Educational Evaluation and Policy Analysis
September 2018, Vol. 40, No. 3, pp. 316–335
DOI: 10.3102/0162373718764828

The Impacts of Reading Recovery at Scale: Results From the 4-Year i3 External Evaluation

Philip Sirinides
Abigail Gray
University of Pennsylvania
Henry May
University of Delaware

Reading Recovery is an example of a widely used early literacy intervention for struggling first-grade readers, with a research base demonstrating evidence of impact. With funding from the U.S. Department of Education's i3 program, researchers conducted a 4-year evaluation of the national scale-up of Reading Recovery. The evaluation included an implementation study and a multisite randomized controlled trial with 6,888 participating students in 1,222 schools. The goal of this study was to understand whether the impacts identified in prior rigorous studies of Reading Recovery could be replicated in the context of a national scale-up. The findings of this study reaffirm prior evidence of Reading Recovery's immediate impacts on student literacy and support the feasibility of successfully scaling up an effective intervention.

Keywords: *evaluation, experimental design, hierarchical linear modeling, literacy, reading, rural education, survey research*

The need for interventions that can help reverse struggling readers' trajectories of low literacy in the early years of school is widely recognized and supported by decades of research. Students who read below grade level at the end of third grade are up to 4 times more likely than their typically achieving peers to drop out of school (Balfanz, Bridgeland, Bruce, & Fox, 2012; Hernandez, 2011), and those who read below grade level after first grade are unlikely to ever catch up (Juel, 1988; Lyon et al., 2001; Shaywitz et al., 1999). Research increasingly demonstrates that low literacy levels in the early grades are associated with decreased rates of high school graduation and increased likelihood of lifelong low literacy (Annie E. Casey Foundation, 2010; Hernandez, 2011).

Producing research evidence of interventions' effectiveness is, therefore, an important goal. However, given that two thirds of all U.S. fourth graders are now reading below grade level (Annie E. Casey Foundation, 2016), establishing effectiveness cannot be the only goal: Research must also produce evidence of the *scalability* of those programs with proven impacts on early literacy. Whether interventions that produce impacts in small or localized trials can show similar effects when expanded on a national scale is a critical policy question.

The Reading Recovery Intervention

Reading Recovery is an example of a widely used early literacy intervention with a research base demonstrating evidence of impact. The program was developed in the 1970s and 1980s by Marie Clay, a developmental psychologist and professor at the University of Auckland, whose theories about early literacy development provide the foundation for the intervention's approach (Clay, 1987, 1991, 2005a). Reading Recovery provides struggling first-grade readers

with intensive, targeted instruction aimed at accelerating their learning progress and bringing them to a level comparable with that of their higher achieving peers within a period of 12 to 20 weeks. Its goal is to address reading difficulties early, before they can permanently affect students' academic trajectories (Lyons, 1998).

Reading Recovery is a supplemental pull-out intervention provided during the regular school day. The program model specifies that the intervention may be provided at any time other than during regular classroom literacy instruction; students receiving the intervention participate fully in regular classroom literacy instruction as well. The program consists of daily, 30-minute, one-to-one instructional sessions delivered by highly trained teachers. Reading Recovery teachers participate in a yearlong graduate course taught by a teacher leader who is generally an experienced Reading Recovery teacher herself and who is overseen by a Reading Recovery trainer based at 1 of the 19 regional university-based training centers. Teachers in training receive on-site coaching and support from their teacher leaders, who also facilitate "behind-the-glass" training sessions, in which observers offer feedback on Reading Recovery lessons taught in real time behind a two-way mirror. After the initial training year, practicing Reading Recovery teachers continue to receive coaching from their teacher leaders and to participate in behind-the-glass sessions periodically.

During Reading Recovery's one-to-one instructional sessions, teachers carefully observe students' literacy behaviors and identify specific learning needs. Instruction is tailored and continually refined in response to students' progress. A key goal of the Reading Recovery intervention is to equip students with literacy strategies that can be applied in the regular classroom, allowing them to continue developing as readers after the intervention has ended. Students who successfully reach a level of reading proficiency, that is, approximately equal to that of their first-grade peers in the course of Reading Recovery session are said to have successfully "discontinued." Students who fall short of reaching grade-level proficiency may exit the program without discontinuing; in this case, Reading Recovery teachers are able to provide schools with detailed information about the student's progress and needs to help specify appropriate future supports (Reading Recovery Council of North America, 2012).

Typically, a student's Reading Recovery intervention period lasts between 12 and 20 weeks, depending on the student's pace of progress toward the established text reading level for discontinuation. Discontinuation targets are established based on average first-grade reading levels in a given school. Students might reach the target level and, therefore, complete the program successfully at any time within the 12- to 20-week intervention window, at which point their lessons would be concluded. Once a student reaches the target text level, her readiness to terminate Reading Recovery lessons is confirmed via the *Observation Survey of Early Literacy Achievement* (OS), an individually administered, multidimensional assessment developed for use with Reading Recovery (Clay, 2002, 2005b, 2016). Those who progress more slowly might receive the full 20 weeks of lessons. Reading Recovery's program model requires that lessons are terminated after 20 weeks, whether or not a student reaches the target text level.

Reading Recovery lessons follow a prescribed structure, within which the teacher exercises discretion in selecting specific instructional activities. Each lesson begins with the student rereading familiar books while the teacher documents the student's errors using a running record. Next, the teacher leads the student through an activity targeting skill-development needs in letter or word recognition. Third, the teacher guides the student through the composition of a brief story. Fourth, the student works to assemble a cut-up, previously composed story by correctly ordering the words and punctuation. Finally, the teacher introduces a new book, which the student practices reading. The student is expected to practice reading the new book at home.

While facilitating these activities, Reading Recovery teachers record detailed observations about the students' development and errors and use this information to plan individualized instruction in specific strategies designed to support their progress. Phonemic awareness and phonics skills are targeted throughout the lessons via instructional activities targeted to individual students' specific developmental needs (Clay, 2005a).

Prior Research on Reading Recovery's Impacts

Reading Recovery has been widely studied over its more than 30-year history in the United States, and a considerable volume of research examines its impacts on student achievement (Allington, 2005; Ashdown & Simic, 2000; Bates, D'Agostino, Gambrell, & Xu, 2016; Center, Wheldall, Freeman, Outhred, & McNaught, 1995; D'Agostino & Murphy, 2004; Holliman & Hurry, 2013; Pinnell, 1989; Pinnell, Lyons, DeFord, Bryk, & Seltzer, 1994; Quay, Steele, Johnson, & Hortman, 2001; Rodgers, Gómez-Bellengé, Wang, & Schulz, 2005; Rodgers, Wang, & Gómez-Bellengé, 2004; Schwartz, 2005; Slavin, Lake, Davis, & Madden, 2011). However, the vast majority of prior studies of the intervention's impacts do not support causal inference (U.S. Department of Education, Institute of Education Sciences, 2013). By and large, prior studies on Reading Recovery used nonexperimental designs with no comparison group, or quasi-experimental designs in which the treatment and control groups were not equivalent at baseline. In a review of 126 studies of Reading Recovery published shortly prior to the current study, only three studies of Reading Recovery's impacts are identified as meeting What Works Clearinghouse (WWC) evidence standards without reservations (U.S. Department of Education, Institute of Education Sciences, 2013).[1] Our own review of the literature leads us to echo other scholars' observation that, although there has been a great deal of evaluative research focusing on Reading Recovery, the vast majority of this literature is lacking in methodological rigor (D'Agostino & Murphy, 2004).

The three previous impact studies that met WWC evidence standards (Pinnell, DeFord, & Lyons, 1988; Pinnell et al., 1994; Schwartz, 2005) are all randomized control trials with small samples (*n*s of 91, 79, and 74, respectively). They report variable but positive impacts on each of the following domains of early literacy development: alphabetics, comprehension, reading fluency, and general reading achievement. At least one statistically significant impact is reported in each outcome domain across the three studies, and effect sizes are generally medium to large: For instance, Pinnell et al. (1994) reported Cohen's *d* effect sizes of .49 and .51 standard deviations on two standardized reading assessments, the Woodcock-R and Gates-MacGinitie. Schwartz (2005) reported Cohen's *d* effect of .94 standard deviations on the Slosson Oral Reading Test and effects ranging from .90 to more than 2.0 standard deviations on various subtests of the Observation Survey.

Cost is an important consideration for any instructional intervention and is of particular interest for intensive programs, such as Reading Recovery, that require specialized staff training and/or serve relatively few students at a time. Numerous studies have sought to quantify the costs of Reading Recovery (Hiebert, 1994; Reynolds & Wheldall, 2007; Shanahan & Barr, 1995), yielding overall cost estimates that generally range from about US$2,000 to nearly US$5,000 per student served (Shanahan & Barr, 1995). One recent and detailed study computed an estimated cost-effectiveness ratio for Reading Recovery and compared it to that of another first-grade literacy intervention, Fast ForWord Reading. This comparison revealed the cost per unit increase in effect size on alphabetics for Reading Recovery (US$1,480) to be nearly twice that of the other program (US$601). However, the authors join other scholars in observing that comparisons of the cost-effectiveness of early literacy programs are challenging because of the many dissimilarities in programs' structures and varied methods of identifying and measuring reading-related outcomes (Hollands et al., 2016). For instance, they observe, that the comparison between the two first-grade interventions does not account for the fact that Reading Recovery has shown effects in three outcome domains (alphabetics, comprehension, and fluency), and Fast ForWord Reading only one, or for the fact that Reading Recovery covers twice as many weeks as Fast ForWord Reading.

No rigorous research to date examines the long-term effects of Reading Recovery on student literacy achievement. Although a quasi-experimental follow-up study is currently underway, long-term impacts are not the focus of the current study. Rather, this study seeks to examine whether the immediate impacts demonstrated in prior research are maintained in a much larger study and in the context of a national scale-up.

TABLE 1
i3 Scale-Up Goals and Accomplishments

Goal	Scale-up goal	Actual total	% of goal met
Reading Recovery teachers trained	3,675	3,747	102
Teacher leaders trained	15	46	306
Students served with one-to-one Reading Recovery lessons	67,264	62,000	92
Other students served by Reading Recovery teachers	302,688	325,500	108

Source. D'Agostino and Rodgers (2015).

Scaling Up an Intervention With Evidence of Effectiveness

In 2010, the U.S. Department of Education's Office of Innovation and Improvement funded the expansion and evaluation of four established programs—Reading Recovery among them—through its Scaling Up What Works grant program. The Ohio State University (OSU)—the seat of Reading Recovery in the United States—received US$45 million in federal funds and US$10.1 million from private sources to expand Reading Recovery nationally over a period of 5 years. The goal of the scale-up was to train and support new Reading Recovery teachers across the country to expand the number of students served with the intervention. Table 1 provides a summary of the goals and accomplishments of the i3 scale-up.

A majority of schools recruited under i3 were located in the Midwest and northeastern United States with a plurality located in rural and town locales. Most schools—roughly 75% of those that participated in the scale-up—used the grant to expand an existing Reading Recovery implementation by training additional teachers and serving more students. In the remaining i3 schools, the grant was used to implement Reading Recovery for the first time.

The Current Study: Confirming Impacts at Scale

Along with funds to support the expansion, the i3 grant included US$4 million for the evaluation discussed here: a rigorous, 4-year, independent evaluation of Reading Recovery's success in scale-up and implementation goals, as well as the program's impacts on the students who received Reading Recovery lessons from teachers trained with i3 funds. The independent evaluation of Reading Recovery under the i3 scale-up also included a mixed-methods analysis of the program's national scale-up and implementation in the i3 schools and was implemented from 2011 to 2015 by the Consortium for Policy Research in Education (CPRE), at the University of Pennsylvania in collaboration with the Center for Research in Education and Social Policy (CRESP) at the University of Delaware (May et al., 2014; May et al., 2013; May et al., 2015; May, Sirinides, Gray, & Goldsworthy, 2016).

The impact study was designed to address the following primary research question:

> **Research Question 1:** What is the immediate impact of Reading Recovery on the reading achievement of struggling first-grade readers, as compared with business-as-usual literacy instruction and supplemental supports?

In addressing this question, the aim of this study was to offer the strongest evidence to date of the impact of Reading Recovery on the literacy achievement of struggling first graders, and of the feasibility of producing significant impacts in the context of a large, nationwide scale-up.

Method

The immediate impacts of Reading Recovery on the reading achievement of first-grade students in the i3 scale-up were investigated through a multisite randomized controlled trial (MS-RCT). The study was conducted from the 2011–2012 school year through 2014–2015.

TABLE 2

School Year 2012–2013 Attribute Data for Schools in the MS-RCT and i3 Scale-Up

	i3 MS-RCT schools	All i3 schools	All U.S. schools with Grade 1 students
Total schools (*N*)	1,321	2,607	52,739
School population			
Total student enrollment	482.0	482.2	459.7
Total full-time-equivalent classroom teachers	32.8	32.1	28.5
Grade 1 student enrollment	82.2	81.9	73.4
Percentage of students receiving lunch assistance	54.8	56.0	55.7
Student race/ethnicity (%)			
American Indian/Alaskan Native	0.9	0.7	1.0
Asian	5.1	5.4	4.6
Blank, non-Hispanic	14.6	16.7	15.2
Hispanic	20.3	20.2	27.1
Hawaiian Native/Pacific Islander	0.2	0.2	0.4
White, non-Hispanic	55.1	52.8	48.1
Two or more races	3.8	3.9	3.5
School locale (%)			
Urban	28.3	32.6	29.8
Suburban/town	45.2	41.9	44.3
Rural	26.5	25.5	25.9

Note. A total of 2,424 i3 schools could be matched to 2012–2013 NCES records. MS-RCT = multisite randomized controlled trial; NCES = National Center for Education Statistics.

Participants

A total of 9,784 students were identified to participate in the MS-RCT. Over the 4 years of the study, a total of 1,490 schools were randomly selected from among the population of schools participating in the i3 scale-up to randomize students for the MS-RCT. Of these, 1,254 schools (84%) participated in the study. To better understand whether random selection produced a representative sample of schools, and the extent to which i3 schools were similar to other U.S. elementary schools, NCES school-level data were obtained for the 2012–2013 school year; 93% of i3 schools from all 4 years of the evaluation could be matched. Table 2 provides a summary of the characteristics of schools in the MS-RCT compared with all schools in the i3 scale-up, and to all U.S. schools with Grade 1 students.

Compared with all U.S. schools with Grade 1 students, the population of i3 schools was similar in the percentage of students receiving free or reduced lunch assistance and school urbanicity. Schools in the i3 scale-up were slightly larger in membership and with a lower percentage of Hispanic and Black, non-Hispanic students, a trend that holds for the sample of schools in the MS-RCT study. Students in schools participating in the MS-RCT were identified to be part of the experiment at the beginning of each of the study's 4 school years via the following screening process: Reading Recovery teachers first screened a pool of candidates for Reading Recovery intervention using the OS. Candidates included first-grade students who were identified by school staff—including kindergarten, first grade, and intervention teachers—as struggling readers. The eight students with the lowest OS scores in a given school were then selected to participate in the MS-RCT. This screening and identification process is consistent with Reading Recovery's normal procedures.

TABLE 3

Sample Size and Retention Rate

	Treatment	Control	Total
Students randomly assigned	4,892	4,892	9,784
Students dropped from the analysis	1,448	1,448	2,896
Due to own missing assessment test data	756	1,173	1,929
Due to missing data on matched student	692	275	967
Students included in the analytic sample	3,444	3,444	6,888
Sample retention rate	70.4%	70.4%	70.4%

TABLE 4

Analytic Sample Characteristics by Subgroup

Variable	Randomized sample	Analytic sample	Percent attrition
English language learners			
Treatment	898	664	26.1
Control	861	639	25.8
All	1,759	1,303	25.9
Students in rural schools			
Treatment	1,720	1,367	20.5
Control	1,720	1,367	20.5
All	3,440	2,734	20.5
All students			
Treatment	4,892	3,444	29.6
Control	4,892	3,444	29.6
All	9,784	6,888	29.6

Note. Attrition for ITBS total, reading words, and comprehension impact analyses. ITBS = Iowa Tests of Basic Skills.

Of the 1,490 schools randomly selected to contribute students for the MS-RCT, 236 either failed to implement random assignment of students or dropped out of the i3 project altogether. We performed sensitivity checks to determine whether the schools attended by randomized students were significantly different from those attended by nonrandomized students. We observed no differences between the schools attended by randomized versus nonrandomized students based on factors such as rural classification ($p = .60$), Title I status ($p = .63$), and the average percentage of students in a school who are identified as in need of Reading Recovery ($p = .13$). We, therefore, found no evidence from available data that students who were not randomized attended schools that were observably different from those attended by randomized students. In addition, we surveyed Reading Recovery teachers at the 236 schools that were selected to randomize students but did not do so about the reasons why their schools did not randomize. In a majority of cases, schools' failure to randomize was a result of misunderstandings about when and how to participate in randomization. In five instances, the Reading Recovery program was temporarily suspended due to extenuating circumstances such as Reading Recovery teacher maternity or medical leave. The results of both of these efforts provide no evidence that the impact estimates yielded by the MS-RCT are not generalizable to the population of students in i3 schools.

A total of 4,892 students were randomized to treatment and 4,892 to control. These students were matched into pairs, within school, according to a process detailed below (see "Procedures" section). Both pretest and posttest data were available for 7,855 of these students (4,136 treatment and 3,719 control). Pairs in which either student was missing assessment data were dropped from the study, leaving a total of 6,888 students (3,444 matched pairs in 1,122 schools). These students comprise the analytic sample, which represents 70% of the students who were randomized to treatment and control. Because an entire pair was dropped in the event that one student in a pair was missing outcome data, there is no differential attrition.

Because of the high penetration of Reading Recovery in rural settings, and because of the interest in effective literacy interventions for nonnative English-speaking students, the evaluation of Reading Recovery includes a specific focus on rural schools and students who are English language learners (ELL). Tables 3 and 4 provide additional detail on the analytic sample, as well as subsamples for ELL students and students in rural schools for the 4 years of the study.

We performed statistical tests of differences in student demographics for students included in the analytic sample ($n = 6,888$) and those dropped

TABLE 5

Teacher Experience in Year of Participation in RCT

	First year as a reading recovery teacher (%)	2 or more years as reading recovery teacher (%)	Total (%)
Never elementary teacher	3.1	1.3	4.4
1 year prior teaching	1.8	1.0	2.8
2–5 prior years	13.5	5.1	18.6
6–9 prior years	13.1	5.7	18.8
10+ prior years teaching	41.4	13.9	55.4
Total	73.0	27.0	100.0

Note. Percentages based on survey responses from 2,345 teachers. RCT = randomized controlled trial.

due to their own or their matched-pair partners' incomplete data (n = 2,896). Chi-square tests of independence in student characteristics for those students who were included and excluded from the analytic sample suggest no significant differences in pretest OS text reading levels (p = .54), sex (p = .80), or ELL status (p = .21). Students who were dropped from the analysis were disproportionately non-White (35% vs. 43%) and the difference was significant (p < .001).

Procedures

Once students were identified for inclusion in the study sample, they were randomly assigned to experimental condition via an online random assignment tool. The tool matched students into pairs, within school, by first matching any students with ELL designations. Next, the tool matched the students with the two lowest scores on the Text Reading Level (TRL) subscale of the OS, the students with the next two lowest scores, and so on.[2] The blocking of students in matched pairs produced, essentially, four mini-experiments per school. A randomizing algorithm then assigned one student in each matched pair to the treatment group and the other to control. Assignments could not be altered after randomization.

Students who were randomly assigned to the treatment group began receiving Reading Recovery lessons as soon as possible after assignment—typically in September or October of first grade. These students were expected to receive a typical Reading Recovery intervention, which consists of one-to-one lessons for 30 minutes per day for 12 to 20 weeks as a supplement to regular classroom literacy instruction. Table 5 presents the number of years of experience that i3 Reading Recovery teachers had both as a teacher in the program and as a general elementary teacher prior to becoming a Reading Recovery teacher. Schools that participated in the i3 scale-up did not necessarily contribute to the MS-RCT in the first year of training new teachers under the grant; thus, the data reported in Table 5 reflect level of teacher experience in the year of MS-RCT participation and not necessarily in the first year of i3 participation.

Students in the control condition received regular classroom literacy instruction and also had access to any literacy supports that were normally provided to low-achieving first-grade readers by their schools, other than Reading Recovery.

The study was designed as a delayed-treatment randomized trial. Treatment students received the Reading Recovery intervention during the first 12 to 20 weeks of the school year. When the treatment student in each pair completed the intervention, both students in the pair were assessed. After this point, the matched pair's participation in the experiment was complete and the control student began receiving Reading Recovery lessons.

Measures

The MS-RCT estimated the impact of Reading Recovery on students' reading achievement using the Iowa Tests of Basic Skills (ITBS). The confirmatory analysis used the Total Reading standard score from the ITBS. Exploratory objectives include the estimation of impacts on ITBS

subtests (reading words and reading comprehension) and Total OS scores. The ITBS is a group-administered, norm- and criterion-referenced, standardized assessment designed to "assess the extent to which a child is cognitively ready to begin work in the academic aspects of the curriculum" (Hoover et al., 1994, as cited in Tang & Gómez-Bellengé, 2007), and to "measure growth in fundamental areas of school achievement" (Hoover et al., 2003, p. 1). The ITBS technical manual provides sound evidence to support the instruments' content validity and high discriminant ability and contains multiple reliability coefficients (internal consistency, equivalent forms, test–retest), most of which range between the mid .80s to low .90s (Hoover et al., 2003).

An additional exploratory analysis examined Reading Recovery's impact on the OS, which is the primary screening, diagnostic and monitoring instrument for Reading Recovery. It is a one-to-one, teacher-administered, standardized assessment that includes six subscales: Letter Identification, Concepts About Print, Ohio Word Test, Writing Vocabulary, Hearing and Recording Sounds in Words, and Text Reading Level (TRL). Reported test–retest and internal consistency reliability estimates range from moderate to high on the individual OS subscales (Clay, 2002, as cited in Denton, Ciancio, & Fletcher, 2006); measures of the interassessor reliability of the Text Reading and Writing Vocabulary tasks yielded coefficients of .92 and .87 (Denton et al., 2006). Across several studies that assess the construct and criterion validity of the OS subscales, researchers have found that scores can be validly interpreted for measurement of early reading constructs (Gómez-Bellengé, Gibson, Tang, Doyle, & Kelly, 2007; Tang & Gómez-Bellengé, 2007); prediction of the attainment of performance benchmarks (Denton et al., 2006); and identification of at-risk students (Rodgers et al., 2005). Total OS scores were used as both baseline and outcome measures in the exploratory analysis of the intervention's impacts on the OS. The TRL subscale of the OS provides a single, consolidated measure of students' reading abilities by ascertaining the level at which a student can read with at least 90% accuracy. TRL was used to match students during the random assignment process, as the baseline measure, and as the pretest covariate in the statistical models of impacts on the ITBS.

The full OS measure, including all six subscales, was administered to all students at baseline by a Reading Recovery teacher working in the school. The OS was administered again immediately postintervention, along with the ITBS. All outcome data analyzed for this study were collected by highly trained teachers who can be presumed to be reliable administrators of standardized instruments with standardized instructions. As an additional safeguard, students were never tested by their own Reading Recovery teachers. Rather, posttests were always administered by another Reading Recovery teacher at the school or by a teacher leader. This is standard practice in Reading Recovery and helps ensure the validity of student scores.

Training in administration of standardized assessments is part of the Reading Recovery training.

Analyses

Analyses performed for this study included (a) test of equivalence between treatment and control groups at baseline; (b) estimation of main impacts on the confirmatory outcome, exploratory measures, and subgroup effects; (c) assessment of implementation fidelity; and (d) description of the control condition and extent of control contamination. The methods for each analysis are described briefly here.

Analysis of Baseline Balance

Baseline balance was assessed for the analytic sample to confirm the effectiveness of the randomization process in creating equivalent experimental groups on observed characteristics including sex, ELL status, race, and prior reading performance

Analysis of Impacts

Impacts on student reading performance were estimated in all analyses by comparing immediate postintervention reading achievement scores of students randomly assigned to participate in Reading Recovery at the beginning of first grade to students randomly assigned to the control condition. We used a three-level hierarchical linear model (HLM; Raudenbush & Bryk, 2002) with

TABLE 6

Baseline Balance Tests for Student Demographics

Pretreatment variable	Treatment group	Control group	*p* value for difference
Sex (*n* = 6,867)			
Male	60%	61%	.18
Female	40%	39%	
ELL status (*n* = 6,851)			
ELL	19%	19%	.47
Non-ELL	81%	81%	
Race (*n* = 6,820)			
Black	12%	13%	.06
Hispanic	20%	19%	
White	42%	44%	
Other	26%	24%	
TRL score (*n* = 6,888)			.56
M	1.04	1.02	
SD	(1.36)	(1.28)	

Note. Some student demographic data were unavailable. *p* values for sex, ELL status, and race based on χ^2 test of independence. *p* value for TRL based on independent *t* test. ELL = English language learners; TRL = text reading level.

students nested within matched pairs, and matched pairs nested within schools. The multisite, matched-pairs design of this random assignment study means that each school and each pair is an independent mini-experiment, and that the analytic sample has no differential attrition through student and school nonparticipation. Please refer to Tables 2, 4, and 6 for analyses of the representativeness of schools and students in the final analytic sample and the equivalence of the two experimental groups at baseline.

Differences in the posttest performance of the treatment and control students were estimated after controlling for pretest performance. For the confirmatory analysis of Reading Recovery's impacts on the ITBS Reading Total scores for the full sample, as well exploratory analyses of ELL and rural subgroup impacts, and for impacts on ITBS subtests, this HLM included TRL scores as a covariate. For exploratory analyses of impacts on the OS, the OS Total score was used as a covariate. All models also included a binary indicator of treatment condition, a four-category fixed effect for year, an interaction effect for treatment by year, a random effect for matched pair, a random effect for overall school performance (i.e., school intercepts), and a school-level random effect for the impact of Reading Recovery (i.e., school-specific treatment effects). A completely general (unstructured) covariance matrix was used, which included a correlation between random effects for school-level intercept and slope. In addition, a grouped residual variance was included to account for differences in dispersion of outcome scores within the treatment versus control groups. Models were estimated using PROC MIXED in SAS 9.3 via Restricted Maximum Likelihood (REML), with model-based standard errors and degrees of freedom based on within- and between-cluster sample sizes.

Equation 1: Statistical model for the Multisite Randomized Control Trial

The mathematical form of the primary impact model for the MS-RCT study is

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}\left(Pretest_i\right) + \beta_{20k}\left(Trt_i\right) \\ &+ \sum_{m=3}^{5}\beta_{m0k}\left(Year_i\right) + \sum_{n=6}^{8}\beta_{n0k}\left(\mathrm{Trt}_i \times Year_i\right) \\ &+ \gamma_j + \alpha_k + \varphi_k\left(\mathrm{Trt}_i\right) + \varepsilon_{ijk}, \end{aligned}$$

with

$$\begin{aligned} &\gamma_j \sim N\left(0, \omega^2\right), \\ &\begin{pmatrix}\alpha_k \\ \varphi_k\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\tau^2 & \rho\left(\tau\times\xi\right) \\ \rho\left(\tau\times\xi\right) & \xi^2\end{pmatrix}\right), \\ &\text{and } \varepsilon_{ijk} \sim N\left(0, \sigma^2_{T,C}\right), \end{aligned}$$

where Y_{ijk} is the posttest outcome score for student *i* from pair *j* in school *k*; β_{0jk} is the model intercept;, β_{1jk} is the slope coefficient for the pretest covariate; β_{20k} is the main effect for treatment; Trt is the treatment assignment indicator, with 1 = *treatment* and 0 = *control*; $\beta_{\{3,4,5\}0k}$ are the main effects for year, with school year (SY) 2014–2015 as reference; *Year* is the year indicator; $\beta_{\{6,7,8\}0k}$ are the interaction effects for year and treatment; γ_j is the random intercept associated with matched pair *j*, with variance ω^2; α_k is the random intercept associated with school *k*, with variance τ^2; φ_k is the random treatment effect associated with school *k*, with variance ξ^2; ρ is the correlation between random school intercepts and treatment effects; and ε_{ijk} is the

student-level residual, with a separate variance estimate, $\sigma^2_{T,C}$, for treatment versus control groups.

The parameter estimates from the HLM model were used to calculate overall mean differences in outcomes between treatment and control groups after controlling for baseline scores. These model-based average student scores (i.e., least squares means) on the outcome measures were estimated for treatment and control groups pooled across all years, schools, and pairs, along with the group contrasts and associated standard errors. The results reflect group mean test score differences that can be converted to standardized effect sizes and benchmarked relative to effect sizes typically found in evaluations of reading interventions.

The impact estimates were then standardized using the standard deviation of the outcome for the control group to produce Glass's Δ. We also calculated a population-based Cohen's *d* standardized effect size, which is calculated by dividing the estimate of treatment impact by the standard deviation of the outcome measure for the national norming sample. This allowed the impact of Reading Recovery to be benchmarked against the full population of first-grade students, not just the struggling readers in the study sample.

Analysis of Implementation Fidelity

Despite broad consensus about the importance of assessing implementation fidelity for ensuring the validity and interpretability of RCT findings, there is inconsistency in the education literature about how fidelity is best defined and measured (Century, Cassata, Rudnick, & Freeman, 2012; Century, Rudnick, & Freeman, 2010; Dhillon, Darrow, & Meyers, 2015; Mowbry, Holter, Teague, & Bybee, 2003). As a result, it often falls to the evaluator to grapple with questions about how fidelity should be understood and assessed in the context of a given intervention (Century, Rudnick, & Freeman, 2010; Goodson, Price, & Darrow, 2015; Summerfelt, 2003). Our understanding of implementation fidelity in the context of Reading Recovery is informed by the *Standards and Guidelines of Reading Recovery in the United States, 6th ed.* (*Standards and Guidelines*; Reading Recovery Council of North America, 2012). The *Standards and Guidelines* codify Clay's understandings about the activities and practices that constitute adherent implementation of the Reading Recovery program (Clay, 2005a). More specifically, we focused our fidelity study on the 51 program standards that are related to the core activities of Reading Recovery: the training of teachers and the delivery of one-to-one lessons.

Consistent with recent literature on implementation fidelity (Dhillon et al., 2015; Goodson et al., 2015; Laurentson, Oh, & LaBlanca, 2015), our approach to measuring fidelity of implementation in the i3 scale-up of Reading Recovery involved a five-step process:

1. Operationalize each of the relevant program standards as measurable program indicators.
2. Construct a logic model that defines the core activities of Reading Recovery by grouping program indicators into four key components:

- **Staff background and selection:** This component includes standards that specify the selection criteria for teachers trained in Reading Recovery and teacher leaders.
- **Teacher leader and site capacity:** This component includes standards that specify the training experience of teacher leaders as well as the standards that characterize the training environment.
- **Reading recovery teacher training and ongoing professional development:** This component includes standards that specify the training and continuous professional development experience of trained and in-training Reading Recovery teachers.
- **One-to-one reading recovery lessons:** This component includes standards that specify the selection, assessment, and instruction of individual Reading Recovery students.

These key components enabled us to represent, as concisely as possible, the complex set of required activities reflected in the *Standards and Guidelines*.

3. Define minimum thresholds for adequate implementation. The minimum threshold established for fidelity indicator adequacy was 80%, meaning that among respondents for whom the indicator represented an applicable standard, 80% reported faithful implementation of a given indicator. The minimum threshold established for key component fidelity was also 80%, meaning that no fewer than 80% of the number of the fidelity indicators within a given key component were adequately implemented.
4. Collect data on each program indicator directly from implementers. Data were collected via survey instruments that asked teachers and teacher leaders involved in the i3 scale-up to report on specific activities reflected in the key components.
5. Measure adherence to the program indicators and assess adequacy of implementation for each; calculate number of program indicators with adequate implementation within each key component to determine fidelity of implementation.

Steps 4 and 5 were repeated in each year of the evaluation, and findings are reported separately for each year.

Analyses of the Control Condition and Control Contamination

In this experiment, we collected data to determine the extent to which control students received other supplemental intervention services (non–Reading Recovery) in addition to classroom instruction. This is important because this study compares Reading Recovery plus classroom instruction and other supplemental supports with classroom instruction *plus* any other interventions or supports the schools normally provide to struggling first-grade students. We surveyed their first-grade teachers in the second and third years of the study and asked them to report, individually for each control group student in their classroom, what supplemental instructional services or interventions the students received during the experiment (i.e., when the treatment students were participating in Reading Recovery). Of the 3,579 first-grade teachers of control students in the study, we received surveys from 1,898 (54% response rate). From these responses, we obtained information on 1,245 (57%) of all students assigned to the control group in Years 2 and 3 of the MS-RCT.

We also analyzed lesson implementation data to assess control contamination, or the extent to which control students received one-to-one Reading Recovery lessons during the time when they were assigned to control. This is important because control contamination could attenuate the impacts. The research team analyzed intervention records—which include pre- and post-testing dates, intervention start and exit dates, and the total number of lessons provided to each student—to assess control contamination.

Results

Baseline Balance

Baseline balance was assessed for the analytic sample to confirm that the treatment and control groups were equivalent on observed characteristics. Table 6 presents results of baseline balance tests for student demographics and baseline TRL of the pooled analytic sample (6,888 students in 1,122 schools).

No significant differences were found between the groups on sex, ELL status, or race, suggesting that random assignment produced treatment and control groups that were well-balanced immediately prior to the start of treatment students' lessons. Baseline balance on prior reading performance, as measured by the TRL subscale, was also assessed. Again, no significant differences were found between the treatment and control groups at baseline. It is only possible to test for differences on measured characteristics; therefore, the possibility remains that there are systematic differences between the groups on characteristics we did not measure. However, this analysis provides strong evidence that the baseline characteristics and reading achievement of treatment and control groups were effectively identical on average, immediately prior to the start of the experiment.

In light of the significantly higher attrition among non-White students (see "Participants" section), following estimation of main impacts (see section "Findings"), sensitivity checks showed that student race (represented as a binary indicator for White/non-White) was nonstatistically significant

as a moderator of the treatment. As such, using available data from the analytic sample, there is no evident bias on the average treatment effect that is associated with higher attrition among non-White students in the study.

In the spring of Year 3, an online survey was administered to a sample of 145 Reading Recovery teachers in the study schools with missing assessment data. The purpose was to understand the reasons for missing student data and to explore any systematic differences between pairs that were retained and those that were not. Sixty-six Reading Recovery teachers responded (46% response rate) with detailed information about their students for whom no assessment data were available. In 48% of instances (37 out of 77 students) no assessment data were recorded because the child was unavailable (i.e., student moved or did not complete their full series of Reading Recovery lessons). Most of the reported instances of students moving or leaving the school occurred in the control group. The remaining 52% of students (n = 40) were reported to have missing assessment data for logistical reasons related to the assessment—for instance, because the teacher never received the testing materials or was not aware that she or he was required to administer and record the assessment. These survey results suggest that a large percentage (possibly half) of the missing test data were a result of natural attrition through student mobility.

Fidelity

Reading Recovery was implemented in the i3 scale-up with high fidelity to the program model, as codified in the *Standards and Guidelines* (Reading Recovery Council of North America, 2012). Overall, 90% of the indicators used to assess implementation had adequate implementation, and all four key components of implementation identified as essential for fidelity (staff background and selection, teacher leader and site capacity, training, and Reading Recovery lesson delivery) were implemented faithfully across the scale-up every year. In some areas, we consistently observed complete (100%) or very high adherence to program standards. Deviations from the program model were observed in a few specific areas, including minimum number of site visits from a trainer during the teacher leader's first year in the field; minimum number of professional development sessions each year, including a minimum number of behind-the-glass sessions; and administration of the OS and start of service to children within 2 weeks of school opening. We observed small deviations from the *Standards and Guidelines'* criteria for student-selection into Reading Recovery in 1 year of the study only. Overall, we observed that Reading Recovery was implemented with high fidelity to the program model, suggesting that the i3 scale-up successfully replicated Reading Recovery in the schools involved in the expansion and that the estimated impacts were indeed the result of faithful implementation of the intervention. A full description of the program standards included in the fidelity analysis, by key implementation component, is reported separately (May et al., 2016).

Dosage

In an analysis of treatment dosage, we calculated the percentage of the intervention that was received for each student based on the student's recorded number of Reading Recovery lessons and final text reading level. For instance, a student who reached Level 16 or completed 60 or more lessons (equal to five lessons per week for 12 weeks) was considered for this i3 evaluation to have received 100% of the treatment; a student who completed 30 lessons without reaching Level 16 was considered to have received 50% of the treatment. We found that 91.5% of students in the intervention received 100% of the treatment and that the average treatment dosage for students in the treatment group was 98.1%. Although our dosage calculation is an approximation, findings indicate that the estimated effects in this evaluation represent the effect of near-complete delivery of the intended treatment on students' reading achievement.

Control Group Instructional Supports

First-grade teachers in schools in the MS-RCT were surveyed about the control group's instruction during the experiment. The results of the survey revealed that 39% of the control students

TABLE 7

Descriptive Statistics for ITBS Scores and OS Total Scores for Treatment and Control Groups

Postintervention outcomes	Treatment group (n = 3,444)	Control group (n = 3,444)
ITBS Total Scale scores		
M (*SD*)	138.8 (7.5)	135.4 (7.2)
Mean percentile rank[a]	36	18
ITBS Comprehension Scale scores		
M (*SD*)	140.0 (9.5)	136.0 (9.0)
Mean percentile rank[a]	39	23
ITBS Reading Words Scale scores		
M (*SD*)	140.7 (9.0)	137.1 (8.2)
Mean percentile rank[a]	43	27
OS Total scores[b]		
M (*SD*)	496.5 (44.2)	451.4 (49.0)
Mean percentile rank	33	7

Note. ITBS = Iowa Tests of Basic Skills; OS = Observation Survey of Early Literacy Achievement.

[a]Percentile ranks based on ITBS Grade 1 midyear norms (Hoover et al., 2006).

[b]For OS scores, treatment n = 3,371; control n = 3,322.

received regular classroom instruction with no supplemental interventions during the experiment, 37% participated in some individual or small-group intervention (other than Reading Recovery) provided by a Reading Recovery–trained teacher, 23% participated in a literacy intervention delivered by a teacher who was not trained in Reading Recovery, and 8% received ELL or special education supports. Seven percent of control students received a combination of more than one of the supplemental instructional services listed above. These findings indicate that the majority of control group students (61%) did experience some form of supplemental literacy support in addition to regular classroom instruction. Therefore, in this study, we are comparing the effectiveness of Reading Recovery plus classroom instruction and other supplemental supports to that of classroom instruction and a range of other support services that schools provide to struggling first-grade readers. Furthermore, because most control students received some supplemental instructional supports during the evaluation period, we do not conclude that teachers' decisions regarding supplemental services were affected by the Reading Recovery evaluation. As with any RCT, it is impossible to assure that assignment to either the treatment or control condition will not affect students' experiences.

Control Group Contamination

We found only a small number of matched pairs—31 pairs out of 3,444 (<1%)—in which the control student was exposed to Reading Recovery before the posttest measures were administered. In these cases, the period of overlap (when both treatment and control students in a pair were receiving the intervention) ranged from 10 days to 177 days. In 14 of the 31 cases, the control student's intervention start date preceded that of the treatment student, indicating noncompliance with random assignment (0.4% noncompliance). In the 17 remaining cases, the treatment student began first but overlapped with control (0.5% control contamination). Given the very limited amount of noncompliance and control contamination in our analytic sample (less than 1% combined), the impact estimates we present here would remain practically unchanged after adjustment for noncompliance. As such, we present only the "intent-to-treat" estimates.

Impact Findings

Table 7 presents descriptive statistics for the treatment and control groups on scale scores from the posttest administration of the ITBS and OS measures. For each outcome, means in the treatment group are one third to one half of a standard deviation larger than the control group means. Differences in percentile ranks of group means are +18 for ITBS total scores, +16 for ITBS reading words, +16 for ITBS reading comprehension, and +26 for OS total scores.

Model Estimates for Main Impact Analyses. The full set of parameter estimates and standard errors from the HLM model are included in Table 8. Table 9 shows the model-based average scores for ITBS Total Reading Scale scores pooled across all years of the MS-RCT. The impact estimate for the difference between treatment and control students' ITBS total reading scores was 3.41 points ($p < .0001$; 95% confidence interval [CI] = 3.09, 3.72). Dividing this impact estimate by the standard deviation of the control group yields a Glass's Δ effect size of +0.48 standard

TABLE 8

HLM Parameter Estimates for Impacts on ITBS and OS Scores

	ITBS Total score	ITBS RW	ITBS Comp	OS total score
Fixed effects				
Intercept (β_0)	**133.71** (0.24)	**135.16** (0.27)	**134.23** (0.30)	**157.76** (4.88)
Pretest (β_1)	**1.30** (0.06)	**1.47** (0.08)	**1.39** (0.08)	**0.79** (0.01)
Treatment effect (β_2)	**3.70** (0.25)	**3.83** (0.30)	**4.32** (0.32)	**44.51** (1.48)
Year				
SY 2011–2012 (β_3)	0.44 (0.48)	0.61 (0.54)	0.38 (0.59)	**6.49** (2.56)
SY 2012–2013 (β_4)	0.48 (0.40)	0.55 (0.45)	0.70 (0.49)	2.14 (2.10)
SY 2013–2014 (β_5)	0.32 (0.38)	0.44 (0.43)	0.24 (0.47)	3.15 (2.00)
SY 2014–2015	Ref.	Ref.	Ref.	Ref.
Treatment Effect × Year				
SY 2011–2012 (β_6)	0.07 (0.53)	0.21 (0.63)	−0.31 (0.67)	−4.44 (3.09)
SY 2012–2013 (β_7)	−0.60 (0.44)	−0.39 (0.52)	−0.98 (0.56)	1.03 (2.54)
SY 2013–2014 (β_8)	−0.70 (0.42)	−0.79 (0.50)	−0.73 (0.53)	−2.78 (2.42)
SY 2014–2015	Ref.	Ref.	Ref.	Ref.
Random effects				
School Intercept (τ^2)	**14.20** (1.12)	**16.20** (1.40)	**18.44** (1.73)	**338.21** (30.5)
School Impact (ξ^2)	**10.01** (1.35)	**11.18** (1.89)	**13.53** (2.22)	**389.77** (44.9)
School Intercept/Impact (ρ)	**–0.24** (0.07)	**–0.25** (0.08)	–0.14 (0.09)	**–0.55** (0.05)
Matched Pair (ω^2)	**2.66** (0.67)	**2.69** (1.02)	**5.21** (1.19)	**48.19** (19.8)
Treatment Residual (σ^2)	**30.28** (1.10)	**51.02** (1.78)	**51.80** (1.92)	**825.02** (31.2)
Control Residual (σ^2)	**29.61** (1.09)	**42.98** (1.60)	**52.55** (1.93)	**935.48** (34.1)

Note. The analytic sample for the ITBS consists of 6,888 students in 1,122 schools. The analytic sample for the OS consists of 6,644 students in 1,122 schools. Significant parameter estimates ($p < .05$) are marked in bold type. See Equation 1 above for definitions of each model parameter. Year 1 coded 0. HLM = hierarchical linear model; ITBS = Iowa Tests of Basic Skills; RW = reading words; OS = Observation Survey of Early Literacy Achievement; SY = school year.

deviations. Alternatively, dividing the impact estimate by the standard deviation of the ITBS 2005 national norming sample of first-grade readers yields a Cohen's *d* effect size of +0.37 standard deviations relative to the national population of first graders.

Impacts on Target Subgroups. The results for rural schools were very similar to the overall results, and the average treatment effect was not significantly different from that in nonrural schools. The estimated difference between rural treatment and control students' total reading scores on the ITBS of 3.00 ($p < .0001$; 95% CI = 2.49, 3.51). Dividing that impact estimate by the standard deviation of the rural control group yields a Glass's Δ effect size of +0.43 standard deviations.

The results for ELL students were also very similar to the overall results, and the average treatment effect was no significantly different from that among non-ELL students. The impact estimate for the difference between ELL treatment and control students' expected total reading scores on the ITBS was 4.08 with a significant *p* value. Dividing that impact estimate by the standard deviation of the control group yields a Glass's Δ effect size of +0.57 standard deviations.

Variation in Impact Estimates. The significant variance components for random effects in the HLM models of impacts on ITBS scores suggest that the magnitude of the Reading Recovery impact estimates varies substantially across schools. Findings show an overall treatment effect of 3.41 points, with a random effect variance estimate of 10.01 points for the school-level impacts (see Table 8). Taking the square root of this variance estimate yields a standard deviation

TABLE 9

Impact Estimates on ITBS Total, ITBS Subtests, and Total OS

Midyear outcomes	Treatment group (n = 3,444)	Control group (n = 3,444)	Difference	Glass's Δ[a]	Cohen's d[b]
ITBS total reading scores					
M	138.71	135.30	+3.41		
(*SE*)	(0.16)	(0.15)	(0.16)	+0.48	+0.37
Mean percentile rank[c]	36	18	+18		
ITBS Reading Words Scale Scores					
Adjusted *M*	140.55	136.98	+3.57		
(*SE*)	(0.19)	(0.17)	(0.20)	+0.43	+0.35
Mean percentile rank[c]	43	27	+16		
ITBS Comprehension Scale Scores					
Adjusted *M*	139.82	135.92	+3.90		
(*SE*)	(0.21)	(0.18)	(0.21)	+0.43	+0.38
Mean percentile rank[c]	39	23	+16		
OS Total Raw Scores[d,e]					
Adjusted *M*	495.37	451.88	+43.49		
(*SE*)	(0.76)	(0.79)	(0.95)	+0.89	+0.99
Mean percentile rank	31	7	+24		

Note. ITBS = Iowa Tests of Basic Skills; OS = Observation Survey of Early Literacy Achievement; RW = reading words.
[a]Control *SD*: ITBS-T *SD* = 7.16; ITBS-C *SD* = 8.98; ITBS-RW *SD* = 8.23; OS-T *SD* = 49.43.
[b]Population *SD*: ITBS Level 6, Fall: ITBS-T *SD* = 9.1; ITBS-C *SD* = 10.2; ITBS-RW *SD* = 10.2; OS-T *SD* = 43.96.
[c]Percentile ranks based on ITBS Grade 1 midyear norms (Hoover et al., 2006).
[d]Percentile ranks based on U.S. Norms for OS Midyear (D'Agostino, 2012).
[e]Treatment n = 3,371; control n = 3,322.

of 3.2 points. We also find that the correlation between school-specific intercepts and impacts was negative and statistically significant (ρ = –.24, $p < .0001$). This suggests that the impacts of Reading Recovery tended to be larger in schools where students had lower average reading performance overall.

The Impact of Reading Recovery on ITBS Subtests and OS Total Scores. Table 9 presents the model-based average scores for treatment and control groups on ITBS subtests and Total OS pooled across all years of the MS-RCT. The full set of parameter estimates and standard errors from the HLM model are included in Table 8.

The impact estimate for the difference between treatment and control students' Reading Words scores on the ITBS was 3.57 points ($p < .0001$). Dividing that impact estimate by the standard deviation of the control group yields a Glass's Δ effect size of +0.43 standard deviations. Alternatively, dividing the impact estimate by the standard deviation from the ITBS national norming sample yields a Cohen's d effect size of +0.35 standard deviations. Analyses of impacts on the ITBS reading comprehension subtest showed similar results. The impact estimate for the difference between treatment and control students' reading comprehension scores on the ITBS was 3.90 points ($p < .0001$), which translates to a Glass's Δ effect size of 0.43 standard deviations and a Cohen's d of 0.38 standard deviations. Finally, the impact estimate for the difference between treatment and control students' OS Total scores was 43.5 ($p < .001$), which translates to a Glass's Δ effect size of 0.89 standard deviations and a Cohen's d of 0.99 standard deviations (D'Agostino, 2012; s = 43.96).

Study Limitations

Although this evaluation is the largest and most rigorous examination of Reading Recovery's impacts to date, it has several limitations that

warrant discussion. First, as the focus of this study is on assessing the feasibility of scaling up an effective program and rigorously confirming immediate impacts at scale, it does not address the long-term impacts of Reading Recovery. Clearly, understanding the maintenance or fade-out of the impacts observed in this study is an important next step. Members of the research team for the i3 evaluation have begun to examine this issue through a new, separately funded project using data from approximately 20,000 students in 1,200 schools in more than 30 states. At the end of this new project, we will be able to present a full understanding of Reading Recovery's short- and long-term impacts at scale.

A second limitation relates to the measurement of student data. Reading Recovery teachers and/or teacher leaders administered the ITBS and OS to participants. The large scale of the study presented practical challenges to blinded observations. To guard against potential bias in the event that assessment administrators were aware of students' assignment status, the primary outcome selected for this study was a well-established standardized assessment and Reading Recovery teachers did not administer the assessment to students in matched pairs for whom they provided instruction. However, teachers or teacher leaders indeed may have been aware of students' assigned condition presenting the potential for unintentional bias.

A final limitation of this study is its inability to explain substantial variation in program effect that were observed across schools. Future research on Reading Recovery should seek to identify the reasons for this variation.

Discussion

Despite these limitations, this study's consistent findings of medium to large effects on student achievement in reading over the course of this 4-year study offer evidence that Reading Recovery is an effective intervention that can help reverse struggling readers' trajectories of low literacy. Estimated treatment effects in each of the 4 years revealed effect sizes on the ITBS and its subtests that are large relative to typical effect sizes found in educational evaluations. In their paper on the interpretation of effect sizes, Lipsey et al. (2012) offered a number of useful benchmarks for understanding the magnitude of these effects. For randomized studies that use "broad scope" standardized tests as the outcome measure for interventions at the elementary level, the authors report average effects of 0.08 standard deviations (Lipsey et al., 2012). This benchmark suggests that the total standardized effect sizes (using Cohen's *d*) for Reading Recovery of 0.37 was 4.6 times greater than average for studies that use comparable outcome measures. Based on their analysis of 181 different samples, Lipsey et al. (2012) also presented mean effect sizes for different types of educational interventions. They report a mean standardized effect size of 0.13 for "curricula or broad instructional programs." The authors specifically include Reading Recovery in this group. This indicates that Reading Recovery's effects were 2.8 times greater than the reading outcomes of other instructional interventions. Similarly, the impacts of Reading Recovery were 3.5 times larger than the average effects of Title I programs reviewed by Borman and D'Agostino (1996).

It is also helpful to benchmark the treatment effects against expected gains on the ITBS for the national sample of students used to norm the ITBS tests. This permits the interpretation of impacts as an increase in growth rate during the study period. Table 10 shows the expected gains on the ITBS benchmarked against the national sample, the gains in terms of additional months of learning, and the growth rate for Reading Recovery students compared with the national average for beginning first graders.

From the start of first grade through the fifth month of the school year (the period during which the treatment students received Reading Recovery instruction), ITBS Reading Total Scale scores for the average student in the United States are expected to increase from 133 to 144 (Hoover et al., 2003). This increase of 11 points over a 5-month period suggests that the additional gains of 3.41 points experienced by Reading Recovery students of our evaluation is roughly equivalent to an additional 1.6 months of learning and translates to a growth rate that is 31% greater than the national average growth rate for beginning first graders. Table 8 also includes data for the Reading Comprehension and Reading Words subtests.

TABLE 10

Reading Recovery Treatment Effects as Compared With National Benchmarks for First Graders

	ITBS Total	ITBS RW	ITBS Comp
Treatment Effect (growth in ITBS scores)	3.41	3.57	3.90
Cohen's *d*	0.37	0.35	0.38
Treatment students' additional months of learning, over national growth average for first graders	1.55	1.62	1.77
Treatment students' growth rate, as a percentage of national average for first graders	131	132	135

Note. From the start of first grade through the fifth month (i.e., the period during which the treatment students received Reading Recovery instruction), ITBS Reading Total Scale scores are expected to increase from 133 to 144 for the average student in the United States (Hoover et al., 2003). The treatment effect represents additional gains experienced by students who received Reading Recovery. ITBS = Iowa Tests of Basic Skills; RW = reading words.

The average ITBS Total Reading Recovery score for the Reading Recovery (treatment) group was equivalent to the 36th percentile for students nationally, whereas the average score for the control group was equivalent to the 18th percentile for students nationally—a difference of +18 percentile points. A similar pattern of large gains in test scores for the Reading Recovery students relative to their control group counterparts was observed using subtests of the ITBS and the total OS. These findings were generally similar for students attending schools in rural settings and their counterparts in nonrural areas as well as for ELL students and their non-ELL counterparts.

These impacts are impressive. However, it is not enough for programs targeting early literacy to be efficacious. Effective interventions must also demonstrate impacts at scale. The evaluation of Reading Recovery's i3 scale-up is one of the most comprehensive and rigorous studies to date of the impacts of an educational intervention at scale. Its findings support the feasibility of successfully scaling up of an effective intervention.

Authors' Note

The opinions expressed are those of the authors and do not represent the views of the Office of Innovation and Improvement (OII), the Institute of Education Sciences (IES), or the U.S. Department of Education.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Office of Innovation and Improvement (OII), U.S. Department of Education, through Grant U396A100027 to The Ohio State University, and in part by the Institute of Education Sciences (IES), U.S. Department of Education, through Grant R305B090015 to the University of Pennsylvania.

Notes

1. Two other studies (Baenen et al., 1997; Iverson & Tunmer, 1993) were originally found to meet WWC evidence standards but were later disqualified.

2. In cases where there was an odd number of English language learners (ELL) students, the ELL student with the highest TRL score was matched with a non-ELL student with a lower TRL score.

References

Allington, R. (2005). How much evidence is enough evidence? *The Journal of Reading Recovery*, *4*(2), 8–11.

Annie E. Casey Foundation. (2010). *Early warning! Why reading by the end of third grade matters*. Baltimore, MD: Author.

Annie E. Casey Foundation. (2016). *Too many young readers still aren't making the grade*. Retrieved from http://www.aecf.org/blog/too-many-young-readers-still-arent-making-the-grade/

Ashdown, J., & Simic, O. (2000). Is early literacy intervention effective for English language learners? Evidence from Reading Recovery. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing*, *5*, 27–42.

Baenen, N., Bernhold, A., Dulaney, C., & Bankes, K. (1997). Reading recovery: Long-term progress after three cohorts. *Journal of Education for Students Placed at Risk*, *2*, 161–181.

Balfanz, R., Bridgeland, J. M., Bruce, M., & Fox, J. H. (2012). *Building a Grad Nation: Progress and*

challenge in ending the high school dropout epidemic (Annual Update, 2012). Washington, DC: Civic Enterprises.

Bates, C. C., D'Agostino, J. V., Gambrell, L., & Xu, M. (2016). Reading Recovery: Exploring the effects on first-graders' reading motivation and achievement. *Journal of Education for Students Placed at Risk*, *21*, 47–59.

Borman, G. D., & D'Agostino, J. V. (1996). Title I and student achievement: A meta-analysis of federal evaluation results. *Educational Evaluation and Policy Analysis*, *18*, 309–326.

Center, Y., Wheldall, K., Freeman, L., Outhred, L., & McNaught, M. (1995). An experimental evaluation of Reading Recovery. *Reading Research Quarterly*, *30*, 240–263.

Century, J., Cassata, A., Rudnick, M., & Freeman, C. (2012). Measuring enactment of innovations and the factors that affect implementation and sustainability: Moving toward common language and shared conceptual understanding. *The Journal of Behavioral Health Services & Research*, *39*, 343–361.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*, 199–218.

Clay, M. M. (1987). Implementing Reading Recovery: Systemic adaptations to an educational innovation. *New Zealand Journal of Educational Studies*, *22*(1), 35–58.

Clay, M. M. (1991). *Becoming literate: The construction of inner control*. Portsmouth, NH: Heinemann.

Clay, M. M. (2002). *An Observation Survey of Early Literacy Achievement* (2nd ed.). Portsmouth, NH: Heinemann.

Clay, M. M. (2005a). *Literacy lessons designed for individuals: Why? when? and how?* Portsmouth, NH: Heinemann.

Clay, M. M. (2005b). *An Observation Survey of Early Literacy Achievement* (2nd ed.). Portsmouth, NH: Heinemann.

D'Agostino, J. V. (2012). *US norms and correlations for an Observation Survey of Early Literacy Achievement*. Columbus: National Data Evaluation Center: The Ohio State University, College of Education, School of Teaching and Learning Reading Recovery.

D'Agostino, J., & Murphy, J. (2004). A meta-analysis of Reading Recovery in United States' schools. *Educational Evaluation and Policy Analysis*, *26*, 23–38.

D'Agostino, J., & Rodgers, E. (2015). *Scaling Up What Works, Reading Recovery*. Final performance report. Washington, DC: U.S. Department of Education.

Denton, C. A., Ciancio, D. J., & Fletcher, J. M. (2006). Validity, reliability, and utility of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly*, *41*, 8–34.

Dhillon, S., Darrow, C., & Meyers, C. V. (2015). Introduction to implementation fidelity. In C. V. Meyers & W. C. Brandt (Eds.), *Implementation fidelity in education research: Designer and evaluator considerations* (pp. 8–22). New York, NY: Routledge.

Gómez-Bellengé, F., Gibson, S., Tang, M., Doyle, M., & Kelly, P. (2007). *Assessment and identification of first-grade students at risk: Correlating the dynamic indicator of basic early literacy skills and an Observation Survey of Early Literacy Achievement*. Available from https://www.idec-web.us/

Goodson, B., Price, C., & Darrow, C. (2015). Measuring fidelity: The present and future. In C. V. Meyers & W. C. Brandt (Eds.), *Implementation fidelity in education research: Designer and evaluator considerations* (pp. 8–22). New York, NY: Routledge.

Hernandez, D. J. (2011). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation*. Baltimore, MD: Annie E. Casey Foundation.

Hiebert, E. H. (1994). Reading Recovery in the United States: What difference does it make to an age cohort? *Educational Researcher*, *23*(9), 15–25.

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., & Levin, H. M. (2016). Cost-effectiveness analysis of early reading programs: A demonstration with recommendations for future research. *Journal of Research on Educational Effectiveness*, *9*, 30–53.

Holliman, A. J., & Hurry, J. (2013). The effects of Reading Recovery on children's literacy progress and special educational needs status: A three-year follow-up study. *Educational Psychology*, *33*, 719–733.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Bray, G. B., Naylor, R. J., . . . Qualls, A. L. (2006). *The Iowa tests: 2005 norms and score conversions*. Iowa City: University of Iowa.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., . . . Shannon, G. P. (2003). *The Iowa tests: Guide to research and development*. Itasca, IL: Riverside Publishing.

Iverson, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology*, *85*, 112–126.

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

Laurentson, M., Oh, Y. J., & LaBlanca, F. (2015). STEM21 Digital Academy fidelity of implementations: Valuation and assessment of program components and implementation. In C. V. Meyers & W. C. Brandt (Eds.), *Implementation fidelity in education research: Designer and evaluator considerations* (pp. 8–22). New York, NY: Routledge.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, DC: National Center for Special Education Research.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., & Olson, R. (2001). Rethinking learning disabilities. In C. E. Finn Jr., A. J. Rotherham, & C. R. Hokanson Jr. (Eds.), *Rethinking special education for a new century* (pp. 259–287). Washington, DC: Thomas B. Fordham Foundation.

Lyons, C. A. (1998). Reading recovery in the United States: More than a decade of data. *Literacy, Teaching and Learning*, *3*(1), 77.

May, H., Goldsworthy, H., Armijo, M., Gray, A., Sirinides, P., Blalock, T., & Sam, C. (2014). *Evaluation of the i3 scale-up of Reading Recovery: Year two report*. Philadelphia, PA: Consortium for Policy Research in Education.

May, H., Gray, A., Gillespie, J., Sirinides, P., Sam, C., Goldsworthy, H., & Tognatta, N. (2013). *Evaluation of the i3 scale-up of Reading Recovery: Year one report*. Philadelphia, PA: Consortium for Policy Research in Education.

May, H., Gray, A., Sirinides, P., Gillespie, J., Sam, C., Goldsworthy, H., . . . Tognatta, N. (2015). Year one results from the multisite randomized evaluation of the i3 scale-up of Reading Recovery. *American Educational Research Journal*, *52*, 547–581.

May, H., Sirinides, P., Gray, A., & Goldsworthy, H. (2016). *Reading Recovery: An evaluation of the four-year i3 scale-up*. Philadelphia, PA: Consortium for Policy Research in Education.

Mowbry, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315–340.

Pinnell, G. (1989). Reading recovery: Helping at-risk children learn to read. *The Elementary School Journal*, *90*, 161–183.

Pinnell, G. S., DeFord, D. E., & Lyons, C. A. (1988). *Reading Recovery: Early intervention for at-risk first graders* [Educational Research Service Monograph]. Arlington, VA: Educational Research Service.

Pinnell, G. S., Lyons, C. A., DeFord, D. E., Bryk, A. S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly*, *29*, 8–39.

Quay, L., Steele, D., Johnson, C., & Hortman, W. (2001). Children's achievement and personal and social development in a first-year Reading Recovery program with teachers in training. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing*, *5*(2), 7–25.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: SAGE.

Reading Recovery Council of North America. (2012). *Standards and guidelines of Reading Recovery in the United States* (6th ed.). Worthington, OH: Author.

Reynolds, M., & Wheldall, K. (2007). Reading Recovery 20 years down the track: Looking forward, looking back. *International Journal of Disability, Development, and Education*, *54*, 199–223.

Rodgers, E., Gómez-Bellengé, F., Wang, C., & Schulz, M. (2005, April). *Predicting the literacy achievement of struggling readers: Does intervening early make a difference?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Rodgers, E., Wang, C., & Gómez-Bellengé, F. (2004, April). *Closing the literacy achievement gap with early intervention*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the Reading Recovery early intervention. *Journal of Educational Psychology*, *97*, 257–267.

Shanahan, T., & Barr, R. (1995). Reading Recovery: An independent evaluation of the effects of an early instructional intervention for at-risk learners. *Reading Research Quarterly*, *20*, 958–995.

Shaywitz, S. E., Fletcher, J. M., Holahan, J. M., Shneider, A. E., Marchione, K. E., Stuebing, K. K., & Shaywitz, B. A. (1999). Persistence of dyslexia: The Connecticut Longitudinal Study at adolescence. *Pediatrics*, *104*, 1351–1359.

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*, 1–26.

Summerfelt, W. T. (2003). Program strength and fidelity in evaluation. *Applied Developmental Science*, *7*(2), 55–61.

Tang, M., & Gómez-Bellengé, F. X. (2007, April). *Dimensionality and concurrent validity of the*

Observation Survey of Early Literacy Achievement. Paper presented at the annual meeting of the American Educational Research Association in Chicago, IL.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, July). *Beginning reading intervention report: Reading Recovery®*. Available from http://whatworks.ed.gov

Authors

PHILIP SIRINIDES, PhD, is a research assistant professor at the University of Pennsylvania and senior researcher at the Consortium for Policy Research in Education (CPRE). He is an applied quantitative researcher who specialized in the development and use of integrated data systems for public sector planning and evaluation.

ABIGAIL GRAY, PhD, is senior researcher at the CPRE at the University of Pennsylvania. She leads multiple mixed-methods research studies to inform policy and practice in K–12 schools, particularly in the areas of early literacy and school climate and discipline.

HENRY MAY, PhD, is director of the Center for Research in Education and Social Policy (CRESP), University of Delaware. He specializes in the application of modern statistical methods and mixed-methods in randomized experiments and quasi-experiments studying the implementation and impacts of educational and social interventions and policies.

Manuscript received August 19, 2016
First revision received October 31, 2017
Second revision received January 13, 2018
Accepted February 12, 2018