intervention on student academic achievement. This study involved randomization of schools blocked by district and collection of outcome data at the level of the student. With such a design, estimation of treatment effects at the level of the cluster that was randomized is the appropriate method (Bloom, 2005; Donner & Klar, 2000; Raudenbush, 1997) and outcomes are analyzed with three-level models to account for students nested in schools and districts (e.g., Raudenbush, 1997). This simultaneously accounts for student-, school-, and district-level sources of variability in the outcomes by specifying a multilevel statistical model that estimates the school-level effect of random assignment. This model is specified as a fixed-effects model based on the assumptions of a conditional inference model (Hedges, 2007), and recent results regarding the degree of generalizability from the study sample (Tipton et al., 2016), in which there is not an intention to generalize effects beyond the current sample of purposively selected districts to a larger population.

The fully specified level 1, or within-school model, nested students within schools within the seven randomization blocks. The linear model for this level of the analysis is written as

Level 1:

$$Y_{ijk} = \pi_{0jk} + \sum_{n=1}^{x} \pi_{xjk} + \mathrm{e}_{ijk} \tag{1}$$

Equation 1 represents the spring posttest achievement for student *i* in school *j* and district *k* predicted by the school mean achievement intercept plus the student-specific level 1 residual variance, e_{ijk}. To improve the precision of the impact estimates and account for any baseline differences between treatment and control, we also include important pretreatment student demographic indicators, π_{xjkl}. At level 2 of the model, we estimated OCR treatment effects on the mean posttest achievement outcome in school *j*. As suggested by the work of Bloom, Bos, and Lee (1999) and Raudenbush (1997), the model included a school-level covariate, the applicable school mean 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 is written as

Level 2:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} T_{jk} + \beta_{02k} \mathrm{Pretest}_{jk} + r_{0jk} \tag{2.1}$$

$$\pi_{xjk} = \beta_{10k} \tag{2.2}$$

where the mean posttest intercept for the *j*th school within district *k* is regressed on the block-level mean β_{00k}, the school-level OCR treatment indicator β_{01k}, and the grand mean-centered school mean achievement score β_{02k}, which is the deviation of school *jk*'s mean from the block mean. In this formulation, the intercept is treated as a random effect, and the OCR treatment and the school mean pretest score are treated as fixed covariate effects. The parameter of central interest in the analysis is at level 2 in the model—the school-level effect of assignment to the OCR treatment β_{01k}. The fully specified level-3 model includes fixed effects for district and is written as

Level 3:

$$\beta_{00k} = y_{000} + y_{001}\,District1 + y_{002}\,District2 + y_{003}\,District3 + y_{004}\,District4 + y_{005}\,District5 + y_{006}\,District6 \quad (3.1)$$

$$\beta_{01k} = y_{010} \quad (3.2)$$

$$\beta_{02k} = y_{020} \quad (3.3)$$

$$\beta_{10k} = y_{020} \quad (3.4)$$

where the mean posttest intercept for district k, y_{000}, is modeled as a fixed effect and includes dummy variables for districts.

Subgroup Analyses to Test the Indirect Effects of the Intervention

Although the analyses presented above addressed the overall ITT effects of the intervention, it is possible that particular subgroups of students benefited more from OCR than other subgroups. For instance, the strong emphasis on phonics in early grades may benefit students to a greater extent in kindergarten and Grade 1 than older students in Grades 3 and 4. By adding a cross-level interaction term to these basic models assessing the ITT effects of the treatment on student achievement outcomes, we examined whether key student-level individual differences moderated the treatment effect. In addition to modeling these cross-level interaction effects within the hierarchical linear modeling framework, we examined more basic forms of subgroup analyses as well, for instance by performing the analyses on only those students who belong to particular subgroups. The linear model for the student level of the analysis is written as:

Level 1:

$$Y_{ijk} = \pi_{0jk} + \sum_{n=1}^{x} \pi_{xjk} + \pi_{1jk}\text{Subgroup}_{jk} + \text{e}_{ijk} \quad (4.1)$$

which represents the spring posttest achievement for student i in school j and district k predicted by the school mean achievement intercept plus the student-specific pretreatment student demographic indicators π_{xjkl}, subgroup indicators, and level-1 residual variance e_{ijk}. At level 2 of the model, we estimate OCR treatment effects on the mean posttest achievement outcome in school j. We also include the school mean pretest score to help reduce the unexplained variance in the outcome and to improve the power and precision of our treatment effect estimates. The subgroup analysis is modeled at level 2 with a cross-level interaction. We estimate OCR treatment effects on the subgroup estimate π_{1jk}. The level-2 model is written as:

Level 2:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}T_{jk} + \beta_{02k}Pretest_{jk} + r_{0jk} \quad (5.1)$$

$$\pi_{1jk} = \beta_{10k} + \beta_{01k}T_{jk} \quad (5.2)$$

These interaction effects between the treatment condition variable and particular baseline characteristics of students (e.g., grade, gender, poverty status, ethnic minority, and ELL status) can establish whether the impacts were greater or smaller in magnitude for particular student subgroups.

The fully specified level-3 model includes fixed effects for district and is written as Level 3:

$$\beta_{00k} = y_{000} + y_{001}\, District1 + y_{002}\, District2 + y_{003}\, District3 + y_{004} District4 + y_{005}\, District5 \tag{3.1}$$

$$\beta_{01k} = y_{010} \tag{3.2}$$

$$\beta_{02k} = y_{020} \tag{3.3}$$

$$\beta_{10k} = y_{020} \tag{3.4}$$

where the mean posttest intercept for district k, y_{000}, is modeled as a fixed effect.

Results

Fidelity of Implementation

Sample Description

Table 2 provides descriptive data for the analytic samples of teachers involved in the fidelity of implementation study. Panel A includes all teachers who taught reading instruction in Grades K–5 and were targeted to receive professional development. Panel B includes all consenting teachers of assessed classrooms and is the analytic sample used for analyses on general fidelity. The analytic sample for specific fidelity is described in Panel C and is a subset of Panel B including only treatment teachers. For all fidelity of implementation analytic samples, there were no statistically significant differences between the baseline demographic characteristics of teachers in treatment and control schools.

Professional Development

Professional development and support were provided to treatment schools by MHE consultant trainers. Based on discussions with MHE, we established the typical amount and type of professional development provided to districts purchasing OCR: one day of launch training and up to three follow-up visits conducted during each year of the study.

The findings indicate that all schools received an initial one-day launch training focused on introducing teachers to the research base, materials, and routines within the OCR program. However, attendance may have been less than ideal. Based on sign-in sheets from six of the seven districts, 81% of 468 teachers signed in at the training.[4] Based on administrator interviews and MHE trainer notes, the teachers were unable to attend the launch training for various reasons, including some teachers were not under contract when the training occurred, some teachers were traveling, and a subset of teachers had not yet been hired when the launch training occurred. In the second year of the study, the launch training was offered across all schools and 25% of 620 teachers attended the training. This second year of training was meant to target new teachers in addition to offering an optional refresher course to returning teachers.

[4] Sign-in sheets were not available for one of the seven districts ($n = 31$ teachers).

Table 2. Fidelity of implementation teacher sample descriptives.

	Panel A All teachers in target grades				Panel B General fidelity				Panel C Specific fidelity			
	Year 1		Year 2		Year 1		Year 2		Year 1		Year 2	
	N	Percentage	*N*	Percentage	*N*	Percentage	*N*	Percentage	*N*	Percentage	*N*	Percentage
Total	1,070		1,194		157		147		73		67	
Random assignment												
Treatment	544	51%	579	49%	77	49%	73	50%	73	100%	67	100%
Control	526	49%	615	52%	80	51%	74	50%	0	0%		0%
Sex												
Male	69	6%	76	6%	7	4%	5	3%	3	4%	3	5%
Female	1,001	94%	1,118	94%	150	96%	142	97%	70	96%	64	96%
Race/ethnicity												
White	907	85%	945	79%	147	94%	132	90%	70	96%	60	90%
Black	31	3%	28	2%	3	2%	4	3%	1	1%	3	4%
Hispanic	13	1%	15	1%	0	0%	1	1%	0	0%	0	0%
Other race/ethnicity	22	2%	29	2%	2	1%	7	5%	0	0%	3	4%
Missing	97	9%	177	15%	5	3%	3	2%	2	3%	1	1%
Advanced degree												
No	446	42%	485	41%	77	49%	76	52%	37	51%	33	49%
Yes	529	49%	530	44%	78	50%	70	48%	36	49%	34	51%
Missing	95	9%	179	15%	2	1%	1	1%	0	0%	0	0%
Grade												
K	164	15%	170	14%	84	54%			39	53%		
1	163	15%	177	15%			79	54%				
2	153	14%	164	14%								
3	135	13%	149	12%	73	46%			34	47%		
4	117	11%	122	10%			68	46%			36	54%
5	108	10%	113	9%								
Multiple	230	21%	299	25%								

Table 3. Fidelity components for overall sample and by latent class analysis membership.

	Cohort 1							Cohort 2						
	Overall		NIC Class[c]		OCR Class[d]		Sig	Overall		NIC Class[c]		OCR[d] Class		Sig
Teacher sample	157		84	54%	73	46%		147		80	54%	67	46%	
Random assignment[a]														
Treatment	77		4	5%	73	95%	***	73		6	8%	67	92%	***
Control	80		80	100%	0	0%		74		74	100%	0	0%	
Grade[a]														
K/1	84		45	54%	39	46%		79		43	54%	36	46%	
3/4	73		39	53%	34	47%		67		37	54%	31	46%	
Indicators of fidelity of implementation														
Dosage														
Minutes of reading[a]														
1–59	5	3%	1	1%	4	6%		1	1%	1	1%	0	0%	
60–89	12	8%	4	5%	8	11%		23	16%	13	16%	10	15%	
90–119	51	33%	31	37%	20	27%		65	44%	40	50%	25	37%	
120–149	46	29%	22	26%	24	33%		38	26%	19	24%	19	28%	
150 or more	15	10%	9	11%	6	8%		18	12%	6	8%	12	18%	
Missing	28	18%	17	20%	11	15%		2	1%	1	1%	1	2%	
Teacher stability[a]	150	96%	79	94%	71	97%		145	99%	79	99%	66	99%	
Adherence														
Materials[b]	35%	0.40	1%	0.03	75%	0.21	***	33%	0.37	1%	0.06	71%	0.14	***
Strands coverage[b]	0.89	1.03	0.01	0.07	1.91	0.56	***	0.81	0.95	0.01	0.08	1.76	0.53	***
Quality of delivery														
Teacher effectiveness[b]	3.88	0.20	3.88	0.21	3.89	0.18		3.99	0.03	3.99	0.04	3.98	0.04	
Program delivery[b]	1.75	1.89	0.01	0.12	3.75	0.32	***	1.81	1.96	0.02	0.15	3.92	0.17	***
Student engagement														
Student engagement[b]	3.78	0.33	3.77	0.36	3.79	0.30		3.90	0.21	3.91	0.17	3.89	0.25	

[a] Reported as frequencies.
[b] Reported as mean values.
[c] Not implementing class.
[d] Open Court Reading.
***$p < .001$.

After launch training, teachers received ongoing support through follow-up visits. The primary goals of the follow-up trainings were to expand teachers' knowledge of the program, address teacher and school-based concerns, and support the fidelity of implementation. The vast majority of treatment schools received three follow-up visits, except one school received two follow-up trainings in the first year of the study. During the second year, three schools did not receive follow-up trainings, 10 schools received two follow-up trainings, and 12 schools received three follow-up trainings.

Classroom Fidelity of Implementation

Across the two years of implementation, the general fidelity (treatment and control teachers) sample (Panel B of Table 2) included 304 teachers: 157 teachers from Year 1 and 147 teachers from Year 2. Half of the teachers were from treatment schools (49.4%). In Year 1, 84 teachers taught kindergarten and 73 teachers taught Grade 3. In Year 2, 79 teachers taught Grade 1 and 67 taught Grade 4. Across both years, most teachers were female (95%), most teachers were White (94% and 89%), and half of the teachers had an advanced degree. They reported teaching for more than 14 years on average and at their current school for more than seven years.

Descriptive Statistics

Next, we present descriptive statistics of the four structural and procedural components of classroom fidelity of implementation measured in treatment and control schools. Descriptive statistics of the implementation components for the overall sample can be found in Table 3.

Dosage. Overall, most teachers in Year 1 had high stability and were at the school all year (96%); the rates of stability were similar for teachers in Year 2 (99%). Teachers also generally reported more than an hour and a half of reading instruction every day across both Year 1 and Year 2. There were no statistical differences in dosage by grade or treatment condition.

Adherence. As expected, the presence of OCR materials observed in the classroom varied depending on random assignment. For Years 1 and 2, no OCR materials were observed for teachers in control schools. In Year 1, approximately 75% of the essential and defining OCR materials were observed for teachers in treatment classrooms, and 71% of the materials were observed for Year 2 teachers in treatment schools. There were no statistically significant differences by grade for either year.

The second indicator of adherence, program coverage, also hinged on random assignment status. Classroom observations of teachers in control schools showed no evidence of covering the instructional strands of OCR in either year. Within treatment schools, teachers covered nearly two out of three strands of the OCR curriculum (Year 1 $M = 1.91$ strands; Year 2 $M =$ 1.76 strands). There were no statistically significant differences by grade for either year.

Quality of Delivery. Teachers were rated as highly effective on both measures of quality. On the indicator of teacher effectiveness, on average, teachers scored 3.88 in Year 1 and 3.99 in Year 2 on a four-point scale. The scores were equally high across grade and treatment condition. There were differences in program delivery, however. As there was no evidence of control teachers using the OCR program, teachers in control schools received scores of 0 relative to the pacing, focusing, and managing of the OCR curriculum. Teachers in the treatment condition received an average score of 3.8 on a four-point scale for program delivery in Year 1 and Year 2, suggesting high quality for pacing, focusing, and managing the curriculum. There were no statistically significant differences by grade.

Student Responsiveness. Teachers also scored highly on the indicator of student engagement; the average score was 3.8 in Year 1 and 3.9 in Year 2 on a four-point scale. There were slight differences across grades for years 1 and 2; differences were statistically significant but not substantively large ($M = 3.71$ for kindergarten vs. 3.85 Grade 3 and $M = 3.87$ for Grade 1 vs. 3.93 for Grade 4). There were no statistically significant differences by treatment condition.

Latent Class Analysis

The second stage of the implementation study analyses combined the structural and procedural fidelity components above into a latent construct that identified teachers' patterns of implementation of the OCR curriculum across treatment and control conditions. A series of LCA models was run separately for each year that included up to five latent classes to find the best fit with the fidelity component indicators. For both years, fit indices indicated that a two-class model best fit the data. Although adding a third class did yield marginal improvement in statistical fit, the substantive interpretation was not improved. Table 3 provides the descriptive statistics for the four components of fidelity of implementation by latent class.

The first and largest class (54% of teachers in both years) consisted of teachers who showed no evidence of implementing OCR. They had virtually no OCR materials, program coverage, or quality of program delivery. We refer to this class as the not implementing class (NIC).

The second class (46%) consisted of teachers showing strong evidence of program implementation. On average, nearly two of the three OCR program strands were observed. In Year 1, 75% of program materials were observed in the classrooms and 71% of materials were observed for Year 2 teachers. Teachers in this class also scored very highly on the measure of quality of program delivery and were coded as being on model. This class is referred to as the OCR class.

Although there were striking differences across classes according to program-specific indicators, teachers in both classes were rated as equally strong in areas of teacher effectiveness and student engagement. There were also no statistically significant differences in minutes of reading instruction provided or teacher stability (dosage) across latent classes.

We also explored possible differences by grade and teacher demographics. There were no differences in class membership according to grade taught, tenure, race/ethnicity, educational attainment, or gender. Teacher demographics were equivalent across the NIC and OCR classes. In addition, there were no statistically significant differences across schools and districts in the study.

As expected, there was statistically significant overlap between treatment condition and latent class membership. All teachers in control schools were grouped into the NIC class, showing no evidence of program contamination in control schools. Among teachers in treatment schools, more than nine out of 10 teachers in both Year 1 and Year 2 were members of the OCR class. In Year 1, four out of 77 treatment teachers were classified in the NIC class. In Year 2, six out of 73 treatment teachers were classified in the NIC class. Thus, across both evaluation years, only 10 teachers (7%) from treatment schools showed no evidence of implementing the OCR program. Conversely, 93% of treatment teachers demonstrated program uptake showing relatively strong evidence that the program was implemented at least adequately across a majority of classrooms in the study.

Impacts

Sample Description

The data for these analyses come from students in all 49 schools nested in seven districts. There was no school-level attrition during the first year of the study. However, in Year 2, two schools (one treatment and one control school) were lost after a planned consolidation. These schools were considered to be attritors and no data from these schools are included in Year 2 analyses.

Table 4 provides descriptions of the overall sample and statistical tests of equivalence on baseline data. Random assignment blocked at the district level did result in equivalence across treatment and control groups for both years of evaluation. When comparing school-level data, there were no statistically significant differences between treatment and control schools or students.

Table 4. Equivalence test of school characteristics by treatment assignment.

	Year 1			Year 2		
	All Students	Treatment	Control	All Students	Treatment	Control
School characteristics						
Urbanicity						
City	37%	40%	33%	34%	38%	30%
Rural	31%	36%	25%	32%	38%	26%
Suburban	25%	16%	33%	26%	17%	35%
Town	8%	8%	8%	9%	8%	9%
Region						
Midwest	51%	52%	50%	49%	50%	48%
South	41%	40%	42%	43%	42%	44%
West	8%	8%	8%	9%	8%	9%
Title I eligible	78%	72%	83%	77%	71%	83%
Schoolwide Title I eligible	78%	72%	83%	77%	71%	83%
Full-time teachers (*M*)	30.66	30.521	30.809	31.26	31.111	31.409
Student teacher ratio (*M*)	15.75	15.784	15.72	15.86	15.847	15.864
School enrollment (*M*)	480.58	482.417	478.75	491.61	492.826	490.391
Student demographics						
FRPL eligible	59%	57%	60%	59%	58%	61%
Students in Grade 3 or 4	50%	51%	49%	50%	50%	50%
Male	51%	51%	52%	51%	50%	52%
Race/ethnicity						
Non-Hispanic White	67%	69%	66%	67%	69%	65%
Non-Hispanic Black	14%	12%	15%	14%	13%	16%
Hispanic	12%	12%	13%	11%	10%	11%
Other	7%	7%	7%	8%	8%	8%
Special education	9%	10%	8%	8%	9%	8%
ELL	11%	11%	12%	12%	11%	12%
School sample size	49	25	24	47	24	23

Note. $M =$ mean.

The study sample was also investigated for equivalence by treatment assignment at the student level. There were few significant differences in student characteristics. Within treatment schools, there was a higher proportion of White students relative to Black students and a lower proportion of students who received FRPL. Although random assignment largely resulted in equivalent groups, we will include controls for student-level FRPL and race/ethnicity to increase precision in the multilevel models. Overall student-level attrition was 12% after one year and 37% over two years with less than 1% differential attrition between conditions in Year 1, and less than 3% differential attrition in Year 2. Accrual rates of new students joining the analytic sample were similar to attrition rates, resulting in similar pretest and posttest sample sizes. This suggests that changes in the sample of tested students is unlikely to produce any bias in school-level impact estimates.

Overall Effects of Treatment Assignment on Student Reading Achievement

Data for the analysis of reading outcomes include spring reading assessments from 4,485 Year 1 students and 4,392 Year 2 students. The school level reading pretest was aggregated from baseline assessments for 4,411 Year 1 students and 4,276 Year 2 students. For both Years 1 and 2, more than 85% of students with reading outcome assessments also completed baseline assessments. There were no statistically significant differences between the baseline and outcome samples.

Table 5. OCR multilinear modeling results.

	Year 1		Year 2	
	Estimate	*SE*	Estimate	*SE*
Level 1				
Intercept	102.509***	0.696	103.21***	1.048
Race/ethnicity (White)				
Black	−4.577***	0.726	−3.639***	0.861
Hispanic	−3.903***	0.727	−4.063***	0.897
Other	−0.736	0.822	0.852	0.945
FRPL status	−3.488***	0.457	−5.431***	0.552
Level 2				
Treatment effect	0.379***	0.481	−1.584*	0.757
Pretest mean	0.639	0.061	0.835***	0.097
Level 3				
District (Derby)				
Muskogee	1.827	1.344	2.934	2.069
Nye	−2.514*	1.063	5.217**	1.633
Pike	−2.3*	0.87	1.259	1.336
Pointe Coupee	−1.463	1.156	3.502	1.777
Rapides	−1.799*	0.891	0.598	1.373
Sioux City	−1.322	0.713	1.362	1.118
ICC—School	.10		.10	
Variance (SD)				
School	0.80 (0.89)		3.82 (1.96)	
Residual	177.40 (13.32)		250.16 (15.82)	
Effect size	0.027		−0.09	
MDE	1.393		2.12	
MDE in SD units	0.096		0.13	

Notes. MDE = Minimal Detectable Effect; *SD* = standard deviation; *SE* = standard error.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

The impacts of treatment assignment were evaluated using a three-level model. The intraclass correlation (ICC) for the school level shows statistically significant within-school group variation (Year 1 ICC = .10, $p < .001$; Year 2 ICC = .10, $p < .001$). The impact model, which includes school-level Year 1 pretest, treatment assignment, and additional student-level covariates to improve precision, did not yield statistically significant treatment effects in Year 1 and a small negative impact in Year 2 (d = –.09) analyses (Table 5). The adjusted mean posttest scores indicated that reading achievement increased for students in treatment schools from baseline to spring but at a rate that was not statistically different from what was observed of students in control schools in Year 1 (Figure 2). In Year 2, reading achievement increased at a statistically significant rate for students in treatment and control schools; however, students in control schools had slightly significantly higher outcomes than students in treatment schools.

Effects of Treatment Assignment on Student Reading Achievement by Subgroups

A related research question also assessed whether the treatment may have differential impacts for different groups of students. Subgroup analyses were conducted using Year 1 baseline reading achievement, grade, gender, race/ethnicity, and FRPL (effect sizes are presented in Table 6). Year 1 results yielded statistically significant interactions for treatment with indicators of grade and race/ethnicity. Specifically, OCR had positive effects for kindergarteners (d = .12) and Hispanic students (d = .10). Kindergartners in OCR schools scored statistically significantly higher than kindergarteners in control

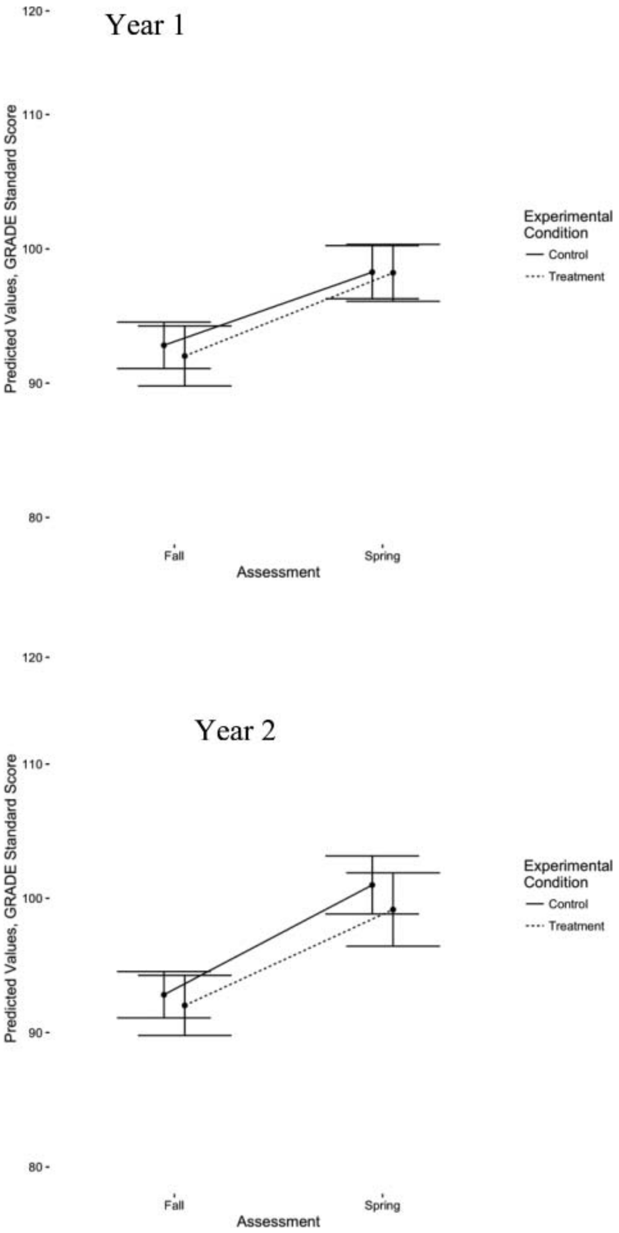

Figure 2. Adjusted mean posttest scores for Years 1 and 2.

schools. Similarly, Hispanic students in OCR schools scored statistically significantly higher than Hispanic students in control schools. Year 2 models showed a statistically significant negative effect for first grade ($d = -.13$), females ($d = -.11$), non-FRPL students ($d = -.19$), and non-ELLs ($d = -.10$).

Table 6. OCR multilinear modeling results by subgroup.

		Year 1		Year 2	
Treatment effects by subgroup		ES	p	ES	p
Grade	Low grade	0.116^{**}	.008	-0.139^{**}	.008
	High grade	−0.062	.154	−0.053	.309
Sex					
	Female	0.002	.963	-0.113^{*}	.033
	Male	0.048	.266	−0.078	.135
Race/ethnicity					
	Black	−0.021	.658	−0.091	.332
	Hispanic	0.101^{*}	.035	−0.086	.380
SES					
	Non-FRPL	−0.004	.936	-0.189^{***}	.001
	FRPL	0.022	.526	−0.026	.616
ELL					
	Non-ELL	0.013	.717	-0.096^{*}	.036
	ELL	0.056	.239	−0.074	.430
Pretest					
	Year 1 achievement	0.024	.742	−0.082	.270

Note. ES = Effect Size
$^{*}p \le .05$. $^{**}p \le .01$. $^{***}p \le .001$.

Discussion

Implementation

This study involved a large and novel effort on the part of the developers, researchers, and participants to roll out and assess the implementation and impacts of adopting a core reading curriculum at scale across almost 50 schools nationwide. Implementation study findings indicate that treatment schools and teachers received high or adequate levels of fidelity of implementation in terms of professional development, with all schools receiving launch training and up to three follow-up visits by coaches in Year 1 and most in Year 2. In addition, a majority of teachers (nearly 95%) demonstrated adequate levels of fidelity of implementation of OCR in terms of use of program materials, quality of delivery, teacher effectiveness, and student engagement.

The implementation findings comparing treatment and control schools indicate that overall the two conditions were very similar. Teacher classroom practices such as teacher effectiveness, student engagement, and reading instruction dosage were not different. There was no indication of contamination in the control schools and teachers in control schools reported using a variety of commonly used core reading curricula.

Impacts

Despite findings indicating high levels of fidelity of implementation to the OCR program and very little contamination or statistically significant differences observed between treatment and control teaching practices, the evidence from this study indicates that overall impacts on reading achievement for the OCR program are not statistically significant when implemented at scale in a large sample of schools after one year and a small and negative impact after two years relative to control schools. OCR schools experienced gains in fall-to-spring reading achievement in both years with statistically significant gains in Year 2, but these gains did not outpace the gains made in equivalent control schools; in Year 2, there was a small negative overall impact. However, the findings provide some evidence that OCR had differential positive impacts on reading outcomes for students in kindergarten and for Hispanic students relative to controls after one year. The findings also revealed a statistically significant and negative effect of OCR for first graders, females, non-FRPL students, and non-ELL students relative to controls in the second year.

Implications and Future Research

This study adds a critical piece to the evidence base for the OCR program—a high-quality third-party experimental study assessing effectiveness. This study is the first effectiveness trial of this long-standing and widely used core reading curriculum and is one of only a few in the field assessing core reading curricula. A couple of points about the differences between this study and the previous study that found significant positive impacts of the program (Borman et al., 2008) are worth pointing out. First, the individual teachers that volunteered were randomized, which helped ensure that teachers had "bought in" to the idea of implementing OCR, whereas with the school-level scale-up RCT, we typically approached recruitment at the district level, or sometimes at the school level, which is more typical of how curriculum adoption happens in schools but it does not necessarily ensure buy-in at the teacher level. Second, the previous study was done as an efficacy trial with close scrutiny by the developer to fidelity of implementation to ensure that a high level of fidelity was achieved, whereas with a scale-up study the fidelity of implementation was monitored and measured but not in order to assure high fidelity. The study impacts represent those to be expected for "typical" levels of implementation at scale and not necessarily for those to be expected under "ideal" implementation conditions. It is also possible that other programs may have "caught up" with OCR, particularly in the area of phonics but more broadly in the five areas identified as critical for reading by the National Reading Panel (2000) report. One could argue that the differences in terms of a theory of change among the core reading programs have decreased over the last 10 years and so all curriculum—whether in the treatment condition or the "business as usual" control condition—may be more alike and not address unique or substantively different content, therefore diminishing the treatment contrast and possible effects. For these reasons, replication of impacts in different contexts over a broader range of outcomes and over a longer time remain important empirical questions to pursue with this and other widely used core reading programs.

The field, including practitioners, researchers, and policymakers, should maintain or even increase the push to ensure that core reading curricula being implemented "at scale" are effective based on sound rigorous evidence in an era of standards-based reform efforts such as the Common Core State Standards Initiative. However, given that fidelity of implementation was relatively high and strong and that consistent overall impacts were not observed, future research and development efforts with this program may be optimized by focusing on efficacy trials with more targeted samples of districts and schools under ideal conditions to assess and establish the program's potential impacts. Alternatively, there may be a useful direction to consider largely from the field of health care that focuses on the role of research in sustainability efforts beyond effectiveness trials (see Chambers, Glasgow, & Stange, 2013). Specifically, these authors and others have characterized the typical problems that impact effectiveness trials, such as "voltage drop," in which interventions are expected to yield lower benefits as they move from efficacy to effectiveness trials in more real-world settings, and, "program drift," in which deviation from manualized intervention protocols in real-world settings is expected to decrease the benefit of the program. There is some indication from our study that each of these may have played a role in lower potential impacts. Chambers et al. (2013) argue that because of this there is a need to extend effectiveness work to increase and support the sustainability of programs in real-world contexts by conducting implementation science research within a dynamic continuous quality-improvement process to attempt to sustain broader impacts by focusing on critical elements to bring a program to scale successfully. Relatedly, there are calls for continuous improvement efforts in the field of education research (e.g., Bryk & Gomez, 2008) which, if coupled with robust fidelity of implementation efforts, could provide useful and quicker information about effective core components and active ingredients in programs implemented "at scale," particularly if the effort includes multiple sites as part of a network or consortium.

Measurement issues are another consideration in the context of measuring the impact of literacy programs. The research team aligned the GRADE with the key outcomes identified for the OCR program (e.g., reading comprehension, vocabulary, letter knowledge, phonological awareness, and listening comprehension) across all of the grades. While the administration of a group assessment is cost effective in large-scale studies, issues also arise due to this method of data collection, especially when testing students across multiple grades. Although the GRADE provides standardized scores for overall performance, the assessment itself is based upon expected grade-level skills and not all subtests provide standardized scores to facilitate comparisons. The GRADE, as a measure, may not be sensitive enough to the nuances of literacy skills across the grade levels assessed. In the current era of accountability in education, standardized assessments are often the primary source of evidence to support the effectiveness of instruction, programs, and success of schools and districts. Additional consideration should be given to the selection of standardized assessments, especially those that are group administered, and their sensitivity to test moderation effects of the outcomes of interest.

Future analyses by this research team will assess potential mediators of the impacts of the program, such as student motivation to read and fidelity of implementation, as well as "treatment on the treated" analyses that will provide more information about the magnitude of the impacts under ideal implementation conditions. Last, future analyses will assess generalizability using propensity score matching (see Tipton et al., 2014) to target populations for the study.

Funding

The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090150 to Southwest Educational Development Corporation, an affiliate of American Institutes for Research. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ARTICLE HISTORY

Received 13 July 2016
Revised 23 May 2017
Accepted 24 May 2017

References

American Guidance Services. (2001). *Technical manual for Group Reading Assessment and Diagnostic Evaluation*. Circle Pines, MN: Author.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches*. New York, NY: 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* (4), 445–469.

Borman, G., Dowling, N. M., & Schneck, C. (2008). A multisite cluster randomized field trial of Open Court Reading. *Educational Evaluation & Policy Analysis, 30*(4), 389–407.

Bryk, A. S., & Gomez, L. M. (2008). Ruminations on reinventing an R&D capacity for educational improvement. In F. M. Hess (Ed.), *The future of educational entrepreneurship: Possibilities of school reform* (pp. 127–162). Cambridge, MA: Harvard University Press.

Chambers, D. A., Glasgow, R. E., & Stange, K. C. (2013). Dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. *Implementation Science*, *8*, 117. Retrieved from http://www.implementationscience.com/content/8/1/117

Cooper, B. R., & Lanza, S. T. (2014). Who benefits most from Head Start? Using latent class moderation to examine differential treatment effects. *Child development*, *85*(6), 2317–2338.

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 cluster randomization trials in health research*. New York, NY: Wiley.

Donovan, S., & Cross, C. (2002). *Minority students in gifted and special education*. Washington, DC: National Academies Press.

Education Market Research. (2002). Elementary reading market—2002–2003 school year data. *The Complete K–12 Newsletter*.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First impact study final report* (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*(3), 257–288.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge, MA: Cambridge University Press.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Harris, A. J., & Sipay, E. R. (1990). *How to increase reading ability: A guide to developmental remedial methods*. New York, NY: Longman.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*(4), 437–447.

Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2003). Latent class and latent transition analysis. *Handbook of Psychology*, *2*, 663–685.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778.

Lyon, G. R., Fletcher, J. M., Fuchs, L. S., & Chhabra, V. (2006). Learning disabilities. In E. Mash & R. Barkley (Eds.), *Treatment of childhood disorders* (3rd ed., pp. 512–591). New York, NY: Guilford.

Mathes, P. G., & Denton, C. A. (2002). The prevention and identification of reading disability. *Seminars in Pediatric Neurology*, *9*(3), 185–191.

McRae, D. J. (2008). *Test score gains from Open Court Schools in California: Results from three cohorts of schools*. Columbus, OH: SRA/McGraw-Hill.

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall, CRC Press.

Muthén, B. O., & Muthén, L. K. (2012). *Mplus Version 7: User's guide*. Los Angeles, CA: Author.

National Center for Education Statistics. (2013). *A first look: 2013 Mathematics and reading (National Assessment of Educational Progress at Grades 4 and 8)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: U.S. Department of Health and Human Services/National Institutes for Health.

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

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Skindrud, K., & Gersten, R. (2006). An evaluation of two contrasting approaches for improving reading achievement in a large urban district. *The Elementary School Journal*, *106*(5), 389–408.

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford Press.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Ruiz de Castilla, V. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal on Educational Effectiveness, 9*(S1), 209–228.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114–135.

What Works Clearinghouse. (2007, August). *What Works Clearinghouse topic report: Beginning reading*. Washington, DC: Author.

What Works Clearinghouse. (2014, October). *Beginning Reading intervention report: Open Court Reading©*. Washington, DC: Author.

Williams, T., Kirst, M., & Haertel, E. (2005). *Similar English learner students, different results: Why do some schools do better? A large-scale survey of California elementary schools serving low-income students*. Mountain View, CA: EdSource.