American Educational Research Journal
June 2015, Vol. 52, No. 3, pp. 547–581
DOI: 10.3102/0002831214565788

Year One Results From the Multisite Randomized Evaluation of the i3 Scale-Up of Reading Recovery

Henry May
University of Delaware
Abigail Gray
Philip Sirinides
Heather Goldsworthy
Michael Armijo
Cecile Sam
Jessica N. Gillespie
Namrata Tognatta
University of Pennsylvania

Reading Recovery (RR) is a short-term, one-to-one intervention designed to help the lowest achieving readers in first grade. This article presents first-year results from the multisite randomized controlled trial (RCT) and implementation study under the $55 million Investing in Innovation (i3) Scale-Up Project. For the 2011–2012 school year, the estimated standardized effect of RR on students' Iowa Tests of Basic Skills (ITBS) Total Reading Scores was .69 standard deviations relative to the population of struggling readers eligible for RR under the i3 scale-up and .47 standard deviations relative to the nationwide population of all first graders. School-level implementation of RR was, in most respects, faithful to the RR Standards and Guidelines, *and the intensive training provided to new RR teachers was viewed as critical to successful implementation.*

KEYWORDS: early literacy, randomized experiment, program evaluation

Instructional initiatives that are able to identify students at greatest risk and help them achieve grade-level proficiency are essential to addressing the epidemic of low literacy in the early elementary grades. Several studies have shown that most students who leave first grade reading below grade level never catch up (Juel, 1988; Lyon et al., 2001; Shaywitz et al., 1999). Others suggest that literacy interventions are most effective if they are instructionally rigorous and focused on the early years of schooling (Johnston, Allington, & Afflerback, 1985; Morrow, 1993; National Research

Council, 1998; Pikulski, 1994; Strickland, 2002). However, schools often lack the structures or expertise to implement an intervention with the intensity required to overcome low literacy. Reading Recovery was designed in direct response to such challenges. Developed in the 1970s by educator and psychologist Marie Clay, Reading Recovery is one of the oldest and most widely implemented literacy interventions in the world.

As part of the 2010 economic stimulus, Reading Recovery was awarded a $45 million "Investing in Innovation" (i3) grant from the U.S. Department of Education, along with an additional $10.1 million from private sources, to fund the scale-up of Reading Recovery across the nation. This five-year grant is intended to expand Reading Recovery in more than 1,400 schools and provide targeted literacy assistance to over 88,000 students. This article presents

---

HENRY MAY, PhD, is director of the Center for Research in Education and Social Policy (CRESP), University of Delaware, 201 Willard Hall Education Building, Newark, DE 19716, USA; e-mail: *hmay@udel.edu*. 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.

ABIGAIL GRAY, PhD, is a senior researcher at the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania. She is a mixed-methods researcher whose work focuses primarily on implementation research in the context of experimental studies, and on exploring the relationship between program implementation and impacts.

PHILIP SIRINIDES, PhD, is a senior researcher at the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania. He specializes in the application of quantitative research methods and the development and use of integrated data systems for public-sector planning and evaluation.

HEATHER GOLDSWORTHY, PhD, is a research specialist at the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania. She specializes in qualitative research, particularly as a method for exploring the structures and impacts of programs and policies.

MICHAEL ARMIJO, PhD, is an assistant research scientist at the College Board and a recent doctoral graduate from the University of Pennsylvania. His research interests include evaluation design, implementation fidelity, and impact analysis using randomized controlled trials and quasi-experimental designs.

CECILE SAM, PhD, is a research specialist at the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania. She designs and conducts studies on academic community and faculty work in a larger K-20 setting, as well as district reform evaluation.

JESSICA N. GILLESPIE, MS, is an impact analyst at the United Way of Greater Philadelphia & Southern New Jersey and was previously a research specialist at the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania.

NAMRATA TOGNATTA, PhD, is an education consultant for the World Bank and a recent doctoral graduate from the University of Pennsylvania. She specializes in quantitative methods for program evaluation, with a focus on education and social policy issues in developing countries.

first-year results from the external evaluation of the Reading Recovery i3 Scale-Up, providing the first evidence as to whether this massive endeavor is producing programs that reflect faithful implementation of the intervention and meaningful improvements in students' reading skills.

Reading Recovery's Theory of Action

Reading Recovery is an intensive intervention targeting the lowest achieving 15% to 20% of first-grade readers. It takes as its underlying principle the idea that individualized, short-term, highly responsive instruction delivered by an expert can disrupt the trajectory of low literacy achievement, produce accelerated gains, and enable students to catch up to their peers and sustain achievement at grade level into the future. Reading Recovery lessons attend to phonemic awareness, phonics, vocabulary, fluency, and comprehension—the critical elements of literacy and reading instruction identified by the National Reading Panel (2000). At its core, Reading Recovery is intended to help students develop a set of self-regulated strategies for problem-solving words, self-monitoring, and self-correcting that they can apply to the interpretation of text. These strategies focus on enabling students to use meaning, structure, letter-sound relationships, and visual cues in their reading and writing (Clay, 1991, 2005).

According to the program model of Reading Recovery (RR), highly trained teachers provide daily instruction to students during 30-minute, one-to-one teaching sessions. Teachers tailor their lessons to a student's individual strengths and needs based on their own observations. According to Clay (2005), RR teachers "must be able to design a superbly sequenced series of lessons determined by the particular children's competencies, and make highly skilled decisions [at each] moment during the lesson" (p. 23). A primary assumption of the Reading Recovery model is that high-quality instruction is key to accelerated progress in literacy learning. Ensuring instructional quality is therefore considered essential to the success of the intervention, and the RR model's multilevel structure, visualized in Figure 1, emphasizes instructional quality at each level. At the top level of the RR model, faculty at University Training Centers (UTC) train RR teacher leaders through an in-residence, postgraduate program of study. Through this training, as well as through mentorship experiences with faculty and experienced teacher leaders, new RR teacher leaders are expected to become expert literacy coaches with a deep understanding of how children learn and become literate and why some children have great difficulty learning to read and write. Teacher leaders also become expert teachers of both children and adult learners; through ongoing work with students and mentoring by university trainers, RR teacher leaders are taught how to design and deliver literacy lessons to individual students with a focus on those students who are making particularly slow progress and also how to deliver training and support to RR

teachers. Developing these core understandings about learning and instruction prepares RR teacher leaders for their role in the Reading Recovery model's second level of instruction. At this level, RR teacher leaders provide a year-long academic program of training and support to the school-based RR teachers who are simultaneously working directly with students. First, all new RR teachers in training complete a week-long summer introduction focused on the administration, scoring, and interpretation of the Observation Survey of Early Literacy Achievement (OS). Second, they complete a year-long academic course taught by a teacher leader, for which they receive a grade and 8 to 10 university credits. This coursework uses as its primary text Marie Clay's (2005) *Literacy Lessons Designed for Individuals, Parts I and II* and incorporates regular opportunities to observe instruction through "behind-the-glass" lessons in which they observe each other's one-to-one lessons through a one-way mirror. As they complete this course, RR teachers in training provide daily one-to-one lessons to four RR students each; they also attend weekly three-hour training sessions conducted by their teacher leader. Finally, both trained RR teachers and those in training receive school-based visits from their teacher leader, who observes them in Reading Recovery lessons and provides immediate feedback and suggestions. This intensive training and support emphasizes the skills that are critical for individualized, highly responsive instruction, including theory of literacy acquisition, assessment, systematic observation, instruction, analysis, and reflection. As they continue past their first year, RR teachers receive ongoing professional development and supervision from their RR teacher leader to ensure continual learning and growth. At the final level of instruction in the RR model, trained RR teachers design and deliver individualized lessons to students identified to receive RR services. These lessons are structured yet unscripted, relying heavily on the teacher's ability, developed through intensive training, to recognize and implement the most effective instructional strategies in immediate response to a student's reading performance during each lesson. Each lesson begins with re-reading familiar books and a running record, then word or letter work on the wallboard, story composition, assembling a cut-up sentence, and finally previewing and reading a new book. These activities are essential elements in the instructional experience of RR students and are expected to occur in every one-to-one RR lesson.

Within the Response to Intervention (RtI) framework, some have labeled Reading Recovery as a "problem-solving" (PS) approach because it relies heavily on teachers' expertise and ability to enact instruction that is responsive to continuous formative data about a student's performance (e.g., Johnston, 2011). While this reliance on instructional expertise is certainly true, the RR intervention is actually structured and delivered in a very consistent manner from one student to the next. This aspect of RR makes it a much more stable and replicable intervention across students and

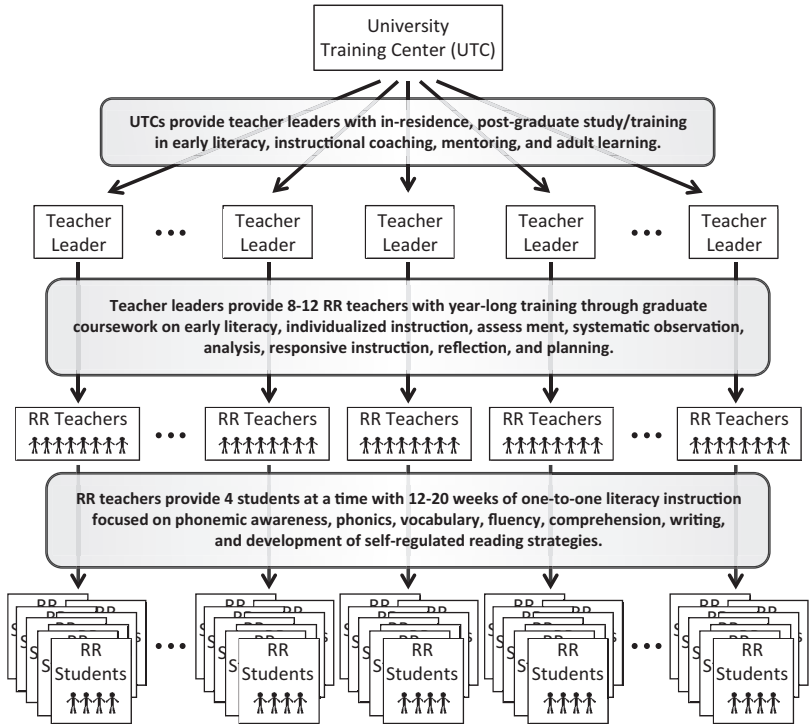

Figure 1. **Theory of action for Reading Recovery (RR) training and intervention.**

contexts—a feature that many would not associate with PS models. Likewise, many feel that the primary alternative approach to RtI, the standard treatment protocol (STP), involves scripted instruction (Johnston, 2011). However, in Fuchs and Fuchs's (2006) description of STP, they imply that there should be considerable flexibility in the intervention as appropriate to each student's needs. In the example cited by Fuchs and Fuchs (2006, p. 95) as an exemplar of the STP approach (i.e., Vellutino et al., 1996), Vellutino and colleagues describe an intervention that is "tailored to the child's individual needs" and whose structure is remarkably similar to that of a Reading Recovery lesson (Vellutino et al., 1996, p. 610). In fact, Vellutino (2010) describes Reading Recovery as "clearly the prototype for RtI approaches to identifying children at risk for long-term reading difficulties" (p. 22). As such, we see Reading Recovery as having features that align very well with Fuchs and Fuchs's original definition of the STP approach to RtI, but with a heavy reliance on the adaptive instruction and teacher

expertise usually associated with PS, and without the scripted instruction that has come to be associated with STP.

Supporting the consistency of implementation of Reading Recovery across sites, lessons, and students, the *Standards and Guidelines of Reading Recovery in the United States* (Reading Recovery Council of North America [RRCNA], 2009) codify Clay's (2005) perspectives on RR and make explicit the practices that constitute faithful implementation. Thus, the *Standards and Guidelines* (RRCNA, 2009) formalize the processes by which individual actors and systems achieve the dimensions of change advocated by Clay and guide the implementation of Reading Recovery across geographically dispersed sites and diverse contexts.

Prior Research on Reading Recovery

Reading Recovery has been the focus of considerable study over the course of its nearly 30-year history. Research has focused on various issues, ranging from the merits and limitations of the intervention's instructional approach (Chapman, Tumner, & Prochnow, 1999; Iversen & Tunmer, 1993; Moats, 2007; Pinnell, DeFord, & Lyons, 1994; Tunmer & Chapman, 2002) to its cost-effectiveness (Dyer & Binkney, 1995; Gómez-Bellengé, 2002; Hiebert, 1994; Rasinksi, 1995). Most relevant to the current study is the subset of the Reading Recovery research that examines the intervention's efficacy in terms of its goal of producing accelerated and long-lasting gains in students' literacy achievement. The vast majority of these studies have used quasi-experimental designs that are subject to issues of selection bias (Allington, 2005; Ashdown & Simic, 2000; Briggs & Young, 2003; Brown, Denton, Kelly, & Neal, 1999; Center, Wheldall, Freeman, Outhred, & McNaught, 1995; Elbaum, Vaughn, & Moody, 2000; Hiebert, 1994; Pinnell, 1989; Quay, Steele, Johnson, & Hortman, 2001; Rodgers, Gómez-Bellengé, Wang, & Schultz, 2005; Rodgers, Wang, & Gómez-Bellengé, 2004; Rowe, 1995; Ruhe & Moore, 2005; Schmitt & Gregory, 2005; Shanahan & Barr, 1995; Torgesen, 2000; Tunmer & Chapman, 2002; Wasik & Slavin, 1993). As such, interpretation of results from these studies is inconclusive given the variation in findings and the inherent weaknesses of the methodologies used. In their meta-analysis of Reading Recovery research, D'Agostino and Murphy (2004) posit that Reading Recovery

> has proven a very difficult program to evaluate, given its student selection and attrition policies, the barriers to locating an equivalent comparison group, the reliance on outcome measures designed for the program, and the problems inherent with accurately measuring students' achievement levels in first grade. (p. 24)

The What Works Clearinghouse (WWC) report on Reading Recovery (WWC, 2008) excluded the aforementioned quasi-experimental studies

from its review, focusing instead on just a few studies that were determined to meet the WWC evidence standards.[1] Four of these studies were randomized controlled trials (i.e., Baenen, Bernhole, Dulaney, & Banks, 1997; Pinnell et al., 1988; Pinnell, Lyons, DeFord, Bryk, & Seltzer, 1994; Schwartz, 2005). In each case, a treatment group of first-grade students who received Reading Recovery services was compared with a control group of first graders who did not participate in the intervention. Treatment and control were randomly assigned. A fifth study reviewed (i.e., Iversen & Tunmer, 1993) was quasi-experimental; the treatment and control groups were not assigned randomly. This study was deemed to meet the WWC standards "with reservations" given that baseline equivalence was documented despite the lack of random assignment.

The studies differed from one another in terms of assessment strategy and sample size and dispersion as well as the scope of the research. On the basis of the empirical impact analyses included in the five studies that met WWC evidence standards, the 2008 WWC report on Reading Recovery found positive effects in two of the key literacy indicators assessed: alphabetics and general reading achievement. The WWC average Improvement Indices[2] were +34 percentile points in alphabetics and +46 percentile points in fluency, which correspond to effect sizes of +1.0 and +1.7 standard deviations. In the other two areas, comprehension and overall reading achievement, the WWC identified "potentially positive effects," with average Improvement Indices +14 percentile points in comprehension and +32 percentile points in overall reading achievement, which correspond to effect sizes of +.35 and +.90 standard deviations. Furthermore, the sample sizes in the studies that met WWC standards were relatively small (i.e., <100), leading to the conclusion of "potentially positive effects, with a small extent of evidence."

While on the whole very positive, the What Works Clearinghouse report's findings—coupled with the shortage of studies that met WWC evidence standards and no large-sample studies to date—strongly suggest the need for a large-scale and highly rigorous examination of the impacts of Reading Recovery and variation in impacts across a large sample of schools.

Methods

The overarching goals of this evaluation are to document the implementation of Reading Recovery under the i3 scale-up and produce rigorous estimates of the causal impacts of the intervention on student literacy achievement. A multisite randomized controlled trial (RCT) was used to estimate program impacts, while a mixed-methods study of program implementation was conducted using surveys, activity logs, interviews, focus groups, observations, and document analysis. The primary research questions for the evaluation are:

Research Question 1: For first-grade students who begin the school year struggling to read, what are the impacts of Reading Recovery on reading achievement at the end of the 12- to 20-week intervention?

Research Question 2: Is training of RR teachers being implemented as intended, and what is the perceived quality of the training?

Research Question 3: Is the Reading Recovery intervention being implemented as intended, and what factors may support or hinder fidelity of implementation?

Sample Selection and Random Assignment

School and Student RCT Sample

Prior to the start of the 2011–2012 school year, 209 schools participating in the i3 scale-up (out of a total of 627 participating schools) were randomly selected for inclusion in a randomized controlled trial. Of these, 25 schools dropped out of the i3 study before random assignment was conducted. The remaining 184 schools selected for the RCT were instructed to implement the following protocol.

At each participating school, a subsample of eight low-performing students eligible for Reading Recovery was identified using the Reading Recovery Observation Survey of Early Literacy Achievement. The names of these eight eligible students, along with their English language learner (ELL) status and pretest Text Reading Level scores from the OS, were entered into an online random assignment tool. The tool then rank-ordered the students by their ELL status and Text Reading Level, and students were matched into pairs according to pretest scores and ELL status. One student in each pair was randomly assigned to the treatment group (i.e., classroom instruction plus RR one-to-one lessons), and the other student in the pair was assigned to the control group (i.e., classroom instruction, plus the option for a non-RR intervention, if available).[3,4] Reading Recovery teachers were directed to begin RR Lessons with the treatment students right away and to begin RR Lessons with each control student only after their treatment counterpart had exited the intervention and only if the control student remained in need of RR at that time.[5]

Qualitative Sample

Our study of the implementation of Reading Recovery during the 2011–2012 school year involved extensive interviews with core RR stakeholders whose professional responsibilities involved the recruitment of i3 participants, the training of i3 teacher leaders and RR teachers, implementation of RR lessons, and the oversight and support of RR implementation activities. Participants in the qualitative study components were recruited as follows: Eighteen out of 19 UTC directors participated in individual interviews conducted in person or by phone. A total of 49 teacher leaders attending the annual Teacher Leader Institute agreed to participate in one of five focus groups. Fifty Reading

Recovery teachers were randomly sampled from the population of i3 RR teachers, and 37 of these agreed to participate in individual phone interviews. An additional 9 RR teachers were interviewed in person during school site visits,[6] and observations of at least two RR lessons were also done with each of these 9 RR teachers. Individual interviews were conducted with 17 school principals, including 9 from the site visit schools and 8 more from a similarly stratified random sample of schools (see note 6). Recruitment of participants involved a process of personal email invitations and follow-up phone calls. No monetary compensation was provided to participants.

Measures and Data Collection

Student Outcomes

Upon the conclusion of the intervention for each treatment group student, both students in that matched pair were assessed at approximately the same time using the reading section of the Iowa Tests of Basic Skills (ITBS). To address possible test administrator bias, the posttest assessments with each student were conducted by someone other than the teacher who provided RR lessons to that student (i.e., typically another RR teacher or a teacher leader). Furthermore, the ITBS reading tests are administered in a standardized format that inhibits the potential for test administrators to influence results.

Teachers were instructed to administer the ITBS test to any treatment group student who exited Reading Recovery as soon as possible after his or her exit. Students who did not complete the minimum 12-week intervention because they were referred for additional services, dropped due to nonattendance, or exited for any other reason were still expected to take the ITBS. Unfortunately, many of these students could not be tested because they had moved away from the school. Of the 13,328 RR students in i3 schools during the 2011–2012 school year for whom final intervention status was recorded in the Reading Recovery International Data Evaluation Center (IDEC), 52.4% of students in Reading Recovery successfully discontinued before the 20th week with no referral for additional intervention. Of the remaining students, 22.4% were referred for additional intervention before or after the 20th week, 4.7% changed schools, and 19.7% received less than 12 weeks of lessons (i.e., typically due to student absenteeism, the end of the school year, or the RR teacher leaving the school midyear).

IDEC provided existing infrastructure to support large-scale data collection at low cost under this evaluation. Since 1998, IDEC has provided data collection, management, and analysis support to Reading Recovery programs throughout the nation. IDEC has developed a two-step data entry and review process. First, RR teachers are required to enter data regarding their personal characteristics, along with student and school information; teacher leaders are then required to review this data as an additional quality

assurance measure. IDEC data are used by UTC directors, teacher leaders, RR teachers, and school administrators for their annual reporting. As part of the scale-up evaluation, we coordinated with IDEC to allow RR teachers to enter pretest OS data and posttest ITBS and OS data for students in the RCT each year. The data files extracted from IDEC included test scores along with student, teacher, and building characteristics. These data were used to: (a) generate sampling frames for the RCT, survey, and qualitative data collection components of the evaluation; (b) report aggregate statistics regarding Reading Recovery's implementation; and (c) compare outcomes for treatment and control students participating in the RCT.

Implementation Surveys

In spring of 2012, we conducted surveys of key RR staff in order to gauge fidelity of implementation. Fidelity data were collected via online surveys, and questions were worded to minimize bias in responses toward the "right" answer. The response rates and sample sizes for site coordinators, teacher leaders, and RR teachers were 64% (n = 105), 89% (n = 169), and 78% (n = 799), respectively. These three separate surveys were designed to collect data on implementation efforts in relation to the *Standards and Guidelines of Reading Recovery in the United States* (RRCNA, 2009), which provides detailed specifications for implementation of Reading Recovery. In addition, i3-supported RR teachers were asked to complete daily activity logs on four randomly selected days from January through April 2012. Seventy-one percent of RR teachers (n = 882) completed at least one log, with the majority completing three or four logs. In total, RR teachers completed 2,623 logs documenting their daily activities from January through April 2012. The survey and log data allowed for quantification of implementation fidelity across sites and by subdomain based on the RR *Standards and Guidelines* (RRCNA, 2009). All surveys and logs were designed and constructed by the research team and administered online using the Qualtrics survey suite.

Interview Protocols

Semi-structured protocols were developed for each set of interview and focus group participants. The interviews explored issues of implementation of the Reading Recovery intervention and how information sharing processes were used to inform and refine practices designed to document, facilitate, and support the attainment of scale-up benchmarks. Specific questions during the interviews addressed implementation issues directly related to the RR *Standards and Guidelines* (RRCNA, 2009), including recruitment of schools and RR teachers; format, frequency, and quality of training classes; conduct of one-to-one RR lessons; and supports and barriers to program implementation.

Data Analyses

Statistical Models of Program Impacts

The impacts of Reading Recovery on student reading performance were estimated by comparing midyear reading achievement of students randomly assigned to participate in Reading Recovery at the beginning of first grade to students randomly assigned to the control condition. Using a three-level hierarchical linear model (HLM) (Raudenbush & Bryk, 2002), students were nested within matched pairs and matched pairs nested within schools. Differences in the posttest performance of the treatment and control students were estimated after controlling for pretest performance. This HLM included the pretest OS Text Reading Level scores as a covariate, random effects for blocks (i.e., matched pairs), a random effect for overall school performance (i.e., random school intercepts), and a random effect for the impact of Reading Recovery (i.e., random treatment effects across schools). The mathematical form of the HLM model is as follows.[7]

$$Y_{ijk}=\beta_0+\beta_1(Pretest)+\beta_2(Trt)+\gamma_j+\alpha_k+\varphi_k(Trt)+\varepsilon_{ijk} \quad (1)$$

$$\text{With}: \gamma_j \sim N(0,\omega^2), \begin{pmatrix}\alpha_k\\ \varphi_k\end{pmatrix} \sim MVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau^2 & \rho\\ \rho & \xi^2\end{pmatrix}\right) \text{and } \varepsilon_{ijk} \sim N(0,\sigma^2_{TC}) \quad (2)$$

where Y_{ijk} is the posttest outcome score for student i from pair j in school k; β_0 is the model intercept; β_1 is the slope coefficient for the pretest covariate (i.e., from the OS); β_2 is the overall treatment effect; *Trt* is the treatment assignment indicator (i.e., with 1 = treatment and 0 = control); 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 variance σ^2_T for the treatment group and σ^2_C for the control group.

Impact estimates from the HLM models represent mean differences in ITBS scale scores between treatment and control groups after adjusting for initial text reading level on the OS. To improve interpretability, we convert these raw impact estimates into three alternative standardized metrics: a sample-based Glass's Δ, a population-based Cohen's d, and a WWC Improvement Index.[8] We also benchmark the Reading Recovery impact estimates against average impacts of Title I interventions and also against the expected growth rate of the average first grader from the national norming sample for the ITBS (Hoover et al., 2006).

The large number of schools involved in this multisite study allows for HLM analyses of school-level variability in program effects. The use of

random treatment effects for schools in a multilevel modeling framework allows estimation of variability in treatment effects across schools and supports estimation of cross-level interactions to explain variability in treatment effects. Analyses in this article include estimates of the standard deviation of program effects across the sample of schools, which are based on the random effects variance estimates from the HLM analyses. We do not include visualizations of site-specific effects (e.g., histograms) or any analyses involving cross-level interactions in this article given the relatively low precision of school-specific estimates and limited sample size; pooling data from this and future years will support more reliable estimation of school-level predictors of impact variation.[9]

Analyses of Implementation Survey and Interview Data

At least one survey item from one or more respondent groups was used to evaluate each standard for program implementation, and response data were used to calculate indices of implementation fidelity, represented as percentage of standards met. Recordings from all interviews and focus groups were externally transcribed, and all identifying information was removed. An initial coding scheme was created that aligned with the evaluation research questions, the RR *Standards and Guidelines* (RRCNA, 2009), and RR's underlying theory of action (see Figure 1). A priori codes were used to focus on the broad dimensions of RR implementation, such as recruitment, training, and lesson instruction; subcodes were also used to explore specific aspects of these dimensions. A subset of interviews was first coded individually by multiple trained members of the research team, and the results were compared to establish interrater reliability. When team members disagreed on how a code should be applied, the code definition was reevaluated and sometimes revised. After the code sets were established, team members individually coded transcripts using Dedoose, a cloud-based application for analyzing qualitative and mixed-methods data that uses strong encryption to ensure data security. Each researcher developed initial analytic memos during the coding process and shared the memos with team members in weekly meetings (Corbin & Strauss, 2008; Emerson, Fretz, & Shaw, 1995; Lofland, Snow, Anderson, & Lofland, 2006). This process generated inductive themes from each set of participant interviews. A second round of analysis examined themes across groups to triangulate perspectives across respondents and data sources (e.g., comparisons to survey results), find common threads, and gain a broader perspective of implementation under the i3 scale-up. Quotes for inclusion in the results were selected as the most clear, concise, and illustrative statements relevant to a given theme, with explicit attention paid to selection of quotes that reflected the variation and/or consistency of perspectives across the full set of respondents.

Results

This section is organized as follows. We first report results from the RCT experiment, including study participation rates and final impact estimates. We then present findings from our implementation study in two sections that integrate data and results from the surveys and interviews. The first section focuses on the implementation and perceived quality of training RR teachers experienced through formal courses run by teacher leaders, while the second section focuses on the implementation of the intervention as specified in the *Standards and Guidelines* (RRCNA, 2009).

RCT Participation Rates

Of the 184 schools selected to participate in the RCT during the 2011–2012 school year, 158 schools (86%) actually carried out the random assignment process. The 26 schools that participated in the i3 study but did not carry out random assignment were excluded from the RCT. Although several modes of direct and indirect communication were used to inform schools of their expected participation in the RCT, these 26 schools did not randomize. The communications included direct emails from IDEC to individual teacher leaders and RR teachers, distribution of documents describing the evaluation design to UTCs and teacher leaders, and inclusion of a video on the IDEC website describing the evaluation design and procedures. The reasons why these 26 schools failed to randomize included staffing changes, data errors (e.g., duplication of schools in the list of participating schools), and miscommunication (e.g., undelivered emails). There was only one case in which the principal of the school explicitly refused to allow random assignment. We found no significant differences between these 26 schools and the other 158 schools in terms of urbanicity or demographics.

Figure 2 is a flow diagram that illustrates how the final analytic sample for the RCT was achieved. Within the 158 schools that participated in the RCT, a total of 1,253 students were randomly assigned to treatment (N = 628) and control (N = 625) conditions. Of these students, a total of 1,241 (622 treatment, 619 control) were successfully matched to data recorded in IDEC. Of these students, a total of 1,002 had ITBS scores recorded in IDEC (530 treatment, 472 control). After linking treatment students to their matched controls, a total of 433 matched pairs in 147 schools included a treatment and a control student with ITBS data. This sample of 866 students (433 matched pairs) represented 69% of the students from matched pairs in schools that carried out the random assignment. The missing data at the student level resulted primarily from student mobility or other factors that led to the inability or failure to administer the ITBS tests to both treatment and control students in a pair. Matched pairs without complete data are not included in the impact analyses presented here. The multisite, matched-pairs design of this random assignment study means that each school and each pair is

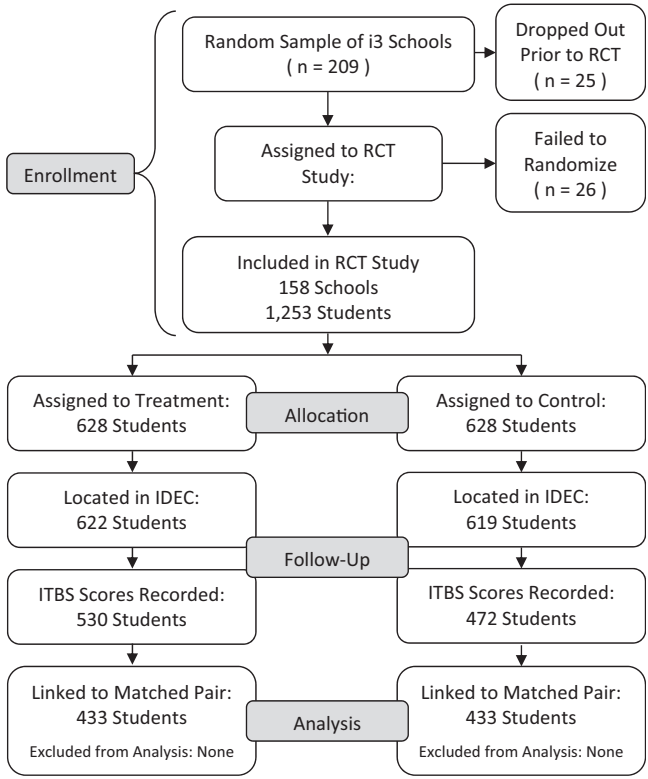

Figure 2. **Consort flow diagram through the Reading Recovery Investing in Innovation (i3) randomized controlled trial, 2011–2012.**

an independent mini-experiment and that the ability to calculate valid causal impacts is less likely to be affected by school nonparticipation or missing data. Although the sample size is reduced and may be less representative of the full population, there is no possibility of differential attrition given that equal numbers of treatment and control students are excluded from analyses (see WWC standards for attrition in RCTs; WWC, 2012, p. 11). Of course, the ability to generalize results to the entire population of i3 schools and students is a critical goal for this impact study given that the key focus is the overall impacts of the intervention on the eligible population of students when implemented at scale. As such, we performed statistical tests of differences in student demographics for students included in the impact analyses and those dropped due to incomplete data. Analyses of differences in student characteristics for those students included and excluded from the analytic sample suggest no significant differences in pretest OS Text Reading

Table 1
Baseline Balance Tests for Student Demographics

Pretreatment Variable	Treatment Group (%)	Control Group (%)	*p* Value for Difference
Gender (n = 862)			
Male	61	61	.96
Female	39	39	
ELL status (n = 860)			
ELL	17	18	.47
Non-ELL	83	81	
Race (n = 856)			
Black	18	19	.93
Hispanic	22	20	
White	57	56	
Other	3	5	
Text reading level			
0	51	52	.87
1	21	19	
2	17	20	
3+	11	10	

Note. p values based on χ^2 test of independence. ELL = English language learner.
Source. Reading Recovery International Data Evaluation Center (IDEC) student demographic and test score data.

Levels, $F(5, 1235) = 1.73$, $p = .13$; gender, $F(5, 1216) = 0.01$, $p = .91$; race, $F(3, 1183) = 0.52$, $p = .67$; or ELL status, $F(1, 1207) = 0.42$, $p = .52$.[10]

Impacts on Student Reading Achievement

Baseline balance tests were performed in order to examine whether the treatment and control groups were equivalent on observed characteristics after random assignment. Table 1 presents results for baseline balance tests for student demographics and prior reading performance of the final analytic sample of 866 students in 147 schools.

No significant differences were found between treatment and control groups on gender, ELL status, race, or prior reading performance. The percentages in each column match up well between the treatment and control groups, suggesting that random assignment produced treatment and control groups that were well balanced immediately prior to implementation of RR for the group of treatment students. This also confirms that the treatment and control groups had initial reading performance that was nearly identical immediately prior to implementation of RR.

Table 2 shows simple descriptive statistics for the treatment and control groups on raw scores and scale scores from the reading sections of the ITBS.

Table 2
Descriptive Statistics for Iowa Tests of Basic Skills (ITBS) Scores for Treatment and Control Groups

Midyear Outcomes	Treatment Group (n = 433)	Control Group (n = 433)	t Statistic and p Value
ITBS Reading Words raw scores			
Mean	21.1	18.2	7.89
Standard deviation	5.1	5.2	<.001
ITBS Comprehension raw scores			
Mean	9.9	7.8	8.11
Standard deviation	3.9	3.2	<.001
ITBS Reading Words Scale scores			
Mean	141.2	136.7	7.73
Standard deviation	9.0	7.6	<.001
Mean percentile rank[a]	43	27	
ITBS Comprehension Scale scores			
Mean	140.0	135.5	7.88
Standard deviation	8.9	7.4	<.001
Mean percentile rank[a]	39	19	
ITBS Total Scale scores			
Mean	139.2	135.0	8.69
Standard deviation	7.6	6.2	<.001
Mean percentile rank[a]	36	18	

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

For both sets of scores, the means are over one-half of a standard deviation larger in the treatment group. Differences in percentile ranks are +16 for Reading Words, +20 for Reading Comprehension, and +18 overall.

Results from the main HLM analysis are presented in Table 3. Analyses of impacts on ITBS Total Reading scores showed a significant positive effect of Reading Recovery overall. The point estimate for the difference between treatment and control students' expected Total Reading scores on the ITBS was 4.24 points ($p < .001$). Dividing that point estimate by the standard deviation of the control group yields a Glass's Δ effect size of 0.69 standard deviations. This effect estimate reflects the impact of RR relative to the population of struggling readers eligible for RR in participating schools. Alternatively, dividing the point estimate by the standard deviation from the ITBS 2005 national norming sample of first graders (i.e., σ = 9.1) yields a Cohen's d effect size of 0.47 standard deviations. This effect estimate reflects the impact of RR relative to the full population of all first graders across the nation. The WWC Improvement Index (see note 2) from this analysis is +25 percentiles, which is slightly smaller than the +32 included in the WWC report for Reading Recovery.

Table 3

Hierarchical Linear Modeling Analysis of Overall Treatment Effects of Reading Recovery on Iowa Tests of Basic Skills (ITBS) Composite Reading Scores

Dependent Variable: Midyear ITBS Total Reading Scores	Estimate	Standard Error	*p* Value
Fixed effects			
Intercept (β_0)	135.13	0.34	<.001
Pretest (β_1)	1.80	0.21	<.001
Treatment effect (β_2)	4.24	0.46	<.001
Random effects			
Matched pair variance (ω^2)	3.85	1.99	.026
School intercept variance (τ^2)	6.04	2.13	.002
School treatment impact variance (ξ^2)	9.33	3.99	.010
School intercept/impact correlation (ρ)	–0.04	0.29	.898
Treatment student-level residual variance (σ_T^2)	31.55	3.43	<.001
Control student-level residual variance (σ_C^2)	24.15	2.88	<.001

Exploratory analyses of impacts on the ITBS Reading Words and Reading Comprehension subscales showed similar results. The point estimate for the difference between treatment and control students' expected Reading Words scores on the ITBS was 4.52 points ($p < .001$) with a corresponding Glass's Δ of 0.61 and a Cohen's d of 0.45 standard deviations. The point estimate for the difference between treatment and control students' expected Reading Comprehension scores on the ITBS was 4.56 points ($p < .001$) with a corresponding Glass's Δ of 0.60 and a Cohen's d of 0.44 standard deviations.

The significant variance components for random effects in the HLM models of impacts on ITBS scores suggests that the magnitude of the impact estimates of RR varies substantially across schools. The results for the overall impact model show an average effect of +4.24 points, with a random effect variance estimate for the school-level impacts of 9.64 points. Taking the square root of this variance estimate yields a standard deviation of 3.1 points. Unfortunately, this information cannot be used to make inferences about the relative prevalence of small/large or positive/negative effects without imposing assumptions about normality on the true site-specific impacts (May, 2014). This is an area of ongoing methodological development, and our hope is to use data from this study to attempt to separate the distribution of true impacts from the normally distributed sampling error that dominates each individual site-specific estimate.

Implementation and Perceived Quality of Training

The next two sections of results describe early results from our study of implementation of Reading Recovery under i3. Although sample sizes from

the first year of this study are still too small to explain variation in impacts across sites, the general findings provide a good picture of whether Reading Recovery under this large and rapid scale-up has been faithful to the intended program design.

Behind-the-Glass Sessions

Perhaps the most crucial component for successful implementation of Reading Recovery is the intensity and quality of training provided to new RR teachers by their teacher leaders, including the behind-the-glass sessions that serve as the foundation for building teachers' instructional skills. Survey responses from RR teachers indicated that 87% of RR teachers in training met the minimum requirement of at least three lessons taught behind the glass. However, there was considerable variability in the number of behind-the-glass sessions observed by RR teachers in training: Only 25% observed more than the 30 behind-the-glass sessions required in the *Standards and Guidelines* (RRCNA, 2009), and 35% reported observing 20 or fewer behind-the-glass sessions. This suggests that while nearly all teachers taught behind the glass often enough to meet RR requirements, a majority of the RR teachers training in 2011–2012 fell short of the required 30 to 36 observations of other teachers' behind-the-glass sessions. Despite these results, behind-the-glass sessions were perceived as a particularly valuable means of training for work with students for both the RR teacher behind the glass and those teachers observing. Teachers who go behind the glass receive immediate feedback and suggestions on an actual lesson with a student and are able to make adjustments to their lessons the very next day. One RR teacher explained how the behind-the-glass sessions are different from more typical teacher observations:

> When you're observed, it's not a contrived, really perfect lesson, but it's actually what you do every day, and you're getting really honest feedback on the things that others see that are working well for the student you were working with. And there are things you may want to try to continue to accelerate their progress. . . . But they're really positive experiences. And I think those conversations that we had as colleagues just were very in-depth in regards to what the students have as far as skills and what they need to accomplish. So I appreciated it very much.

For those teachers who are observing, the behind-the-glass session offers an opportunity to provide constructive feedback to the teacher they observe while learning new skills or strategies they can use themselves. One RR teacher described how observing a behind-the-glass session helped give her new ideas:

> Sometimes you do your Reading Recovery lesson, you always assume you have to be perfect. So, when you watch someone doing

> [something different], you think "Oh, I will try this. Oh, what they're trying is great! I should try it in my next lesson."

Teacher Leader Caseload

Survey responses by teacher leaders indicated considerable variability in the number of RR teachers they supported and monitored during the 2011–2012 school year. The RR *Standards and Guidelines* (RRCNA, 2009) note that operating sites should "limit the number of teachers supported and monitored by the teacher leader to 42, or considerably fewer, depending on factors such as distance, the number of teachers per school, and the number of districts" (Standard 2.24). Survey data indicated that 46% of teacher leaders supported 1 to 20 RR teachers, and 28% supported 21 to 42 RR teachers. Nine percent of teacher leaders exceeded the recommended caseload. It is important to note that these survey responses include the total number of RR teachers—those currently in training as well as those previously trained—being supported by a teacher leader.

Quality of Training and Support

The majority of RR teachers described their teacher leader as an instructional coach who helped develop their individual skills by modeling lessons and providing immediate feedback following observations—two types of support that allowed teachers to make concrete improvements to their teaching. Another RR teacher commented:

> She's a lot like a coach, where she kind of coaches us along. Sometimes the Reading Recovery itself is all about having a constructivist approach. Which means, you kind of let the learner come to the learning themselves. You kind of just lead them there. And so with that in mind, that's kind of how they approach it with us as well. They want us to discover these things and, you know, lead ourselves into understanding.

Several RR teachers mentioned the value of being able to immediately apply what they learned in their training class to their RR lessons the same week, which allowed them to practice their new skills before the next week's training session. One RR teacher explained: "We'll learn something or study an article and then we'll try to implement something within our teaching strategies for that upcoming week." She, and other RR teachers, commented that the opportunity to learn new content and apply it to the RR lessons each week was important as they tried to hone their skills while continuously working with students.

RR teacher training occurs in a cohort model, with groups of RR teachers training together at the same pace. Most RR teachers indicated in interviews

that they felt a sense of community and camaraderie develop among their colleagues during the training process. At school sites where multiple RR teachers worked and trained together, the RR teachers remarked on the benefit of having opportunities to learn from frequent demonstrations and observations. One RR teacher commented: "I like the fact that my teaching partner and I are in the same classroom so we hear each other and that informs our teaching and we ask each other questions." By exchanging feedback, RR teachers at the same school were able to address their questions and concerns frequently.

Nearly 70% of all RR teachers started working with RR students within two weeks of the first day of school. Most expressed that their first few months implementing RR were particularly challenging since they were simultaneously teaching RR students while learning RR instructional methods. The pace of training meant that sometimes teachers encountered instructional tasks or issues that had not yet been covered in class. One teacher explained that each component of Reading Recovery was gradually introduced to trainees; she learned about the Observation Survey first and later how to take a running record. Though most of the time the pacing worked, sometimes the training objectives did not necessarily correspond with her students' progress:

> You now are adding this part and now are adding this part. So it started off we were just reading with them. Then we were reading and writing. Then we were reading off the books, writing, and doing a running record. Each couple weeks we'd add a new component, and I think for me, I needed to see all that as a whole before I jumped into it. So if we were able to maybe start the class in August versus September, go all through August to learn all that stuff, to see what it really looks like so that when I get into it, I really have an idea of what it should be. . . . That's the kind of learner I am though.

It was not until this teacher saw a behind-the-glass session later in her training that she was able to put everything together and see how the full program is designed to be implemented:

> It was very hard for me until I saw people teaching behind the glass, which obviously didn't start right away. When I saw that then it all pulled together. I was like, "Oh ok, I get it now." But I never had that model. To see what a full lesson would have looked like. And it's different reading it in a book versus really seeing that kind of interaction.

While most RR teachers said the simultaneous learning and applying of RR strategy felt like "trial by fire," they still conceded it was the best way to cement their understanding of the program. As one teacher commented:

> I definitely think, because you are a practicing RR teacher during the training year, the first couple of months are a little rough, because

> you're trying to learn the process as you are simultaneously teaching. However I don't think there's another way to do it because you have to do it while learning it in order for things to cement and for you to learn them.

Overall, RR teachers described their experience with and perception of the training they received in positive terms, and most felt the training was high quality. A vast majority of RR teachers reported feeling very well prepared for their RR responsibilities, such as planning and conducting lessons and reporting data. One RR teacher explained:

> I feel really well prepared. This course is really intense and I would say I learned more in this one class than I have in my entire master's program to become a Reading Specialist. . . . I feel like I've learned more about the reading and writing process and what kids might be doing and what they might not be doing and where they might be getting stuck and how I can push their learning forward.

This sentiment was reiterated throughout much of the interview data and substantiated through the RR teacher survey data. The survey data showed that RR teachers' feelings of preparedness after training were generally high but varied by task. For example, 99% of RR teachers reported feeling either adequately, well, or very well prepared to administer the Observation Survey at the beginning of the year. Most RR teachers also felt that their first-year training prepared them to conduct the numerous tasks that are related to conducting a lesson—over 90% responding that they felt adequately, well, or very well prepared for each of the nine elements included in one-to-one RR lessons.

Over the course of their training year, RR teachers felt they gained confidence and developed skills through their work with students. Several RR teachers reported having very different experiences with their first and second cohort of students. "Now that I'm working with my second-round students," commented one teacher, "I feel like I'm doing a better job with them than I did with my first round just because with the first round you're trying to learn so much." Another RR teacher saw the second cohort of students as an opportunity to apply all of the knowledge she gained by working with the first:

> So now just looking back, I'm going to do things differently this time. After being in training since September, I know what works better. Now I get a second chance with this other group of kids.

RR teachers reported that with their second group of students, they had a better sense of what the overall program and individual lessons should entail. A few RR teachers even expressed an interest in keeping their first cohort of students for longer than the allotted time, maybe even the whole

year, to compensate for the amount of learning they had to do at the beginning of their training. This way the first cohort of students is not "short-changed" by a teacher's developing skill. Another RR teacher suggested changing the structure of the training year so that a teacher sees only one or two students at the beginning of the year, then eases into a full load of students once the teacher is more comfortable.

In summary, RR teachers reported very positive perceptions of their training and preparation, with nearly every teacher reflecting on substantial improvements in their skills as a literacy teacher and changes in their perspectives on literacy instruction in general. However, while building teachers' capacity is a major cornerstone of Reading Recovery, there are many additional practices, policies, and routines described in the *Standards and Guidelines* (RRCNA, 2009) that define how the Reading Recovery intervention is intended to be implemented.

Intervention Implementation Fidelity

Overall implementation fidelity results based on surveys of RR teachers, teacher leaders, and site coordinators are presented in Table 4. First reported are numbers of respondents and response rates calculated based on the population sizes that we received from IDEC. Response rates for RR teachers, teacher leaders, and site coordinators were 74%, 88%, and 58%, respectively (note that these response rates are slightly lower than the overall survey response rates due to item-level nonresponse). The relatively low response rate for site coordinators is likely due to the fact that their involvement in Reading Recovery in much less intense than that of teacher leaders and RR teachers (hence a lower incentive to respond) and that no monetary response incentives were provided to any of these groups. Table 4 also presents the percentage of standards that were met by respondents, overall and by subdomain. Results show a high percentage of standards met, suggesting that the Reading Recovery model is being implemented with high fidelity by most RR teachers, teacher leaders, and site coordinators. On average, RR teachers met 95% of the standards, while teacher leaders and site coordinators met 87% and 88% of the standards, respectively.

The selection of RR teachers and teacher leaders was found to have complete fidelity, and near complete fidelity was reported for standards that pertain to conducting Reading Recovery lessons, suggesting that, as reported by RR teachers, the one-on-one lessons are being implemented as required. Data from our teacher logs (n = 2,623) confirm that all of the nine structural components of RR lessons were included in virtually every one-to-one lesson. The activity most likely to be skipped was Assembly of a Cut-up Story, although it was skipped in only 10% of the lessons. Dozens of observations of the delivery of RR lessons in a sample of nine i3 scale-up schools confirmed this finding; strong fidelity to the standard

Table 4
Implementation Fidelity of Investing in Innovation (i3) National Scale-Up of Reading Recovery (RR)

	RR Teachers	Teacher Leaders	Site Coordinators
Number of respondents	742	168	95
Response rate (%)	73	88	58
Percentage standards met			
Overall (% of standards met)[a]	95	87	88
Reading Recovery lessons	98		
Selection	100	100	
Training	81	73	
Data and monitoring	99	100	
Leadership		94	
Research		68	
Professional development		98	68
Background			81
Site preparation and maintenance			96
Communication			79

Source: Surveys of RR teachers, teacher leaders, and site coordinators.
[a]Overall percentages are calculated as unweighted averages across all standards. Given that the number of standards in each domain varies, these percentages are not a simple average across domains.

format of lessons was observed, even in schools that struggled with other aspects of implementation. Across the 10 categories of standards, training was the category for which RR teachers and teacher leaders reported the lowest level of fidelity. Specific issues included training classes that were smaller than the minimum eight RR teachers and conducting fewer than the required 18 training sessions that included two behind-the-glass lessons. In the area of research, results suggest that about one-third of teacher leaders did not submit an annual site report to their UTC. Site coordinators reported high fidelity to site preparation and maintenance (i.e., fiscal responsibilities, program monitoring, tasks related to training and oversight of personnel, and resource provisioning). The areas with lowest fidelity for site coordinators were professional development and communication (i.e., 68% and 79%, respectively), which implies that some site coordinators may not be as invested in learning about and advocating for Reading Recovery as the *Standards and Guidelines* (RRCNA, 2009) suggest.

One area of implementation where we observed substantial variation was in the selection of students for one-to-one intervention. Reading Recovery strives to serve the students at greatest risk of lifelong low literacy, an aspiration that is operationalized in the *Standards and Guidelines* (RRCNA, 2009) by the suggestion that schools provide the intervention to

at least the lowest achieving 15% to 20% of a first-grade cohort. The *Standards and Guidelines* also indicate that the selection of students to participate in the intervention should be guided by Marie Clay's assertion that "all kinds of children with all kinds of difficulties can be included" (Clay, 1991, p. 60). Specifically, Clay's recommendation states that:

> exceptions are not made for children of lower intelligence, for second-language children, for children with low language skills, for children with poor motor coordination, for children who seem immature, for children who score poorly on readiness measures, or for children who have been categorized by someone else as learning disabled. (Clay, as cited in RRCNA, 2009, p. 8)

To facilitate this inclusive selection process, the *Standards and Guidelines* (RRCNA, 2009) require schools to use only the OS to select the lowest achieving first-grade students and to serve the lowest scorers first. We examined student selection through interviews with RR teachers, teacher leaders, first-grade classroom teachers, and principals and observed considerable variation in schools' implementation of this requirement.

While a minority of schools in the scale-up administered the OS to all first graders in order to assign the lowest scorers to Reading Recovery, our exploration revealed that most schools used a two-tiered selection process. Typically, schools first identified a preliminary pool of low-achieving students and then selected the students who ultimately received the intervention from within that pool. Much of the variability observed around student selection occurred in the first tier of this process, with the identification of a pool of prospective participants.

Variations in this first tier of the student selection process included school-to-school and occasionally district-to-district differences in which school staff members were involved in nominating students to the preliminary pool, the criteria on which these nominations were based, and the level (classroom, school, or district) at which the first-grade cohort was defined. In most schools, the first-round identification process was collaborative, informed by input from first-grade and often kindergarten teachers. Student identification teams also generally included interventionists, ELL and special education teachers, instructional coaches, and/or school administrators. Typically, schools nominated students to the preliminary pool based on reading levels determined by commercially available leveling systems or based on assessment data from teacher-administered tests. However, in a number of schools, students were nominated to the pool based on classroom teachers' general observations or "gut" impressions of their needs. An additional source of variation in the first tier of the selection process was inconsistency in the level at which the first-grade cohort was defined. In some schools, teams looked across all first-grade classrooms to identify the lowest first graders in the whole school, even if they were distributed

unevenly across classrooms. In other cases, each classroom teacher was asked to nominate a predetermined number of students, regardless of whether they were among the lowest first-grade students in the school overall. In a minority of cases, students were nominated to the preliminary pool at the district level, resulting in some schools having more candidates for Reading Recovery lessons than others.

We found more consistency in the second tier of the selection process—the selection of students to receive the intervention from the preliminary pool. Consistent with the *Standards and Guidelines* (RRCNA, 2009), interview participants reported that RR teachers were generally responsible for selecting students from the preliminary pool and that they generally selected those who scored lowest on the OS. However, some departures from the *Standards and Guidelines* were observed at this stage as well. For example, in some cases the OS was not the only indicator of students' performance used for the final selection of students. Some RR teachers reported considering other student data—from other assessment scores to attendance records—as well, generally at the insistence of a school or district administrator. There was also variation reported in teacher leaders' involvement in the selection process: Many RR teachers described consulting with their teacher leaders on the interpretation of OS scores or on choosing between students with similar scores. A minority of RR teachers reported that the final selection of students rested entirely in the hands of their teacher leaders.

One important area of variation that emerged through our examination of student selection was the extent to which schools excluded particular groups of students from receiving Reading Recovery. For instance, several schools excluded students with chronic truancy or disruptive behaviors. Decision makers at these schools, respondents reported, preferred to reserve Reading Recovery slots for students they regarded as more likely to benefit from the intervention. Describing her school's selection process, one RR teacher noted that first graders who had 10 or more absences in kindergarten were automatically excluded, resulting in the exclusion of 4 of the school's 12 lowest readers that year from Reading Recovery. Another teacher remarked: "We have two students who are [first grade] repeaters who really need it, but that's something we don't do. They're like the lowest in the whole grade now and we can't serve them."

A significant number of teachers reported that their school or district policies, or their own understanding of Reading Recovery policy, required the exclusion of students who had Individualized Education Plans (IEPs) that included a literacy intervention or who were repeating first grade, even if they were among the lowest scorers on the OS. One teacher explained:

> If a student has an IEP with reading goals on it, they can't be in Reading Recovery. And we have one little girl that was too low to even fit in an LLI [Leveled Literacy Intervention] group, so she was

> a prime candidate for Reading Recovery, but . . . she has an IEP and gets reading services so we couldn't take her, and we wanted to. And there was a boy, we had done the second-time-around Observation Survey on him and we were getting ready to take the second round [of students for one-to-one lessons]. My teaching partner went to go look in his file for something and found that he had been retained in first grade. . . . And if someone is a retainee, you know, from first grade, they don't qualify for Reading Recovery.

There may simply be insufficient clarity regarding Reading Recovery's policies about the inclusion or exclusion of students with IEPs. The *Standards and Guidelines* (RRCNA, 2009) state that all students should be served regardless of disability; however, other Reading Recovery policy documents specifically exclude those with IEPs for literacy if (and only if) they are already receiving "special help in reading" (see RRCNA, 2002). RR teachers themselves and other school-level implementers also described varying understandings of RR's policies in this area. For instance, while some RR teachers reported that their teacher leaders instructed them not to select students with IEPs, others reported that their teacher leaders insisted that no student be excluded. This variation in perspectives on eligibility for Reading Recovery is a major focus in this ongoing evaluation. At this point, while it is clear that exclusions are often made, the vast majority of these are because excluded students are already receiving some other literacy intervention.

Discussion

Our discussion of results from Year 1 of this large-scale randomized evaluation of Reading Recovery under the i3 Scale-Up focuses on the strength of causal inference supported by this study in relation to WWC standards and also on two critical components inherent in Reading Recovery's theory of action (see Figure 1). The first of these is whether and how the training provided to Reading Recovery teachers builds teachers' capacity to deliver intensive literacy instruction that is responsive to individual student's needs. The second is whether the intervention is being delivered as specified in the Reading Recovery *Standards and Guidelines* (RRCNA, 2009) and when adaptations are made, what are the reasons and justifications and who influences these decisions?

Causal Effects of Reading Recovery Under i3

The What Works Clearinghouse (2012) standards of evidence require (a) random assignment to treatment and control conditions (preferably with pre-intervention data to confirmation of baseline equivalence) and (b) low overall and differential attrition from the study. This study clearly meets the random assignment requirement and confirms baseline equivalence on pre-intervention outcomes. The overall student-level attrition from the study

was 31%, and differential attrition was 0% given the decision to drop incomplete matched pairs from the analytic sample.[11] This is well below the WWC's limit of 56% overall attrition when differential attrition is zero (WWC, 2012, p. 12). According to WWC, this level of "attrition is expected to result in an acceptable level of bias even under conservative assumptions, which yields a rating of *Meets Evidence Standards*," suggesting that the quality of causal inference from this study is quite strong.

The estimated impact of Reading Recovery on students' ITBS Total Reading scores was .69 standard deviations relative to the study sample and .47 standard deviations relative to the national population of first graders. Although the effects were slightly smaller than those reported in the WWC 2008 review of Reading Recovery, these standardized effect sizes are large relative to typical effect sizes found in educational evaluations. For example, the impacts of Reading Recovery are up to 5.7 times larger than the average effects of Title I programs reviewed by Borman and D'Agostino (1996; average weighted effect size of .11). Gains in percentile rank scores were also large, with treatment students outperforming control students by up to 20 percentile points. When compared to typical gains of first graders on the ITBS tests, the additional gains experienced by Reading Recovery students is analogous to an additional 1.9 months of learning or a growth rate that is 38% greater than the national average growth rate for beginning first graders.[12]

Developing Teachers' Instructional Skills

A critical element to the faithful implementation of Reading Recovery is the intensive training RR teachers receive during their first year. The goal of this training is to enhance teachers' skills for individualized early literacy instruction, assessment, systematic observation, analysis, responsive instruction, reflection, and planning (see Figure 1). RR teachers interviewed felt that the immediate feedback and intensive support provided by teacher leaders during the training was very effective in enhancing these skills. RR teachers felt that the one-to-one interactions with the teacher leaders, observations and reflections on actual lessons, and interactions with other RR teachers helped them make concrete changes to their RR teaching. Many RR teachers reported that their RR training was transformative in terms of their own instruction and understanding about literacy. Additional survey data confirmed that nearly all RR teachers felt that their training experience and work with their teacher leader prepared them for RR implementation and teaching RR students.

Fidelity to RR Implementation ***Standards and Guidelines***

Behind-the-Glass Training

The behind-the-glass sessions were reported by RR teachers to be one of the most valuable aspects of their training experience. However, not all teachers are getting as much experience with behind-the-glass training as

is required by the *Standards and Guidelines* (RRCNA, 2009). While 87% of new RR teachers reported teaching behind the glass at least three times (as required by the RR *Standards and Guidelines*), the vast majority reported observing other teachers teach behind the glass fewer than the 36 required times. This may have been the result of smaller than usual training classes or the great distances some RR teachers in training had to travel to attend behind-the-glass sessions. The implications of this are not yet clear. Would the skills of RR teachers in training have been developed more substantially if they had observed more behind-the-glass lessons, or is a lower standard acceptable? Is the requirement to observe at least 30 to 36 sessions during the training year (i.e., nearly one per week) unrealistic under this scale-up? These are questions that we hope to answer in subsequent years of this study.

Quality and Structure of Lessons

School-level implementation of RR was, in most respects, faithful to the Reading Recovery *Standards and Guidelines* (RRCNA, 2009). Data from teacher logs and lesson observations confirmed strong fidelity to standards in the execution of RR lessons. Lessons were conducted with regularity, and the standard format of lessons was followed in the vast majority (>90%) of one-to-one lessons. Lower fidelity was observed related to requirements for program documentation and communication, although the majority of standards in these domains were still met.

When identifying students eligible for RR intervention, schools and districts varied in their selection processes. While most schools used scores on the OS to identify low-performing students, the students ultimately selected to receive RR were not always the lowest performing students in the first grade. Where the selection of students with IEPs is concerned, our key finding is that while many schools are not adhering to a strict interpretation of the *Standards and Guidelines* (RRCNA, 2009) by including all of the lowest performing students in Reading Recovery, the vast majority of students who are excluded from Reading Recovery are deemed ineligible because they are already receiving other literacy intervention services. In general, most schools are making a good-faith effort to comply with their understanding of RR policy and with the goal of ensuring that all of the lowest achieving students in their schools receive intensive services of one kind or another. While our findings from this first year indicate that variation around student selection is prevalent and that it affects which students ultimately participate in RR, our findings also highlight a lack of clarity around the goals and policies of the intervention. If students who are already participating in another literacy intervention should not be eligible for Reading Recovery (see RRCNA, 2002), then why is this not clearly stated in the *Standards and Guidelines* (RRCNA, 2009)? Nevertheless, the students selected for Reading

Recovery are, across the board, very low-achieving students. Although they may not always be the students with the absolute lowest scores, the students selected to participate in Reading Recovery are clearly and consistently among the most challenged first-grade readers in their schools.

Study Limitations

There are at least two notable limitations to this study. First, although the RCT in this evaluation provides strong causal inference about the impacts of Reading Recovery for the sample of schools and students that were included in the final analytic sample, the degree to which these results generalize to the full population of i3 schools and students or to the greater Reading Recovery community beyond i3 is not guaranteed. Although we found no significant differences in demographics of participating and nonparticipating schools and students, we cannot conclude that there are no differences on other unobserved variables.

The second notable limitation is that the sample size within each school is very small. With only four treatment students and four control students, estimates of site-specific impacts are incredibly noisy. Simulations for post hoc power and precision suggest that within-school sample sizes would need to increase dramatically (e.g., >100) in order to produce site-specific impact estimates that are reliable enough to differentiate schools with truly large or small impacts of Reading Recovery (e.g., $r > .60$). With our current sample size of eight students per school, variation in site-specific impacts is driven primarily by sampling variability. Fortunately, by pooling data across years one through five of the study, we will produce an analytic sample that includes more than 1,000 schools and several thousand students who will have been randomly assigned. This very large school-level sample size may be sufficient to detect correlations between site-specific impact estimates and implementation measures despite the imprecision in each school's estimate. At this point, we don't yet have enough linkable data on impacts and implementation in individual schools to explain variation in impacts, but this is an essential goal for the final years of this evaluation.

Implications for Policy and Practice

The quality of implementation and large positive effects of Reading Recovery during the first year of this i3 scale-up suggest that the $55 million investment is paying off. Although more specific cost-effectiveness results won't be available until the final year of the project, these early results are very encouraging. It is now up to the Reading Recovery community to ensure that the new RR programs continue to succeed and are sustainable after the end of the i3 project. If the scale-up of Reading Recovery endures, this will provide tangible support to the idea that federal investments in interventions that are

contingent on rigorous evidence (e.g., the i3 fund) can produce widespread and significant changes in educational practice and outcomes.

There are two critical implications for the Reading Recovery community. Low fidelity in training due to small training classes and having too few behind-the-glass observations may simply reflect standards that are unrealistic or too constraining. Is it acceptable to have a training class with 5 teachers instead of 8 to 12? Is it reasonable to expect that each teacher observe 30 to 36 behind-the-glass sessions, or is 20 to 25 per year sufficient? Lastly and perhaps most importantly, the lack of consistency among RR staff in their understanding of policies for student selection and exclusion suggests that the RR community should work to make these policies more clear. For example, is it reasonable to exclude students from Reading Recovery if they are already participating in some other literacy intervention, but not simply because they have an IEP or a history of absenteeism or behavior problems? These issues should be addressed more explicitly in the *Standards and Guidelines* (RRCNA, 2009).

Notes

Funding Sources: Institute of Education Sciences (Grant/Award Number: #R305B090015); Office of Innovation and Improvement (Grant/Award Number: #U396A100027).

[1]To meet What Works Clearinghouse (WWC) standards "without reservations," a study must use random assignment to treatment and control conditions and not exceed WWC's thresholds for differential and overall attrition of study participants. Studies that meet WWC standards "with reservations" include (A) random assignment studies with high attrition and (B) nonrandomized designs, provided that A and B demonstrate baseline equivalence of the groups in the analytic sample. For more details, see the What Works Clearinghouse Procedures and Standards Handbook, Version 3.0 (What Works Clearinghouse, Institute of Education Sciences, 2014) available from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.

[2]The WWC Improvement Index "can be interpreted as the expected change in percentile rank for an average comparison group student if the student had received the intervention" (What Works Clearinghouse, Institute of Education Sciences, 2014, p. 23).

[3]The "blocking" of students in matched pairs was intended to address the variability of the length of the intervention cycle for students. Blocking students into pairs ensured that the outcome for each treatment student was compared to the outcome for a control student who experienced the counterfactual for the same length of time as the treatment. In addition, blocking is a mechanism that increases the likelihood of the baseline equivalence of treatment and control in regards to pretest scores (i.e., text reading levels) and English language learner (ELL) status. Lastly, the use of matched pair blocks allows the integrity of the random assignment design to remain uncompromised (i.e., on observed covariates) by missing outcome data. By dropping all matched pairs with missing outcome data from the analytic sample, there is no risk of differential attrition of students across treatment and control groups with regard to those variables used to create the blocks (i.e., pretest and ELL status). Furthermore, comparisons of students included and excluded from analyses (see section entitled "RCT Participation Rates") confirm no differential attrition by pretest, gender, race, or ELL status. Imputation methods for missing outcome data are not used because they do not guarantee unbiased causal estimates if the data are not missing at random. Furthermore, the criteria for the National Evaluation of the Investing in Innovation Program (NEi3) do not permit imputation of outcome variables (see Abt Associates, 2012).

[4]Students assigned to the control condition could receive any instructional intervention other than Reading Recovery (RR). Although data on control group students'

participation in other interventions were not collected during Year 1 of the randomized controlled trial (RCT), preliminary analyses of data from Year 2 suggest that over three-quarters of control group students received non-RR supplemental literacy instruction, typically in small group settings.

[5]This design ensures that (a) the experimental contrast is maintained until the treatment students conclude their RR intervention and posttest outcomes are measured in each pair, (b) control students are not denied the intervention after posttest outcomes are measured, and (c) RR teachers are not required to provide services to students who no longer need them.

[6]Site visit schools were selected based on a random sample, stratified by region and urbanicity.

[7]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.

[8]The choice to use Glass's Δ is based on the expectation that the impact of Reading Recovery will vary across students and schools, resulting in an increase in not only mean posttest achievement but also an increase in the variance of posttest achievement scores. By using the control group standard deviation, we are better able to benchmark the impact estimate against the counterfactual (i.e., the distribution of potential outcomes in the absence of the intervention) (Cooper, 1998; Lipsey & Wilson, 2001). In addition to Glass's Δ, which represents a standardized effect relative to the distribution of outcomes for only study participants (i.e., the lowest eight students in each school who were selected for the RCT), we present a population-based standardized effect size. This Cohen's *d* effect size is calculated by dividing the raw impact estimate by the standard deviation of Iowa Tests of Basic Skills (ITBS) scores for the national norming sample. This allows 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 (see May, Perez-Johnson, Haimson, Sattar, & Gleason, 2009). The Cohen's *d* impact estimates should be smaller than the Glass's Δ estimates because the variance in outcomes for the full population of first graders is larger than the variance for struggling readers. Lastly, we calculate the WWC Improvement Index (see note 2) in order to make direct comparisons to results from the WWC review.

[9]Raw empirical Bayes estimates of site-specific impacts have two major limitations. First, they suffer from underestimated variance due to over-shrinkage (Bloom, Raudenbush, & Weiss, 2012; Raudenbush & Bryk, 2002, p. 88). Second, empirical Bayes estimates can be highly imprecise when within-site sample sizes are small (as is the case in this study), and the distribution of these effects can be biased toward normality due to measurement error. While it is possible to calculate adjusted site-specific estimates (see Bloom et al., 2012), this does not address the issue of imprecision and non-normality of true effects. As such, we do not include histograms or other visualizations of site-specific impacts in this article.

[10]Comparisons of students included and excluded from analyses are based on models similar to those used for the impact analyses (i.e., generalized linear mixed models with random effects for schools and matched pairs estimated via GLIMMIX in SAS 9.3).

[11]Even if potentially participating students from the schools that failed to randomize are included in the overall attrition, the rate is still only 41% overall.

[12]From the start of first grade through the fifth month (i.e., the period during which the treatment students received RR 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., 2006). This increase of 11 points over a 5-month period suggests that the additional gains of 4.2 points experienced by Reading Recovery students is equivalent to an additional 1.9 months of learning. Alternatively, the additional 4.2 points translates to a growth rate that is 38% greater than the national average growth rate for beginning first graders.

References

Abt Associates. (2012). *National evaluation of i3: Analysis and reporting plan, version 1.0.* Washington, DC: Author.

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

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*(1), 27–42.

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

Bloom, H. S., Raudenbush, S. W., & Weiss, M. (2012). *Estimating variation in program impacts: Theory, practice and applications*. Manuscript submitted for publication.

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.

Briggs, C., & Young, B. (2003). Does Reading Recovery work in Kansas? A retrospective longitudinal study of sustained effects. *The Journal of Reading Recovery*, *3*(1), 59–64.

Brown, W., Denton, E., Kelly, P., & Neal, J. (1999). Reading Recovery effectiveness: A five-year success story in San Luis Coastal Unified School District. *ERS Spectrum: Journal of School Research and Information*, *17*(1), 3–12.

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

Chapman, J. W., Tunmer, W. E., & Prochnow, J. E. (1999). Why Reading Recovery does not work: A longitudinal study in a whole language context. *Thalamus*, *17*, 52–53.

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

Clay, M. (2005). *Literacy lessons designed for individuals: Part one*. Portsmouth, NH: Heinemann.

Cooper, H. (1998). *Synthesizing research: A guide for literature review*. Thousand Oaks, CA: Sage.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research* (3rd ed.). Thousand Oaks CA: Sage.

D'Agostino, J., & Murphy, J. (2004). A meta-analysis of Reading Recovery in United States' schools. *Educational Evaluation and Policy Analysis*, *26*(1), 23–38.

Dyer, P. C., & Binkney, R. (1995). Estimating cost-effectiveness and educational outcomes: Retention, remediation, special education, and early intervention. In R. L. Allington & S. A. Walmsley (Eds.), *No quick fix: Rethinking literacy programs in America's elementary schools* (pp. 45–60). New York, NY: Teachers College Press.

Elbaum, B., Vaughn, S. M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, *92*(4), 605–619.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing ethnographic field notes*. Chicago, IL: University of Chicago Press.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*(1), 93–99.

Gómez-Bellengé, F. X. (2002). Measuring the cost of Reading Recovery: A practical approach. *The Journal of Reading Recovery*, *2*(1), 47–54.

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

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, IA: University of Iowa.

Iversen, S., & Tunmer, W. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology*, *80*(4), 437–447.

Johnston, P. A., Allington, R. L., & Afflerback, P. (1985). The congruence of classroom and remedial reading instruction. *The Elementary School Journal*, *85*, 465–478.

Johnston, P. H. (2011). Response to Intervention in literacy: Problems and possibilities. *The Elementary School Journal*, *111*(4), 511–534.

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.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lofland, J., Snow, D., Anderson, L., & Lofland, L. H. (2006). *Analyzing social settings: A guide to qualitative observation and analysis* (4th ed.). Belmont, CA: Wadsworth.

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, A. J. Rotherham, & C. R. Hokanson (Eds.), *Rethinking special education for a new century* (pp. 259–288). Washington, DC: Progressive Policy Institute and Thomas B. Fordham Foundation.

May, H. (2014). *Distortions in distributions of impact estimates in multi-site trials: The central limit theorem is not your friend*. Paper presented at the fall meeting of the Society for Research in Educational Effectiveness, Washington, DC.

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Moats, L. (2007). *Whole-language high jinks: How to tell when "scientifically based reading instruction" isn't*. Washington, DC: Thomas B. Fordham Institute.

Morrow, L. M. (1993). *Literacy development in the early years: Helping children read and write*. Needham Heights, MA: Allyn & Bacon.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health.

National Research Council. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy of Sciences.

Pikulski, J. (1994). Preventing reading failure: A review of five effective programs. *The Reading Teacher*, *48*, 30–39.

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., Lyons, C., DeFord, D., Bryk, A., & Seltzer, N. (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.

Rasinski, T. V. (1995). Commentary: On the effects of Reading Recovery: A response to Pinnell, Lyons, DeFord, Bryk, and Seltzer. *Reading Research Quarterly*, *20*(2), 264–270.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rodgers, E., Gómez-Bellengé, F., Wang, C., & Schultz, M. (2005). *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). *Closing the literacy achievement gap with early intervention*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Rowe, K. (1995). Factors affecting students' progress in reading: Key findings from a longitudinal study. *Literacy Teaching and Learning: An International Journal of Early Literacy*, *1*(2), 57–110.

Reading Recovery Council of North America. (2002). *A principal's guide to Reading Recovery*. Columbus, OH: Author.

Reading Recovery Council of North America. (2009). *Standards and guidelines of Reading Recovery in the United States* (5th ed.). Columbus, OH: Author. Retrieved from http://www.readingrecovery.org/pdf/implementation/Standards_Guidelines-09_Full_Version.pdf

Ruhe, V., & Moore, P. (2005). The impact of Reading Recovery on later achievement in reading and writing. *ERS Spectrum*, *23*(1), 20–30.

Schmitt, M. C., & Gregory, A. E. (2005). The impact of an early literacy intervention: Where are the children now? *Literacy Teaching and Learning: An International Journal of Early Literacy*, *10*, 1–20.

Schwartz, R. (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). A synthesis of research on Reading Recovery. *Reading Research Quarterly*, *30*, 958–996.

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*(6), 1351–1359.

Strickland, D. S. (2002). The importance of effective early intervention. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 69–86). Newark, DE: International Reading Association.

Torgesen, J. K. (2000). Individual responses to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, *15*, 55–64.

Tunmer, W. E., & Chapman, J. W. (2002). The relation of beginning readers' reported word identification strategies to reading achievement, reading-related skills, and academic self-perceptions. *Reading and Writing*, *15*(3–4), 341–358.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S., Chen, R., Pratt, A., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology*, *88*, 601–638.

Vellutino, F. R. (2010). Learning to be learning disabled: Marie Clay's seminal contribution to the Response to Intervention approach to identifying specific reading disability. *The Journal of Reading Recovery*, *10*(1), 5–23.

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly*, *28*, 179–200.

What Works Clearinghouse. (2008). *WWC intervention report: Reading Recovery*. Washington, DC: US Department of Education, Institute of Education Sciences.

What Works Clearinghouse. (2012). *Procedures and standards handbook: Version 3.0*. Washington, DC: US Department of Education, What Works Clearinghouse. Retrieved from http://ies.ed.gov/ncee/wwc/

Manuscript received January 31, 2014
Final revision received July 11, 2014
Accepted September 10, 2014