CPRE

CONSORTIUM FOR POLICY RESEARCH IN EDUCATION

Evaluation of the i3 Scale-up of Reading Recovery

Year Two Report, 2012-13

Henry May
Heather Goldsworthy
Michael Armijo
Abigail Gray
Philip Sirinides
Toscha J. Blalock
Helen Anderson-Clark
Andrew J. Schiera
Horatio Blackman
Jessica Gillespie
Cecile Sam

December 2014
RR-79

A COLLABORATIVE PUBLICATION BETWEEN
CPRE | University of Delaware | Center for Research in Education & Social Policy

The research reported here was supported by the Office of Innovation and Improvement (OII), U.S. Department of Education, through Grant #U396A100027 to The Ohio State University, and in part by the Institute of Education Sciences (IES), U.S. Department of Education, through Grant #R305B090015 to the University of Pennsylvania. The opinions expressed are those of the authors and do not represent the views of OII, IES, or the U.S. Department of Education.

About Consortium for Policy Research in Education (CPRE)

Since 1985, the Consortium for Policy Research in Education (CPRE) has brought together renowned experts from major research universities to improve elementary and secondary education by bridging the gap between educational policy and student learning. CPRE researchers employ a range of rigorous and innovative research methods to investigate pressing problems in education today. Having earned an international reputation for quality research and evaluation, CPRE researchers have extensive experience conducting experimental studies, large-scale quasi-experimental research, qualitative studies, and multi-state policy research.

CPRE's member institutions are the University of Pennsylvania; Teachers College Columbia University; Harvard University; Stanford University; University of Michigan; University of Wisconsin-Madison; and Northwestern University.

About the Center for Research in Education & Social Policy (CRESP)

The Center for Research in Education and Social Policy (CRESP) within the College of Education and Human Development at the University of Delaware conducts rigorous research to help policymakers and practitioners in education, health care and human services determine which policies and programs are most promising for improving outcomes in children, youth, adults and families.

Although research in prevention sciences and health care have long used rigorous designs to assess the effectiveness of programs, it was not until the Education Sciences Reform Act of 2002 that we witnessed a dramatic increase in the quantity and quality of research to evaluate the effects of education programs and policies. The education community began to focus on research that could measure the impact of these programs through randomized experiments and other research designs that support causal conclusions and can determine whether, how well, for whom, and why new programs and interventions work.

CRESP specializes in experimental and quasi-experimental research that uses quantitative and mixed methods to evaluate how and how well programs and interventions work to improve educational, family, and health outcomes in schools and communities.

Table of Contents

Executive Summary

Reading Recovery is a short-term early intervention designed to help the lowest-achieving readers in first grade reach average levels of classroom performance in literacy. Students identified to receive Reading Recovery meet individually with a specially trained Reading Recovery teacher every school day for 30-minute lessons over a period of 12 to 20 weeks. The purpose of these lessons is to support rapid acceleration of each child's literacy learning. In 2010, The Ohio State University received a Scaling Up What Works grant from the U.S. Department of Education Investing in Innovation (i3) Fund to expand the use of Reading Recovery across the country. The award was intended to fund the training of 3,675 new Reading Recovery teachers in U.S. schools, thereby expanding service to an additional 88,200 students.

The Consortium for Policy Research in Education (CPRE) was contracted to conduct an independent evaluation of the i3 scale-up of Reading Recovery over the course of five years. The evaluation includes parallel rigorous experimental and quasi-experimental designs for estimating program impacts, coupled with a large-scale mixed-methods study of program implementation. This report presents the findings of the second year of the evaluation. The primary goals of this evaluation are: a) to provide experimental evidence of the impacts of Reading Recovery on student learning under this scale-up effort ; b) to assess the success of the scale-up in meeting the i3 grant's expansion goals; and c) to document the implementation of the scale-up and fidelity to program standards.

This document is the second in a series of three reports based on our external evaluation of the Reading Recovery i3 Scale-Up. This report presents results from the impact and implementation studies conducted over the 2012-2013 school year–the third year of the scale-up effort and the second full year of the evaluation.

In order to estimate the impacts of the program, a sample of first graders who had been selected to receive Reading Recovery were randomly assigned to a treatment group that received Reading Recovery immediately, or to a control group that did not receive Reading Recovery until the treatment group had exited the intervention. The reading achievement of students in this sample was assessed using a standardized assessment of reading achievement–the Iowa Tests of Basic Skills (ITBS). The data for the implementation study include extensive interviews and surveys with Reading Recovery teachers, teacher leaders, site coordinators, University Training Center directors, members of the i3 project leadership team at The Ohio State University, and principals and first-grade teachers in schools involved in the scale-up. Case studies were also conducted in nine i3 scale-up schools to observe how Reading Recovery operates in different contexts.

Key findings from Year Two of this evaluation include the following:

Impacts on Student Reading Performance

- Treatment students who participated in Reading Recovery outperformed students in the control group on each subscale of the ITBS Reading test.
    - The mean of Reading Recovery students' posttest ITBS Total Reading scores was at the 36th percentile nationally, while students in the control group had posttest scores at the 22nd percentile–a difference of +14 percentile points.
    - The mean of Reading Recovery students' posttest ITBS Reading Words scores was at the 43rd percentile nationally, while students in the control group had posttest scores at the 27th percentile–a difference of +16 percentile points.
    - The mean of Reading Recovery students' posttest ITBS Reading Comprehension scores was at the 39th percentile nationally, while students in the control group had posttest scores at the 23rd percentile–a difference of +16 percentile points.
- The estimated standardized effect of Reading Recovery on students' ITBS Total Reading Scores was 0.42 standard deviations relative to the population of struggling readers eligible for Reading Recovery under the i3 scale-up, and 0.33 standard deviations relative to the nationwide population of all first graders. These standardized effect sizes are large relative to typical effect sizes found in educational evaluations.
- When benchmarked against expected gains on the ITBS for the test's national norming sample, Reading Recovery students who participated in the randomized control trial in Year Two exceeded expected growth by 3.03 points. This is equivalent to an additional 1.4 months of learning over the intervention period, over and above what beginning first graders typically achieve during that timeframe. These effects are large relative to the estimated treatment effects typically reported in studies of instructional interventions. They are 2.8 times the average effects reported in studies of similar interventions analyzed by Lipsey et al. (2012).
- Effect estimates were similarly large for both the ITBS Reading Words and Reading Comprehension subscales.
- The impact estimates of Reading Recovery vary substantially across schools, with most schools having moderate to large positive impact estimates (greater than 0.40 standard deviations).

Scale-up: Successes and Challenges

- Reading Recovery is on target to meet its scale-up goals. At the end of the third year of scale-up, 2,079 teachers had been recruited and trained in Reading Recovery. This number represents 105% of the scale-up goal for this time period (the first three years of grant funding). In addition, 23,720 students had been served with one-to-one Reading Recovery lessons by an i3-trained teacher. This number is 94% of the goal for years one through three of the scale-up. A total of 113,976 other students had been served by these teachers through classroom or small-group instruction, representing 109% of the scale-up goal for this timeframe.
- Of the 2,079 teachers trained in Reading Recovery since the start of the scale-up, 524 exited by the end of the 2012-2013 school year. This number represents a total attrition rate of 25%.
- As in Year One of this evaluation, the primary barrier to recruitment and retention in 2012-2013 was school and district administrators' concerns about the affordability of Reading Recovery in a difficult economic climate. Competing instructional priorities—including schools' growing focus on Common Core implementation—and philosophical differences with Reading Recovery's instructional approach represented additional challenges. UTC directors and teacher leaders continued to strategize new ways to address these obstacles.

Implementation: Fidelity

- As in Year One of this study, CPRE observed strong implementation fidelity overall. This leads us to believe that the impact estimates we observed in both years are trustworthy.
- A few deviations from fidelity were noted. As in Year One, the most notable of these pertained to the selection of students to receive the intervention. While multiple sources of data suggest that the students served by Reading Recovery are consistently among the lowest-achieving first graders in their schools, many Reading Recovery teachers reported that their schools' processes for selecting Reading Recovery students are not wholly aligned with the program's *Standards and Guidelines*. In particular, inconsistencies were observed around schools' decisions to include or exclude students receiving special education services. This will be explored further in the final year of the evaluation.[1]

1 The student selection issue has no bearing on randomization or compliance with random assignment.

Implementation Fidelity: Quality

» Over two years we have learned that Reading Recovery implementers understand fidelity to *Standards and Guidelines* to be necessary, *but not necessarily sufficient,* for high-quality implementation. In recognition of this reality, in Year Two we paired our examination of implementation fidelity with an exploration of other features of implementation that vary from school to school and appear to impact program quality. Our analysis of implementation is ongoing, and will be included in the final report.

I. Introduction

Overview of Study

This report presents the second-year findings of a four-year research effort to evaluate the scale-up, implementation, and impacts of Reading Recovery in the United States. This research is being conducted by the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania, in collaboration with the Center for Research in Education and Social Policy (CRESP) at the University of Delaware. It is funded by a Scaling Up What Works grant from the U.S. Department of Education Investing in Innovation (i3) Fund. The Ohio State University received this i3 grant in 2010 to expand the use of Reading Recovery across the US by training 3,675 new teachers in the program, thereby extending the intervention to an additional 88,200 students. CPRE was contracted to conduct an independent evaluation of the i3 scale-up of Reading Recovery over the course of the five-year grant period.[2]

CPRE's evaluation of Reading Recovery includes parallel rigorous experimental and quasi-experimental designs for estimating program impacts, coupled with a large-scale, mixed-methods study of program implementation under the i3 scale-up. The primary goals of the evaluation are:

- To provide experimental evidence of the short- and long-term impacts of Reading Recovery on student learning in schools that are part of the i3 scale-up;
- To assess the implementation of Reading Recovery under the i3 grant, including fidelity to the program model and progress toward the scale-up goals; and
- To explore the relationship between school-level impacts and program implementation.

The impact evaluation comprises a multi-site randomized controlled trial (RCT) for estimating short-term impacts and a regression discontinuity design (RDD) for estimating long-term impacts. The RCT component involves hundreds of schools over four years, and is described in detail below. Because the first cohort of students served under the i3 grant is only now in third grade, the RDD component of the evaluation will begin in 2013-2014, and long-term program impacts will be presented in the final report. The implementation study involves a combination of qualitative and quantitative research executed on a large scale over the same four-year timeframe. This study is also described below.

The report on Year One of this project was published in August, 2013, and is available for download at: http://readingrecovery.org/images/pdfs/Reading_Recovery/Research_and_Evaluation/RRi3_Year1Eval_ Report.pdf

2 While the grant period is five years, research activity began at the start of the second year of the grant (2011-2012), which we refer to as Year One of the evaluation.

Reading Recovery

The Year One report includes a detailed description of Reading Recovery's history, structure, and program model. Briefly, Reading Recovery is a first-grade literacy intervention designed to help the lowest-achieving readers reach average levels of classroom performance in literacy. Students identified to receive Reading Recovery meet individually with a teacher trained in Reading Recovery each school day for 30-minute lessons over a period of 12 to 20 weeks. Reading Recovery instruction is intended to be supplemental; students receiving the lessons typically continue to participate in regular classroom literacy instruction while the intervention is in progress.

The purpose of Reading Recovery lessons is to support the rapid acceleration of literacy learning. The intervention's underlying principle is that 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 instruction attends to phonemic awareness, phonics, vocabulary, fluency, comprehension, and composition. It ultimately strives 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 information in their reading and writing processes (Clay, 1991, 2005a, 2005b, 2005c). The Reading Recovery model is based on theory that asserts that once equipped with these strategies for independent processing, struggling readers can achieve at average reading levels and maintain proficiency in the regular classroom without special intervention.

The teachers who provide Reading Recovery instruction must complete a rigorous training process that includes a full year of graduate-level coursework and site-based support, followed by ongoing professional development. This training emphasizes the program's theories of literacy learning and works to develop teachers' observation, diagnostic, and instructional skills. The intent is to produce strong Reading Recovery teachers who are expert observers able to make nuanced, in-the-moment decisions about instruction based on students' subtle literacy behaviors.

Both the Reading Recovery coursework and the site-based support are provided by a specially trained instructor–known as a teacher leader–who has deep expertise in Reading Recovery instruction. Typically, teacher leaders are responsible for teaching Reading Recovery classes to new teachers and supporting them in their schools, and for providing ongoing professional development–known as continuing contact–to experienced teachers. Teacher leaders are trained and supported at the regional level by university faculty members associated with one of 20 Reading Recovery University Training Centers (UTCs) nationally. Site coordinators–typically school district administrators charged with overseeing Reading Recovery–support training and implementation activities at the school district level.

The implementation of Reading Recovery at the site and school level is guided by the Standards and Guidelines of Reading Recovery in the United States, 6th Edition (2012). This document outlines the key tasks that are to be performed by implementers. The role of the Standards and Guidelines in governing school-level operations is detailed in the Year One report, and is further discussed later in this report.

Timeline of Study

The i3 grant was awarded to The Ohio State University in October of 2010, after the start of the 2010-2011 school year. As a result of this timing, it was not possible to begin the experimental impact study in the first year of the grant. The 2010-2011 school year was therefore devoted to study design, planning, and instrument development. The second year of the grant period (i.e., 2011-2012) marked the first year of the external evaluation and is the focus of CPRE's Year One report. The RCT and RDD data collection processes and the implementation study were all launched that year.

This is the second in a series of three reports on the i3-funded evaluation of Reading Recovery. Second-round impact findings are reported here. The third and final report, to be released in 2015, will present final estimates of both short- and long-term impacts on student outcomes. It will also present the findings of the RDD, which focuses on long-term program impacts; this analysis cannot be conducted until the first cohort of students who received Reading Recovery under the i3 grant completes third grade, in 2014-2015. The final report will also include a cumulative perspective on the scale-up effort, and recommendations for maximizing the quality of implementation and the magnitude of impacts for Reading Recovery.

Year One: Key Findings

Key findings from Year One guided our work in Year Two. These findings, which are described in-depth in the Year One report, include the following:

Impacts on student reading performance

- Students who received Reading Recovery services outperformed students in the control group on each subscale of the Iowa Test of Basic Skills (ITBS) in Reading, which was administered to both groups at the end of the treatment period.
    - The mean of Reading Recovery students' posttest ITBS Total Reading scores was at the 36th percentile for the national Grade 1 norming population for the ITBS. Students in the control group had posttest scores at the 18th percentile–a difference of +18 percentage points.
    - The mean of Reading Recovery students' posttest ITBS Reading Words scores was at the 43rd percentile nationally, while students in the control group had posttest scores at the 27th percentile–a difference of +16 percentage points.

    - The mean of Reading Recovery students' posttest ITBS Reading Comprehension scores was at the 39th percentile nationally, while students in the control group has posttest scores at the 19th percentile—a difference of +20 percentage points.

- The estimated standardized effect of Reading Recovery on students' ITBS Total Reading Scores was .68 standard deviations relative to the population of struggling readers eligible for Reading Recovery under the i3 scale-up, and 0.47 standard deviations relative to the nationwide population of all first graders. These standardized effect sizes are large relative to typical effect sizes found in evaluations of educational interventions.
- Reading Recovery students' average gain on the ITBS Reading Total score was 4.2 points greater than that of students in the control group over the five-month experiment period. This is equivalent to an additional 1.9 months of progress, or a growth rate that is 38% faster than the national average for beginning first graders.
- Effect estimates were similarly large for both the ITBS Reading Words and Reading Comprehension subscales.
- The impact estimates of Reading Recovery vary substantially across schools, with most schools having moderate to large positive impact estimates (greater than 0.40 standard deviations).

School-level implementation

- School-level implementation of Reading Recovery was, in most respects, faithful to the Reading Recovery Standards and Guidelines. One notable exception concerns the selection of students to receive Reading Recovery. Schools manage this selection process in different ways, and in some cases these diverge from the Standards and Guidelines.
- Reading Recovery teachers have very demanding schedules, but report high levels of satisfaction with their jobs and their training in Reading Recovery.
- Teacher leaders play a critical role in supporting the work of Reading Recovery teachers; however, they are challenged by providing adequate support in the time they have available.
- There is great variation in the extent to which Reading Recovery is supported by and integrated into schools' processes and cultures.

The following section details the key questions that emerged from these findings to guide our work in Year Two.

II. Year Two: Research Questions and Methods

This report documents the findings of research conducted throughout the 2012-2013 academic year, the second full year of the Reading Recovery i3 evaluation. During Year Two, CPRE continued to monitor the progress of the scale-up and assessed the intervention's impacts on student learning.

A goal of our research in Year Two was to begin the work of linking implementation with impacts. Over the course of the year, we worked to build hypotheses about the aspects of Reading Recovery implementation that appear to impact intervention quality and may contribute to variation in program impacts. This effort was guided by our foundational work in Year One, which yielded a deep understanding of the roles, structures, and processes that support the implementation of Reading Recovery at the school level, and gave us a starting point for a more targeted investigation of program implementation. This work was further informed by our efforts, over two years of research, to understand and document implementation fidelity in the context of Reading Recovery's i3 scale-up. The research methods and conceptual framework that guided the development of hypotheses linking implementation and impacts are discussed in detail later in this section.

Three questions guided our research in Year Two:

1. What were Reading Recovery's impacts on student achievement in Year Two?
2. What did Reading Recovery implementation look like in Year Two? Were the scale-up objectives met, and was the program implemented with fidelity?
3. Which aspects of Reading Recovery implementation appear to affect variation in school-level impacts?

Research Question #1: Assessing Program Impacts

As in Year One, Reading Recovery's short-term impacts on students' reading achievement were estimated via a multi-site randomized experiment. Prior to the start of the 2012-2013 school year, 348 schools participating in the i3 scale-up were randomly selected for inclusion in the RCT.[3] At each selected school, a subsample of low-performing students was identified using the Observation Survey of Early Literacy Achievement (OS), developed by Marie Clay (2005a). These students were then rank-ordered according to their text reading levels as established by the OS. The eight eligible students with the lowest scores were matched into pairs according

3 Please refer to the Year One report for a detailed description of the sampling process for both the RCT and the RDD.

to pretest scores and English Language Learner (ELL) status. One student in each pair was randomly assigned to the treatment group, which received Reading Recovery services in the first half of the school year in addition to regular classroom literacy instruction. The other student in each pair was assigned to the control group, which received regular classroom literacy instruction. At the conclusion of the 12- to 20-week intervention period, both students in each matched pair were assessed using the reading sections of the ITBS. The control student from each pair was then eligible to receive Reading Recovery in the second half of the school year.

This "blocking" of students in matched pairs was intended to address the variability of the length of the intervention cycle for students. Blocking students in 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 students increased the likelihood of baseline equivalence of treatment and control groups in regards to pretest scores (OS text reading levels) and ELL status.

The outcome measures: The Observation Survey and Iowa Test of Basic Skills

The OS is the pretest measure for the impact study. The OS is the primary screening, diagnostic and monitoring instrument for Reading Recovery. It is a one-to-one, teacher-administered, standardized assessment that includes six sub-scales: Letter Identification, Concepts about Print, Ohio Word Test, Writing Vocabulary, Hearing and Recording Sounds in Words, and Text Reading Level (Clay, 2005a). The Text Reading Level subtest is used to block students during the random assignment process, and later as a pretest covariate in the statistical models of impacts.

CPRE selected the ITBS as the outcome measure for the impact study. This measure was used in each of the first two years of the RCT. The ITBS is a well-regarded, group-administered, norm- and criterion-referenced, standardized assessment designed to "assess the extent to which a child is cognitively ready to begin work in the academic aspects of the curriculum" (Hoover, Hieronymus, Frisbie, & Dunbar, 1994) and to "measure growth in fundamental areas of school achievement" (Hoover, et al., 2003, 1). Please refer to the Year One report for a more detailed discussion of the components and technical characteristics of both the ITBS and the OS.

Statistical analyses of impacts on student reading achievement

Impacts on student reading performance were estimated by comparing mid-year 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) with students 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 OS text reading level scores as a covariate, random effects for blocks (matched pairs), a random effect for overall school performance (random school intercepts), and a random effect for the impact of Reading Recovery (random treatment effects across schools). 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.

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. These raw impact estimates are standardized using the standard deviation of the outcome for the control group to produce Glass' D. The choice to use Glass' D is based on the expectation that the impact of Reading Recovery would 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 were better able to benchmark the impact estimate against the counterfactual (i.e., the distribution of potential outcomes in the absence of the intervention).

In addition to Glass' D, 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), in this report we present a population-based Cohen's D standardized effect size. This effect size was calculated by dividing the raw impact estimate by the standard deviation of ITBS scores for the national norming sample. This allowed the impact of Reading Recovery to be benchmarked against the full population of first-grade students, not just the struggling readers in the study sample. One would expect these impact estimates to be smaller than the Glass' D estimates because the variance in outcomes for the full population of first graders was larger than the variance for struggling readers.

Research Question #2: Investigating the Implementation of Reading Recovery

CPRE's second research question for Year Two concerns the implementation of Reading Recovery, including progress toward scale-up goals and school-level enactment.

Progress toward scale-up goals

CPRE compared scale-up results for Year Two with projections made at the outset of the grant period in three areas:

1. Recruitment and training of teachers in Reading Recovery under the i3 grant;
2. Retention of teachers from Year One to Year Two; and,
3. Number of students served.

A comprehensive discussion of progress toward all the goals of the scale-up–teacher and teacher leader recruitment, and students served–will be included in the final report.

To track progress on teacher recruitment, we worked in partnership with Reading Recovery's International Data and Evaluation Center (IDEC), which is housed at The Ohio State University. IDEC maintains administrative data on all Reading Recovery schools, students, and personnel, and provided CPRE with count data on teacher recruitment and retention, and students served. CPRE analyzed these data, comparing the totals with prior years' numbers, to assess scale-up progress. Additionally, we gathered qualitative data from interviews with UTC directors and teacher leaders to explore strategies used for recruitment, as well as perceived challenges to the recruitment and retention of schools and Reading Recovery teachers. Further details of the methods and the findings of our analysis of scale-up progress are discussed in Section III.

School-level implementation

As in Year One, our investigation of school-level implementation of Reading Recovery in Year Two involved a comprehensive set of qualitative and quantitative research methods and activities. Table 1 provides an overview of data sources used in Year Two.

This study draws on the perspectives of key players involved with the Reading Recovery implementation at hundreds of schools using interviews, focus groups, document review, daily activity logs, and surveys. Detailed descriptions of the interview protocols and sampling methods for the project are included in the Year One report. These were consistent from Year One to Year Two with two exceptions. In Year One, independent random samples were drawn for the Reading Recovery teacher and principal interviews, and focus groups were conducted to obtain the perspectives of teacher leaders. Based on the findings of the Year One research, we decided to use a school-based cluster sampling process in Year Two. We randomly sampled 30 Reading Recovery teachers, and then worked to recruit the principals and teacher leaders at their schools to participate in interviews. This approach enabled us to develop a more detailed picture of how implementation unfolds at the school level, and is reflective of the evolution of our mixed-methods approach, further described below. This method also allowed us to explore the role of the teacher leader in more depth than was possible with the focus group approach.

Additionally, as in Year One, schools were recruited to be studied as individual cases of Reading Recovery implementation. This case study work continued in Year Two, and will continue in Year Three. The field work for these cases has included interviewing Reading Recovery implementers and school-level personnel, shadowing Reading Recovery teachers or teacher leaders, and observing classroom instruction and one-to-one Reading Recovery lessons. We had also selected a subset of these schools to be studied over multiple years in order to examine changes in Reading Recovery implementation over time. The data gathered through case studies of individual schools are not included in this report. A cross-case analysis is ongoing, and will be presented as a separate report.

Table 1: Implementation Study Data Sources

Qualitative Data Sources	Quantitative Data Sources
» 45 interviews with Reading Recovery teachers » 24 interviews with teacher leaders » 18 interviews with University Training Center directors » 3 interviews with project directors/staff of the i3 office at The Ohio State University » 21 interviews with principals » 9 site-based case studies, which included: » Observation of Reading Recovery lessons » Review of Reading Recovery lesson documentation » Discussion of Reading Recovery lesson data with Reading Recovery teacher » Observation of first-grade classroom literacy instruction » 12 interviews with Reading Recovery teachers » 15 interviews with first-grade classroom teachers » 9 interviews with teacher leaders » 8 interviews with principals » 7 interviews with district site coordinators	» Surveys of Reading Recovery teachers (1506 responses for a response rate of 76%) » Surveys of first-grade teachers (599 responses for a response rate of 61%) » Surveys of teacher leaders (210 responses for a response rate of 80%) » Surveys of site coordinators (133 for a response rate of 72%) » More than 5,000 daily activity logs completed by Reading Recovery teachers on 10 randomly selected school days throughout the year

In Year Two, interviews and observations were designed to explore how and why Reading Recovery implementation differs from one school or site to another. All interview protocols explicitly probed for a deeper understanding of quality of implementation; participants were asked to discuss what they thought constituted a high-quality Reading Recovery implementation, and their understanding of the relationship between fidelity to the Standards and Guidelines and quality. Coding and analysis of qualitative data involved multiple coders, integrated reliability checks, and multiple stages of analysis. For more details on the process used to analyze school-level qualitative data, please refer to Section IV.

Research Question #3: Linking Implementation and Impacts

An overarching goal of CPRE's evaluation is to explore the relationship between program implementation and school-level impacts. In Year One, we focused heavily on fidelity of implementation, generating hypotheses about how adherence to/departures from fidelity may be related to impacts. The work of testing these hypotheses is ongoing, and this analysis will be included in the final report.

In Year Two, we began to focus on quality of implementation in addition to fidelity, based on qualitative data gathered over the first two years of the evaluation. Our Year Two research into Reading Recovery's implementation grew out of two key understandings that emerged from our work in Year One:

1. There is significant variation in Reading Recovery's school-level impacts that does not seem to be explained by the fidelity analysis; and,

2. Because Reading Recovery's implementation requires the involvement of multiple players performing a range of nuanced tasks–some quite removed from the program priorities of lesson delivery and teacher training–there are important aspects of implementation that are not easily codified as Standards and are not captured by our measures of fidelity.

Understanding quality in the context of Reading Recovery's implementation therefore became a goal of our research in Year Two–one we believe is a critical step in building hypotheses about sources of variation in school-level impacts. We continue to develop and test our hypotheses regarding factors that influence quality of implementation and their relationship to school-level impacts; our findings will be presented in the final report.

III. Impacts on Student Achievement

This section presents the statistical analysis of ITBS reading test scores in order to estimate the impacts of Reading Recovery on students participating in the 2012-2013 randomized experiment. As described in Section II, hierarchical linear modeling was used to analyze differences between the ITBS reading scores of students in the treatment and control groups. The primary analyses focused on the ITBS Total Reading Scale scores. Additional exploratory analyses examined impacts on the Reading Words and Reading Comprehension subscale scores, and impacts on Total Reading scores for ELL students and those in rural schools. Detailed information about the pretest and outcome measure are provided in Appendix D.

School Participation and Data Availability

Of the 348 i3 schools randomly selected in July 2012 to participate in the RCT during the 2012-2013 school year, 267 schools actually carried out the random assignment process (See Appendix E). Several modes of direct and indirect communication were used to inform schools of their expected participation in the random assignment study, including direct emails from IDEC to individual teacher leaders and Reading Recovery 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.

Results from follow-up data collection indicate there are several reasons why schools did not carry out the RCT. Of the 81 schools that were selected but did not participate in the RCT, 39 dropped out of the i3 project before implementing Reading Recovery. This puts the overall school compliance rate at 86% (267 out of 309 schools). Reasons why the remaining 42 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, misinterpretation of instructions). The 39 schools that dropped out of the i3 project and the 42 schools that did not carry out random assignment are not included in the impact analyses presented here.

Across the 267 schools that implemented random assignment, a total of 2,092 students were randomly assigned to treatment (N=1,048) and control (N=1,044) conditions.[4] Of these students, a total of 1,893 (980 treatment, 913 control) had available pretest data (fall OS scores). Of those with pretest data, 1,697 (872 treatment, 825 control) had available posttest data (ITBS scores). After linking treatment students to their matched controls, a total of 1,430 students were able to be matched into pairs of treatment and control (715 matched pairs in 233 schools). This sample represents 68% of the students in schools that carried out random

4 There are a few extra students assigned to the treatment condition in schools with less than eight eligible students. Any eligible student without a matched pair was automatically assigned to the treatment condition, although their data is not included in any impact analyses due to the absence of a matched control.

assignment. The missing data at the student level primarily resulted from student mobility or other factors that prohibited administration of the ITBS to both treatment and control students in a pair.

The multi-site, matched-pairs design of this random assignment study means that each school and each pair is an independent mini-experiment, and that the ability to calculate valid causal impacts is less prone to problems associated with school non-participation or missing data. Although the sample size is reduced and may be less representative of the target population, the impact estimates for the reduced sample are still valid indicators of causal impacts for those schools and students in matched pairs that actually participated in the experiment.

Because the purpose of this study is to assess the impact of Reading Recovery at scale, generalizability of findings to the overall population of i3 schools is a critical goal. 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 levels ($p = .63$), gender ($p = .55$), race ($p = .94$), or ELL status ($p = .68$). As additional students and schools participate in the RCT in the coming year, we will collect additional data to further explore the reasons why schools and students are missing data, and the degree to which this may or may not limit generalizability of results.[5]

Table 2: Baseline Balance Test for Student Demographics

Pretreatment Variable	Treatment Group	Control Group	p-value for Difference
Gender[a]			
Male	58%	60%	0.42
Female	42%	40%	
ELL Status[b]			
ELL	21%	21%	0.81
Non-ELL	79 %	79%	
Race[c]			
Black	16%	17%	0.93
Hispanic	21%	21%	
White	55%	55%	
Other	8%	7%	

Notes: The analytic sample consists of 1,430 students in 233 schools. [a]N = 1,428; [b]N=1,427; [c]N=1,428.

5 Free and reduced lunch status (FRL) data are not included in these analyses given that many school districts' privacy policies do not allow reporting of FRL data to IDEC. As such, the rates of missing data are too high to support meaningful analyses.

Baseline balance tests

Baseline balance tests were performed in order to examine whether the treatment and control groups were equivalent on observed characteristics after random assignment. Table 2 presents results for baseline balance tests for student demographics of the final analytic sample of 1,430 students in 233 schools.

No significant differences were found between treatment and control groups on gender, ELL status, or race. 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 Reading Recovery for the group of treatment students.

Table 3 shows a baseline balance test for prior reading performance of students in the analytic sample. Again, no significant differences were found between the treatment and control groups, and the percentages in the two columns match up well. This confirms that the treatment and control groups had initial reading performance that was nearly identical immediately prior to implementation of Reading Recovery for the treatment group students.

Table 3: Baseline Balance Test for Student Pretest Reading Performance

Pretreatment Variable	Treatment Group	Control Group	p-value for Difference
Text Reading Level			
0	49%	50%	0.95
1	21%	20%	
2	17%	18%	
3	12%	12%	

Note: The analytic sample consists of 1,430 students in 233 schools.

The control group experience

In order to better understand what is reflected in the impact estimates from this study, it is crucial to know how the experience of students in the control group compares to that of students in the treatment group. This is especially important in this study because control group students were allowed to receive any available supports or intervention services other than Reading Recovery during the time that the treatment group received Reading Recovery instruction; therefore, in this study, we are not comparing Reading Recovery to a true control condition. Instead, we are comparing the effectiveness of Reading Recovery to that of other support services that schools might provide to struggling readers.

To document the experience of control group students, we surveyed their first-grade teachers and asked them to record, individually for each control group student in their classroom, what

supplemental instructional services the students' received during the first half of the school year (i.e., when the treatment students were participating in Reading Recovery). We were able to collect these data on 464 (65%) of the 715 control group students. Among these students, 24% received no supplemental supports, 30% participated in an alternate literacy intervention, 31% participated in small group work with a Reading Recovery teacher, and 17% received other supports under ELL or special education programs. This confirms that the vast majority of control group students experienced substantial support in addition to regular classroom instruction.

Overall impacts on ITBS reading scores

Table 4 shows simple descriptive statistics for the treatment and control groups on raw scores and scale scores from the reading sections of the ITBS. 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, +16 for Reading Comprehension, and +14 overall. Table 4 below provides ITBS raw and scale scores.

Statistical tests of significance of differences in ITBS scale scores between the treatment and control groups were performed using HLM models as described in Section II and Appendix A. Results from these analyses are presented in Tables 5, 6, and 7.

Analyses of impacts on ITBS Total Reading scores showed a significant positive effect of Reading Recovery overall. As shown in Table 5, the point estimate for the difference between treatment and control students' expected Total Reading Scores on the ITBS was 3.03 with a p-value significant at greater than 99% confidence. The school intercept/impact correlation was negative ($\rho = -.39$) and statistically significant at the 99% confidence level. This suggests that Reading Recovery impact estimates tended to be larger in schools where participating students have lower average ITBS scores.

Table 4: Descriptive Statistics for ITBS Scores for Treatment and Control Groups

Mid-Year Outcomes	Treatment Group	Control Group	Difference
ITBS Total Reading Scores			
Scale Scores			
Mean	138.6	135.5	3.1***
(Standard Deviation)	(7.1)	(7.3)	-
Mean Percentile Rank[a]	36	22	14
ITBS Reading Words Subscale			
Raw Scores			
Mean	20.8	18.5	2.3
(Standard Deviation)	(5.2)	(5.7)	-
Scale Scores			
Mean	140.6	137.2	3.4***
(Standard Deviation)	(8.8)	(8.3)	-
Mean Percentile Rank[a]	43	27	16
ITBS Comprehension Subscale			
Raw Scores			
Mean	9.8	8.3	1.5
(Standard Deviation)	(4.0)	(3.8)	-
Scale Scores			
Mean	139.7	136.4	3.3***
(Standard Deviation)	(9.3)	(9.0)	-
Mean Percentile Rank[a]	39	23	16
Sample N	715	715	

***p<.0001

Notes: The ITBS Test data were obtained through tests completed by students in pull-out sessions at study schools during January of 2014. All test administrators were trained in the test developer's administration procedures and adhered to the stipulated administration time frames and procedures, including allocating extra time to students who had Individual Education Plans (IEPs). [a] Percentile ranks based on ITBS Grade 1 midyear norms (Hoover et al., 2006).

Table 5: HLM Analysis of Overall Treatment Effects of Reading Recovery on ITBS Total Reading Scores

Dependent Variable: Mid-Year ITBS Total Reading Scores	Estimate	Standard Error	p-value
Fixed Effects			
Intercept (β_0)	135.52	0.33	<.0001
Pretest[a] (β_1)	1.50	0.16	<.0001
Treatment Effect (β_2)	3.03	0.37	<.0001
Random Effects			
Matched Pair Variance (ω^2)	4.00	1.35	.0015
School Intercept Variance (τ^2)	15.83	2.47	<.0001
School Treatment Impact Variance (ξ^2)	14.77	3.01	<.0001
School Intercept/Impact Correlation (ρ)	-0.38	0.10	.0002
Student-Level Residual Variance (σ^2)	25.43	1.60	<.0001

Notes. The analytic sample consists of 1,430 students in 233 schools. See Appendix A for the mathematical / symbolic form of the HLM model of program impacts and definitions for each model parameter. [a] The raw student-level correlation between pretest and posttest scores was .32.

Dividing the point estimate by the standard deviation of the control group yields a Glass' D effect size of 0.42 standard deviations. This effect estimate reflects the impact of Reading Recovery relative to the population of struggling readers eligible for Reading Recovery in participating schools. Alternatively, dividing the point estimate by the standard deviation from the ITBS 2005 national norming sample of first graders (i.e., s = 9.1) yields a Cohen's D effect size of 0.33 standard deviations. This effect estimate reflects the impact of Reading Recovery relative to the full population of all first graders across the nation.

Analyses of impacts on the ITBS Reading Words subscale showed similar results. The model estimates are shown in Table 6. The point estimate for the difference between treatment and control students' expected Reading Words scores on the ITBS was 3.36 points ($p<.0001$). Dividing this point estimate by the standard deviation of the control group yields a Glass' D effect size of 0.40 standard deviations. Alternatively, dividing the point estimate by the standard deviation from the ITBS 2005 national norming sample of first graders (i.e., s = 10.2) yields a Cohen's D effect size of 0.33 standard deviations.

Table 6: HLM Analysis of Overall Treatment Effects of Reading Recovery on ITBS Reading Words Subscale Scores

Dependent Variable: Mid-Year ITBS Reading Words Scores	Estimate	Standard Error	p-value
Fixed Effects			
Intercept (β_0)	137.16	0.38	<.0001
Pretest[(a)] (β_1)	1.73	0.20	<.0001
Treatment Effect (β_2)	3.36	0.42	<.0001
Random Effects			
Matched Pair Variance (ω^2)	5.59	2.10	0.0039
School Intercept Variance (τ^2)	18.49	3.15	<.0001
School Treatment Impact Variance (ξ^2)	13.06	3.86	0.0004
School Intercept/Impact Correlation (ρ)	-0.29	0.13	0.0292
Student-Level Residual Variance (σ^2)	40.67	2.56	<.0001

Notes. The analytic sample consists of 1,430 students in 233 schools. See Appendix A for the mathematical / symbolic form of the HLM model of program impacts and definitions for each model parameter. [a] The raw student-level correlation between pretest and posttest scores was .30.

As seen in Table 7, analyses of impacts on the ITBS Reading Comprehension subscale showed similar results. The point estimate for the difference between treatment and control students' expected Reading Comprehension scores on the ITBS was 3.26 with a p-value significant at greater than 99% confidence. Dividing that point estimate by the standard deviation of the control group yields a Glass' D effect size of 0.36 standard deviations. Alternatively, dividing the point estimate by the standard deviation from the ITBS 2005 national norming sample of first graders (i.e., s = 10.2) yields a Cohen's D effect size of 0.32 standard deviations.

The significant variance components for random effects in the HLM models of impacts on ITBS scores suggest that the magnitude of the Reading Recovery impact estimates varies substantially across schools. The results for the overall impact model in Table 5 show an average effect of +3.03 points, with a random effect covariance estimate for the school-level impacts of 14.77 points. Taking the square root of this covariance estimate yields a standard deviation of 3.84 points.

Table 7: HLM Analysis of Overall Treatment Effects of Reading Recovery on ITBS Reading Comprehension Subscale Scores

Dependent Variable: Mid-Year ITBS Reading Comprehension Scores	Estimate	Standard Error	p-value
Fixed Effects			
Intercept (β_0)	136.32	0.42	<.0001
Pretest[a] (β_1)	1.60	0.21	<.0001
Treatment Effect (β_2)	3.26	0.48	<.0001
Random Effects			
Matched Pair Variance (ω^2)	6.16	2.29	0.0035
School Intercept Variance (τ^2)	24.09	3.98	<.0001
School Treatment Impact Variance (ξ^2)	22.84	5.02	<.0001
School Intercept/Impact Correlation (ρ)	-0.28	0.12	0.0171
Student-Level Residual Variance (σ^2)	43.79	2.78	<.0001

Notes. The analytic sample consists of 1,430 students in 233 schools. See Appendix A for the mathematical / symbolic form of the HLM model of program impacts and definitions for each model parameter. [a] The raw student-level correlation between pretest and posttest scores was .27.

Impacts on ITBS Reading Scores in rural schools and for English language learners

The i3 scale-up of Reading Recovery includes a specific focus on rural schools and students who are English language learners. Exploratory analyses of treatment effects on ITBS Total Reading Scores for these two subgroups are presented below.

Table 8 shows simple descriptive statistics for ITBS Total Reading Scores of students in rural schools by treatment and control groups. Once again, the means are over one-half of a standard deviation larger in the treatment group.

Statistical tests of significance of differences in ITBS Total Reading scores between the treatment and control groups in rural schools were performed using HLM models as described in Section II and Appendix A. Table 9 shows the results of these tests.

The results for rural schools were very similar to the overall results. Analyses of impacts on ITBS Total Reading scores showed a highly significant positive effect of Reading Recovery in rural schools. The point estimate for the difference between rural treatment and control students' expected Total Reading scores on the ITBS was 3.51 with a p-value significant at greater than 99% confidence. Dividing that point estimate by the standard deviation of the control group yields a Glass' D effect size of 0.49 standard deviations. Additional analyses using the full

sample with a cross-level interaction between the ELL indicator and Treatment assignment indicator showed no significant difference in impacts for rural and non-rural schools. Estimates of the school-level variability in treatment effects were statistically significant at the 99% confidence level, and the variance was even larger in rural schools than the effect variance in the overall analysis. This suggests that while the majority of rural schools' Reading Recovery programs have positive impacts on student performance, they vary greatly in their ability to produce sizable impact estimates. The school intercept/impact correlation was negative and statistically significant at the 99% confidence level. This suggests that Reading Recovery impact estimates tend to be larger in rural schools where participating students have lower average ITBS scores.

Table 8: Descriptive Statistics for ITBS Total Reading Scores for Treatment and Control Groups in Rural Schools

Mid-Year Outcomes	Treatment Group	Control Group
ITBS Total Reading Scores		
Mean	139.6	135.9
(Standard Deviation)	(6.5)	(7.1)

Note. The analytic sample consists of 654 students in 108 schools.

Table 9: HLM Analysis of Treatment Effects of Reading Recovery on ITBS Total Reading Scores in Rural Schools

Dependent Variable: Mid-Year ITBS Total Reading Scores	Estimate	Standard Error	p-value
Fixed Effects			
Intercept (β_0)	135.94	0.49	<.0001
Pretest[(a)] (β_1)	1.23	0.23	<.0001
Treatment Effect (β_2)	3.51	0.57	<.0001
Random Effects			
Matched Pair Variance (ω^2)	5.74	1.98	0.0018
School Intercept Variance (τ^2)	15.77	3.56	<.0001
School Treatment Impact Variance (ξ^2)	18.12	4.68	<.0001
School Intercept/Impact Correlation (ρ)	-0.56	0.11	<.0001
Student-Level Residual Variance (σ^2)	23.06	2.16	<.0001

Notes. The analytic sample consists of 654 students in 108 schools. See Appendix A for the mathematical/symbolic form of the HLM model of program impacts and definitions for each model parameter. [a]The raw student-level correlation between pretest and posttest scores was .23.

Table 10 shows simple descriptive statistics for ITBS Total Reading Scores of ELL students by treatment and control groups. Once again, the means are approximately one-half of a standard deviation larger in the treatment group.

Table 10: Descriptive Statistics for ITBS Total Reading Scores for ELL Students by Treatment and Control Groups

Mid-Year Outcomes	Treatment Group (N=111)	Control Group (N=111)
ITBS Total Reading Scores		
Mean	136.4	132.8
(Standard Deviation)	(7.6)	(7.4)

Note. The analytic sample consists of 222 students in 66 schools.

Table 11 : HLM Analysis of Treatment Effects of Reading Recovery on ITBS Composite Reading Scores for ELL Students

Dependent Variable: Mid-Year ITBS Total Reading Scores	Estimate	Standard Error	p-value
Fixed Effects			
Intercept (β_0)	133.84	0.77	<.0001
Pretest[a] (β_1)	2.35	0.52	<.0001
Treatment Effect (β_2)	3.50	0.87	<.0001
Random Effects			
Matched Pair Variance (ω^2)	4.08	4.11	0.161
School Intercept Variance (τ^2)	17.35	6.10	0.002
School Treatment Impact Variance (ξ^2)	17.04	10.40	0.051
School Intercept/Impact Correlation (ρ)	-0.26	0.28	0.347
Student-Level Residual Variance (σ^2)	24.74	5.04	<.0001

Notes. The analytic sample consists of 222 students in 66 schools. See Appendix A for the mathematical/symbolic form of the HLM model of program impacts and definitions for each model parameter. [a]The raw student-level correlation between pretest and posttest scores was .31.

Statistical tests of significance of differences in ITBS Total Reading scores of ELL students in the treatment and control groups were performed using HLM models as described in Section II and Appendix A. Results from that analysis are presented in Table 11.

The results for ELL students were also very similar to the overall results. Analyses of impacts on ITBS Total Reading scores of ELL students showed a highly significant positive effect of Reading Recovery. The point estimate for the difference between ELL treatment and control students'

expected Total Reading Scores on the ITBS was 3.50 with a p-value significant at greater than 99% confidence. Dividing that point estimate by the standard deviation of the control group yields a Glass' D effect size of 0.47 standard deviations. Additional analyses using the full sample with a cross-level interaction between the ELL indicator and Treatment assignment indicator showed no significant difference in impacts for ELL and non-ELL students.

Despite the reduced sample size, the estimates for the estimate of school-level variability in treatment effects was still statistically significant at the 95% confidence level. This suggests that while the vast majority of schools' Reading Recovery programs have positive impacts on ELL students' reading performance, they vary greatly in their ability to produce sizable impact estimates for ELL students. There was not a statistically significant correlation between school average ITBS scores and the impact of Reading Recovery for ELL students.

Control group contamination

Several important considerations should be applied to the interpretation of findings from any experimental study. First is the extent to which impact estimates may be affected by deviations from the experimental design. Significant deviations–such as noncompliance with random assignment–can threaten a study's internal validity. Even less serious deviations can affect the magnitude of impact estimates. CPRE analyzed IDEC intervention records–which include start dates, exit dates, and the total number of lessons provided to each student–to assess both fidelity to the design, and control contamination.

Using the intervention records, we examined whether each student assigned to treatment received Reading Recovery lessons prior to the control student in his or her matched pair, as planned. Our analysis revealed a small number of matched pairs–13 pairs out of 715 in the Year Two RCT–in which the control student was exposed to Reading Recovery before the treatment-control contrast was estimated. In these cases, the period of overlap (when both treatment and control students in a pair were receiving the intervention, prior to the ITBS administration) ranged from 10 days to 64 days. In two of the 13 cases, the control student's start date preceded that of the treatment student. This might indicate non-compliance with random assignment; however, in both of these instances there were apparent data errors (i.e., the intervention end date for a student preceded the same student's start date), suggesting these may be instances of faulty data entry rather than non-compliance.

In the case of an intervention that is demonstrated to produce positive impacts, like this study of Reading Recovery, control contamination can degrade the magnitude of impact estimates. It is therefore possible that the treatment impacts we observed in Year Two were smaller than they would have been absent any contamination. However, given the very limited scope of exposure to the treatment we observed among control students in the analytic sample, it is unlikely that our impact estimates were significantly affected by contamination of the Reading Recovery intervention in the control condition.

Control group contact with Reading Recovery teachers

To further inform interpretation of the estimated program impacts, we analyzed data on the non-Reading Recovery interactions that control group students had with Reading Recovery teachers during the treatment period. This contact does not represent contamination since it does not involve one-to-one Reading Recovery Lessons. Rather, it provides information about the experiences of control group students prior to the treatment contrast. CPRE collected survey data from Reading Recovery teachers about any instructional interaction they had with control group students in the first half of the school year. Students in the treatment and control group both receive regular classroom instruction, and Reading Recovery teachers spend half of their time engaging in instructional activities outside of their Reading Recovery lessons (e.g., small-group work, co-teaching, push-in support). However, to clearly understand what the treatment contrast represents, it is important to determine if the control students had any instructional interaction with a Reading Recovery teacher before the treatment student within their matched pair exits the intervention.

All Reading Recovery teachers in the 267 schools that implemented random assignment were asked about instructional interactions with each control group student. Survey data was collected from each Reading Recovery teacher in an RCT school on the precise types of interaction (whole class, small group, individualized) and frequency of interaction (from never to daily) they had with each of their four control group students. Averaging across all teacher survey responses, we collected data on 629 (66%) of the control students in 229 schools (98%) in the analytic sample. To investigate the representativeness of the subsample we explored the relationship with student ITBS scores and found that data is more likely to be present for students with slightly lower reading scores (-1.54, $p<.05$).

Within the sample of control students for whom we have Reading Recovery Teacher contact data, we found that 37% of the control group students had some exposure to a Reading Recovery teacher in a whole-class setting. This is likely due to the Reading Recovery teacher supporting whole-class instruction, for example as a co-teacher. This is not surprising given the multiple roles that Reading Recovery teachers play in schools. We also found that 9% of control group students had exposure to a Reading Recovery teacher in a small-group setting at least once during the first half of the school year. We view these interactions with a Reading Recovery teacher in whole-class or small-group settings as acceptable elements in the counterfactual, not forms of contamination. The only restriction in the RCT was that students in the control condition could not participate in one-to-one Reading Recovery lessons; this did not preclude them from receiving any other instructional intervention or supports available (see above section on the control group experience). We also found that 12% of the control group students received individualized support from a Reading Recovery teacher at some point during the first half of the school year. This individualized instructional interaction with

a Reading Recovery teacher has the highest potential for imitating the intervention. However, since the rate of this occurrence is so low, it is reasonable to conclude that the potential for contamination is also low.

Exploratory analyses of the effect of control contact on treatment effect found neither a significant main effect nor a significant interaction with treatment. Although this analysis was limited to pairs in which control students had slightly lower ITBS scores, we found that direct contact between Reading Recovery teachers and control students did not explain variation in program effects.

Benchmarking the Effects of Reading Recovery

CPRE's Year Two impact analysis revealed standardized effect sizes between 0.36 and 0.42 standard deviations. These are large relative to typical effect sizes found in educational evaluations. In their recent paper on the interpretation of effect sizes, Lipsey et al. (2012) offer a number of useful benchmarks for understanding the magnitude of these effects. For randomized studies that use "broad scope" standardized tests as the outcome measure for interventions at the elementary level, the authors report average effects of 0.08 standard deviations (Lipsey et al.,2012, 34). This benchmark suggests that the Year Two standardized effect sizes for Reading Recovery, between 0.36 and 0.42, were at least 4.5 times greater than average for studies that use comparable outcome measures. Based on their analysis of 181 different samples, Lipsey et al. (2012) also present mean effect sizes for different types of educational interventions. They report a mean standardized effect size of 0.13 for "curricula or broad instructional programs." The authors specifically include Reading Recovery in this group. This indicates that Reading Recovery's Year Two effects were 2.8 times greater than the reading outcomes of similar programs. Similarly, the impacts of Reading Recovery in Year Two were 3.5 times larger than the average effects of Title I programs reviewed by Borman and D'Agostino (1996).

It is also helpful to benchmark the treatment effects against expected gains on the ITBS for the national sample of students used to norm the ITBS tests. This permits the interpretation of impacts as an increase in growth rate during the study period. Table 12 shows the expected gains on the ITBS benchmarked against the national sample, the gains in terms of additional months of learning, and the growth rate for Reading Recovery students compared to the national average for beginning first graders. From the start of first grade through the fifth month of the school year (the period during which the treatment students received Reading Recovery instruction), ITBS Reading Total scale scores for the average student in the U.S. are expected to increase from 133 to 144 (Hoover et al., 2003). This increase of 11 points over a five-month period suggests that the additional gains of 3.03 points experienced by Reading Recovery students in Year Two of our evaluation is roughly equivalent to an additional 1.4 months of learning, and translates to a growth rate that is 27% greater than the national average growth rate for beginning first graders. Table 12 also includes data for the Reading Comprehension and Reading Words Subscales.

Table 12: Reading Recovery treatment effects as compared with national benchmarks for first graders

	Treatment Effect (growth in ITBS scores)	p-value	Effect Size	Treatment students' additional months of learning, over national growth average for first graders	Treatment students' growth rate, as a percentage of national average for first graders
Reading Total	3.03	<.0001	0.33 SD	1.38	127%
Reading Words Subscale	3.36	<.0001	0.33 SD	1.52	131%
Reading Comprehension Subscale	3.26	<.0001	0.32 SD	1.48	130%

Notes: From the start of first grade through the fifth month (i.e., the period during which the treatment students received Reading Recovery instruction), ITBS Reading Total scale scores are expected to increase from 133 to 144 for the average student in the U.S. (Hoover et al., 2003).The treatment effect represents additional gains experienced by students who received Reading Recovery .

To date, only a handful of studies investigating the impacts of Reading Recovery have been found to meet evidence standards of the Institute of Education Sciences' What Works Clearinghouse. One such study, conducted by Pinnell et al. (1994), found effect sizes ranging from .49 to 1.50 standard deviations for Reading Recovery. Their findings were based on a number of different evaluation measures, including the Gates-MacGinitie Reading Test and the Woodcock Reading Mastery Tests. Schwartz (2005) also reported large, positive effect sizes (.90 to 2.02) for Reading Recovery using Clay's Observation Survey of Early Literacy Achievement. The findings of the current evaluation to date are consistent with the positive impacts observed in these prior studies.

IV. Implementation: Scale-Up

Along with understanding Reading Recovery's impacts on student achievement, a goal of the implementation study is to examine the progress of the nationwide scale-up. CPRE's research in this area is focused on the three aspects of the scale-up effort that best reflect the i3 grant's overall goal of dramatically increasing the number of U.S. students impacted by the program. These aspects are: 1) the number of new teachers trained in Reading Recovery; 2) the number of students served with one-to-one Reading Recovery lessons; and, 3) the number of students served by Reading Recovery-trained teachers through other instructional activities, including small-group instruction and regular classroom instruction.

CPRE obtained teacher recruitment, attrition, and student-service data from IDEC to evaluate progress in each of these areas. In addition, the scale-up process was explored in interviews with UTC directors and teacher leaders. The focus of this qualitative research was on identifying facilitators and barriers to the expansion and sustainability of the Reading Recovery system, and the strategies used to address barriers.

Progress toward scale-up goals

For purposes of CPRE's evaluation, the pertinent goals of the five-year scale-up are to train 3,675 new Reading Recovery teachers; to provide one-to-one Reading Recovery lessons to an additional 67,264 students; and, to provide other services–specifically, classroom and small-group instruction–to 302,688 students. These goals are consistent with the terms of the i3 grant to The Ohio State University.[6] Table 13 provides an overview of the progress of the scale-up to date, as compared with these goals.

At the end of its third year, the scale-up had reached 105% of its target for recruitment and training of teachers (2,079 teachers trained). The teachers trained during this period provided one-to-one Reading Recovery to 23,720 students, representing 94% of the scale-up goal for lesson delivery. In addition, these teachers served 124,480 students with small-group or classroom instruction, representing 109% of the goal for students served outside of Reading Recovery lessons.

Recruitment and retention: Strategies and challenges

The Year One report includes a detailed description of the Reading Recovery community's efforts to meet the scale-up goals over the first two years of the grant period. In that report, we observed that schools recruited under the i3 grant varied on characteristics such as Title

6 The targets cited reflect adjustments to the original scale-up targets. These adjustments were negotiated between The Ohio State University and the U.S. Department of Education's Office of Innovation and Improvement to account for teacher attrition and the delayed release of i3 funds in the first year of the grant.

I status, ELL population, and free and reduced lunch population. We also observed that recruitment strategies are locally controlled, devised by the UTC director and teacher leaders in each region. In Year Two of the evaluation (the third year of the scale-up), retention of recruited teachers and schools received increased attention from i3 project leadership at Ohio State University (OSU), UTC directors, and teacher leaders. Updates on the challenges to recruitment and retention, and Reading Recovery personnel's evolving efforts to address them, are provided here.

Table 13: Scale-up progress through Year 3 of the i3 grant

Goal	5-Year Scale-up Goal	Goal for Years 1 - 3	Total for Years 1 - 3	Percent of Year 3 goal met
Reading Recovery Teachers trained	3,675	1,980	2,079	105
Students served with one-to-one lessons[a]	67,264	25,328	23,720	94
Other students instructed by Reading Recovery teachers	302,688	113,976	124,480	109

[a]Targets for student services—both one-to-one lessons and other instruction—assume attrition from each cohort at 15% reduction in the second year, 15% more reduction in the third year, 10% more in the fourth year, and 5% more in the fifth year.

Recruitment and the fiscal climate

In the first two years of the scale-up, the primary challenges to recruitment were a difficult fiscal climate, limited access to decision-makers, and negative perceptions of Reading Recovery. Several of these issues have persisted into the third year of the scale-up: UTC directors and teacher leaders continued to name the economic climate, administrator turnover, and limited understanding of Reading Recovery as difficult obstacles to overcome. They reported that the most significant challenge to recruiting schools and teachers continued to be funding.

UTC directors and teacher leaders reported that they hear a great deal of interest in Reading Recovery from school and district administrators; however, the perceived cost of implementation remained a significant barrier to adoption of the program. One UTC director explained that "often the barrier to people signing on the dotted line is 'How do we fund teacher salary?'" Another remarked that "nearly every decision that's being made is being made based on finances."

In some schools, the reported challenge was finding the funds to create a new teacher position. Reading Recovery is designed to be a half-day assignment; therefore it is not strictly necessary to create a new full-time position in order to adopt or expand the program. While some schools

elected to have a classroom teacher deliver Reading Recovery lessons for part of her day, our research indicates that most schools created a new position that combined Reading Recovery with small-group intervention work. In these schools, implementing Reading Recovery necessitated investment in a full-time staff member, which presented a significant financial obstacle.

The fiscal challenges of incorporating Reading Recovery into a school were frequently repeated by administrators, teacher leaders, and UTC directors. While they often reported positive impacts from Reading Recovery, many principals we spoke with also expressed concerns about the small number of students served through the intervention. One principal commented: "I feel like because of the budget and the reality of that, people have found it hard to commit to Reading Recovery when they think about it in terms of the number of individuals that are being served." A UTC director explained that "the principal has to decide that he or she has the budget to totally fund Reading Recovery. And that's where it gets back to a staffing issue. They only have so many spots for personnel." Though the availability of i3 grant funds, which fully cover training costs, has catalyzed many successful recruitment efforts, it is not always enough of an enticement to overcome administrators' financial concerns.

In recognition of this challenge, UTC directors and teacher leaders tailored their recruitment strategies to address the financial issue head-on. A common approach used in recruitment was to help individual schools and districts problem-solve the integration of an additional teacher into their budget. One UTC director explained that her most successful recruitment strategy was to work on a case-by-case basis with district and school administrators to develop "creative staffing structures," emphasizing other roles Reading Recovery teachers can play in a building (e.g. small-group interventionist) and the number of students the teachers serve through those other contacts. In Year Two of the study, a number of UTC directors and teacher leaders reported that this strategy helped to counter claims that Reading Recovery teachers serve too few students to justify their cost.

Recruitment and competing district priorities

The UTC directors and teacher leaders charged with recruiting schools to the scale-up also reported a different kind of obstacle: other policies competing for administrator and teacher attention. More frequently than in Year One, in Year Two respondents cited the Common Core State Standards as an example of a policy that was overwhelming schools' capacity to take on new initiatives. One UTC director observed that "people are feeling… like they can't take something else on in the school when they're trying to take hold of Common Core State Standards as well."

In an atmosphere of increasing concern about Common Core-aligned testing, several UTC directors specifically mentioned crafting recruitment strategies that highlighted how Reading Recovery can help schools meet the standards. One director described that her latest

recruitment strategy was "by hell or high water, to help these people understand how [Reading Recovery] can advantage their teachers and their children, and fulfill their [Response to Intervention], Common Core, political, educational… all their needs."

As described in the Year One report, UTC directors and teacher leaders often explained that a persistent challenge to recruiting schools and districts was a perceived philosophical disagreement with Reading Recovery's approach to literacy, or misunderstandings about its goals and methods. These respondents reported that they heard principals and district administrators describing Reading Recovery as a "whole language" approach to literacy instruction–a characterization Reading Recovery disavows. As one teacher leader explained:

> Sometimes our mindsets cause people to decide not to go with [Reading Recovery], because they don't view it as the right way to teach reading. That's the pendulum constantly swinging in education. And I think that tends to hurt the [recruitment] process.

UTC directors reported countering negative perceptions with student progress data. As one teacher leader explained, "if they saw what we do, that would be very powerful."

In fact, UTC directors and teacher leaders frequently mentioned student data as the key tool for addressing reservations about Reading Recovery–from cost-efficiency concerns to philosophical disagreements. Reading Recovery advocates frequently expressed the belief that understanding of Reading Recovery, both its practices and its impacts, translates to support for the program. "It's the data," explained one teacher leader when asked what argument has been most successful in her recruitment efforts. "The numbers talk." The use of data as a tool in Reading Recovery, and the relationship between understanding of the program and support for it, are further discussed later in this report.

Combating attrition of teachers and schools

In Year Two of this evaluation, we increased focus the sustainability of Reading Recovery implementations. Two major concerns–which mirror the challenges to recruitment noted above–emerged as key drivers of attrition. First, UTC directors and teacher leaders reported that negative perceptions of the cost-effectiveness of Reading Recovery in challenging fiscal times have contributed to attrition of Reading Recovery teachers, schools, and sometimes entire sites. Second, turnover in school or district administration sometimes led to the departure of key supporters of Reading Recovery implementations, or the arrival of administrators who did not support the program for various reasons.

Attrition and cost-effectiveness

As was the case with recruitment, funding was by far the most significant perceived threat to continuation of existing implementations reported by UTC directors, teacher leaders, and school administrators. When asked about the future of Reading Recovery (broadly, or in

specific schools or districts), many teachers and school administrators reported uncertainty, specifically because they do not know what funding changes lie ahead. Because school and district budgets are frequently in flux, many principals reported feeling that Reading Recovery is in a tenuous position. "We'll hold on to it as long as we can," explained one principal. "But if we get to the point where we have to decide between cutting a literacy teacher and a Reading Recovery teacher, I know that the Reading Recovery teacher will have to be cut, just because of the sheer number of students [classroom teachers] serve."

Attrition and administrator turnover

A second challenge to sustaining implementations is administrator turnover. When district or school leadership changes, the potential for attrition can increase. New leaders may have a different perspective on the cost-benefit analysis of Reading Recovery, or different priorities in terms of curriculum or instruction. UTC directors and teacher leaders tell us that administrator turnover is a significant source of attrition; to date, approximately 52% of the teachers who have exited Reading Recovery have cited school or district decision as the reason. One teacher leader explained how this might happen:

> If you've got someone on the [school] board, or a superintendent that knows Reading Recovery and is a supporter, they'll fight to get it into their district. The same with if you have a superintendent who has different views of it. It can go away really fast.

At the district level, the entrance of an administrator who does not support Reading Recovery can lead to attrition in a few ways. Some simply eliminate the program from the district entirely, while others exert influence on principals, persuading them to stop dedicating funding to Reading Recovery. One UTC director described an example of the latter scenario. She explained that the new assistant superintendent in a district "was very negative about Reading Recovery" based on her involvement with the Reading First initiative.[7] Because of this perspective, she "influenced a lot of new principals that Reading Recovery just wasn't effective, that you were better off to do small-group interventions or to have smaller class sizes. And so that really made a big difference." Other UTC directors also described situations where principals felt pressure to eliminate Reading Recovery based on central administrator preference, even though they were personally inclined to keep the program in their buildings.

At the school level, attrition is largely influenced by how well principals understand Reading Recovery. Those who understand the program are more likely to feel that its benefits justify the expense, and are often less vulnerable to pressure imposed by a central administrator who

7 As detailed in the Year One report, Reading Recovery suffered significant setbacks following the passage of the Reading First legislation in 2001. An investigation by the Office of the Inspector General concluded the implementation of the Reading First program by the U.S. Department of Education involved numerous processes that resulted in conflicts of interest and bias against specific reading programs (OIG, 2006).

does not support the program. However, not all principals are well informed about Reading Recovery, and the entrance of a new school administrator can quickly threaten school-level implementation. One teacher leader described an experience in which a school replaced the principal mid-year: "Once they lost that principal, no one was really supporting [Reading Recovery]. So when they began to have personnel situations, they just pulled those Reading Recovery teachers." Unfortunately, as one UTC director explained, sometimes administrator turnover occurs when a Memorandum of Agreement is already in place at a school. "The major problems related to attrition," she commented, "are caused by administrators… not safeguarding the original or keeping the original agreement." She went on to say:

> The Memorandum of Agreement is only as good as the elementary principal whose interest is in adhering to commitments made… When a new administrator comes in they don't necessarily feel they are obligated to uphold a commitment made by the previous administrator… I just think that administrator commitment is huge. The administrator's behind it, or if they're just going through the motions they can make a teacher's scheduling of the daily lessons difficult… The administrator plays a key role.

To address attrition based on administrator preference or understanding, teacher leaders and UTC directors again rely heavily on local student data. For districts in which Reading Recovery has existed for years, one UTC director observed, there are enough data to be effective protection against decline. "I don't have to advocate much to people about how good the program is," she explained, "because we have twenty years of data showing that it is." However, not every school or district has long-term data – particularly those that have been recruited over the course of the i3 scale-up. Using data to prevent attrition can be difficult in such settings. One teacher leader described this challenge:

> If the data does not support the intervention, they'll find something else… So we are always trying to get data to the superintendents and all the academic officers under the superintendent to make sure that they see the success of Reading Recovery and so that they will keep us around.

The data currently available indicate that, three years after the allocation of the i3 grant, Reading Recovery teacher leaders and UTC directors have made significant progress toward the scale-up goal for recruiting and training new teachers. A comprehensive discussion of progress toward all the goals of the scale-up–teacher, teacher leader, and school recruitment–will be included in the final report.

V. Implementation: Fidelity

A thorough assessment of implementation fidelity is increasingly regarded as a key component of program evaluation (Dusenbury, Brannigan, Falco, & Hansen, 2003; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008). Information about program effects is of limited use without an understanding of how the effects were achieved. In addition, detailed information about how programs are implemented is essential to efforts to replicate effective interventions (Yeaton & Sechrest, 1981).

However, recent studies have documented a lack of consensus in education literature about "what exactly fidelity of implementation is, how it is measured, or how program theory or study design relates to fidelity of implementation" (O'Donnell, 2008, p. 40). In parsing implementation fidelity, it seems, it falls to the evaluator to grapple with questions about how fidelity should be defined in the context of a given intervention and where fidelity to the program model begins and ends (Century, Rudnick, & Freeman, 2010; Summerfelt, 2003).

A research objective in each year of the i3 scale-up evaluation of Reading Recovery is to measure the nature and extent of variation in implementation fidelity. Thus, CPRE has reflected on these questions and the concept of fidelity in the context of Reading Recovery implementation: how implementation fidelity is best defined and measured, where it provides useful insight, and when it may be an inadequate lens on implementation. These questions give shape to our work, and they inform the findings presented in this report. Additionally, our measurement approach and methodologies for constructing fidelity indices were informed by contributions from the National Evaluation of i3 (NEi3) team.

This section details our current approach to conceptualizing implementation fidelity in the context of Reading Recovery, our process for identifying and understanding the components of implementation fidelity, the methods we used to measure and assess fidelity in Year Two, and our findings in this area.

Delineating Implementation Fidelity

Implementation of Reading Recovery comprises activities and processes that span a number of domains, and that depend on the participation of multiple players at the university, district, site, and school levels. These activities and processes are described in detail in the Year One report, and are illustrated by the original Implementation Logic Model CPRE developed in Year One (see Appendix B). Many, though not all, of the activities represented in this model are also reflected in the Standards and Guidelines–the published manual that guides Reading Recovery implementation.

Over the course of our two years of research on Reading Recovery, CPRE has developed a deep understanding of the complex web of activities that constitutes the program's implementation.

This evolving understanding has led to a refinement of our thinking about implementation fidelity over time. In Year One we used the Standards and Guidelines, in their entirety, as the basis for our study of implementation fidelity. This comprehensive fidelity analysis identified some activities that occur at considerable distance from Reading Recovery's core functions of teacher training and lesson delivery–for example, activities related to program oversight at the district level. While we recognize their importance to the functioning of the Reading Recovery system and the quality of the program, it became clear that these activities are not central to the intervention itself.

With these lessons in mind, we approached Year Two with the goal of constructing a framework for implementation fidelity that focused on the activities we regard as most critical to adherent implementation of the Reading Recovery intervention. To that end, we refined our definition of implementation fidelity to focus on activities that meet three criteria:

1. The activities are represented as standards in the Standards and Guidelines.
2. The activities are essential to the core functions of Reading Recovery: the training of teachers and the provision of one-to-one lessons.
3. The activities are performed by core Reading Recovery personnel: university trainers, teacher leaders, and Reading Recovery teachers.

Activities that satisfy all three of these criteria are identified as Implementation Fidelity Activities to distinguish them from other aspects of implementation. Only these Implementation Fidelity Activities are included in our fidelity analysis. This approach to measuring fidelity is informed by, and consistent with, the NEi3 framework for high-quality implementation studies.

In establishing the criteria for Implementation Fidelity Activities, we draw some distinct lines informed by our findings from Year One. For example, we specify that Implementation Fidelity Activities are performed by core Reading Recovery personnel, identified as university trainers, teacher leaders, and Reading Recovery teachers. This excludes some of the other players involved in implementation. Most notably, we excluded site coordinators from this definition even though their role is explicitly outlined in the Standards and Guidelines. This decision resulted from a great deal of consideration. It was driven by our research findings, which suggest that while site coordinators are *in* the Reading Recovery system, they are, importantly, very often not *of* it. Site coordinators are generally district-level administrators with responsibility for overseeing many instructional initiatives and activities. Site coordinators therefore assume Reading Recovery duties as just one small part of a broad and multifaceted role. They do not necessarily have deep knowledge about Reading Recovery, and they often have little or no contact with Reading Recovery teachers or students. While they can play a powerful role in implementation at the system level, we find that site coordinators generally

remain several steps removed from the core teacher-training and lesson-delivery functions we identify as most critical. For these reasons, we do not regard their activities as components of implementation fidelity.

The difficulties of capturing complex program models in fidelity frameworks is well documented (Century et al., 2010; O'Donnell, 2008), and the example of the Reading Recovery site coordinator exemplifies this challenge. We acknowledge the complexity of this balance, and its significance to our work. Our efforts to capture this complexity underlie our decision to pair our focus on implementation fidelity with a parallel examination of other non-fidelity-related features of Reading Recovery that have surfaced through our two years of research, and that we hypothesize are critical to high-quality implementation. Through this approach, we hope to avoid the pitfall of overreliance on fidelity as an explanatory tool, while at the same time examining the critical aspects of implementation concretely enough to support a productive analysis of variation in impacts. Our examination of non-fidelity features of implementation, including the role of the site coordinator, is ongoing and will be included in our final report.

The Implementation Fidelity Logic Model

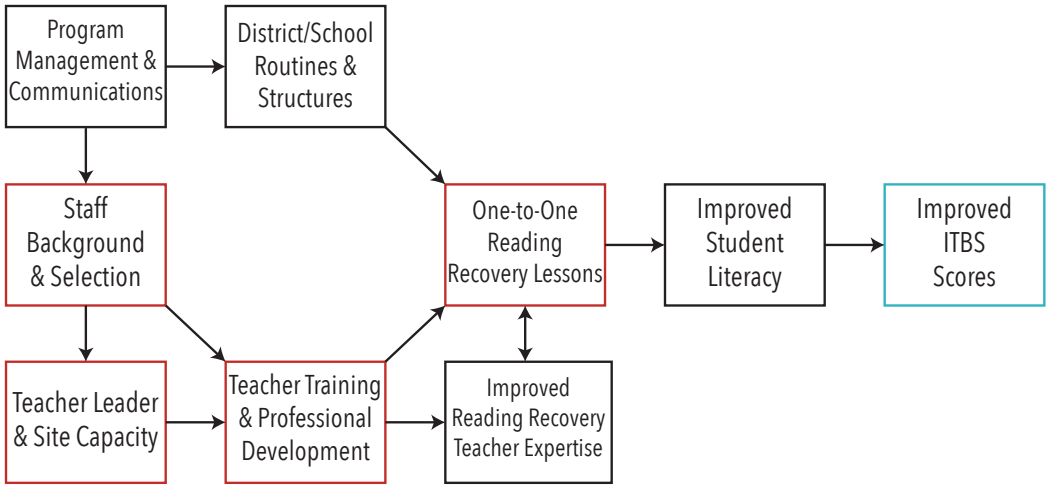
Once CPRE identified the Implementation Fidelity Activities for inclusion in the Year Two fidelity analysis, we found it helpful to group them into broad categories of related activities. These categories both facilitate the analytic approach described later in this section and enable us to represent, as concisely as possible, the complex set of activities reflected in the Standards and Guidelines and the original Implementation Logic Model presented in Section I.

Drawing on both the general structure of the Standards and Guidelines and our Year One research findings, we grouped the Implementation Fidelity Activities into the following four categories:

- Staff Background & Selection: This category includes standards that specify the selection criteria for teachers trained in Reading Recovery and teacher leaders.
- Teacher Leader & Site Capacity: This category includes standards that specify the training experience of teacher leaders as well as the standards that characterize the training environment.
- Reading Recovery Teacher Training & Ongoing Professional Development: This category includes standards that specify the training and continuous professional development experience of trained and in-training Reading Recovery teachers.
- One-to-One Reading Recovery Lessons: This category includes standards that specify the selection, assessment, and instruction of individual Reading Recovery students.

Figure 1 shows the Implementation Fidelity Logic Model CPRE developed over the course of Year Two. This model depicts our expectations regarding the relationships between these four categories of Implementation Fidelity Activities (outlined in red) and the primary program outcome–student achievement–via two mediators: improved Reading Recovery teacher expertise and improvements in student literacy.

Figure 1. Implementation Fidelity Logic Model



As Figure 1 illustrates, staff background and selection activities are expected to facilitate both teacher leader/site capacity and Reading Recovery teacher training by ensuring that qualified individuals are selected for key roles related to the training of teachers and provision of one-to-one instruction. Training and ongoing professional development are expected to support both a teacher's capacity to conduct the lesson and development of that teacher's expertise in early literacy and formative assessment. As such, the one-to-one Reading Recovery lessons are depicted as both a direct outcome of the training and professional development, and an indirect outcome of training via the effect on improved teacher expertise. Improved teacher expertise is also supported through ongoing implementation of the Reading Recovery lessons. This is specified by a reciprocal arrow. Improved teacher expertise is expected to mediate the relationship between training and provision of instruction. Participation in the one-to-one lessons is expected to produce improvements in students' literacy, which in turn produces improvements in ITBS scores. Thus, the streamlined Implementation Fidelity Model shown above represents our understanding of how these broad categories of activities–each encompassing a number of Implementation Fidelity Activities–progress toward and directly support a single, prioritized outcome: gains in student achievement as measured by the ITBS.

While recognizing its importance, we are also mindful of the limitations of fidelity as a lens for understanding implementation. Our two years of research have revealed that assessing implementation fidelity is necessary but not sufficient for understanding high-quality

implementation of Reading Recovery. We therefore underscore the contributions of some other features of Reading Recovery implementation–features that do not fit our criteria for Implementation Fidelity Activities–to the overall quality of the program. Figure 1 reflects this understanding by pairing Implementation Fidelity Activities, represented by the red boxes, with other types of implementation activities, represented by the black boxes. These complementary activities are discussed further in Section VI of this report.

Measuring implementation fidelity

Fidelity data are collected annually in the spring, via online surveys that are administered to the full population of Reading Recovery teachers and teacher leaders involved with the i3 scale-up. Response rate calculations are based on population data received from IDEC.[8] The surveys were designed to collect data on implementation efforts in relation to the expectations detailed in the Standards and Guidelines. The Year Two findings allowed for quantification of implementation fidelity by activity category.

CPRE operationalized the Reading Recovery Standards and Guidelines as questions on various survey instruments. The Standards for the implementation of Reading Recovery are mandated by the program; however, the Guidelines are recommendations that may improve the quality of the program, but are not required. To acknowledge this distinction, CPRE quantified fidelity to the Reading Recovery model using only the required standards, and only those that met the criteria described earlier in this section. The applicable standards for program implementation were each measured by at least one survey item from one or more respondent groups. Fixed-response data from each survey were used to create indices of implementation fidelity, represented as percentage of standards met.

The surveys were administered via Qualtrics™, an online survey platform. Web links were emailed to individual Reading Recovery teachers and teacher leaders. We took several precautions to avoid problems related to self-reported data; for instance, when measuring standards, respondents were asked to report objective facts about Reading Recovery implementation in their context, not to make evaluative judgments about themselves or their own contributions. In addition, to the extent possible, questions were carefully worded so as to not suggest that one response was correct or more appropriate than the others.

Reading Recovery teachers and teacher leaders who are in training are expected to adhere to different Standards and Guidelines than those who are already trained. For example, Reading Recovery teachers in their training year are required to attend weekly classes, whereas trained teachers meet only occasionally. To account for these differences, CPRE calculated the percent of standards met by each respondent using only those relevant to his/her specific position and stage in the training process. In addition, a majority of the teacher leaders were trained

8 In Year Two, response rates for Reading Recovery teachers and teacher leaders were 76% and 80% respectively.

in the first year of the i3 scale-up; we include their training experiences in the analysis of Teacher Leader training fidelity for Year Two as well. Overall percentages were calculated as the unweighted average of standards that were applicable to a given respondent.

Going forward with the evaluation, we will continue to administer the Reading Recovery teacher survey to all teachers whose training was supported by i3 funds at any time during the grant period. We will also continue to administer the teacher leader survey to all who work with at least one teacher supported by i3 funding. Fidelity indices will be calculated based on all response data in a given year.

Fidelity indices

CPRE used a two-step process to assess overall fidelity of implementation in Year Two of the scale-up evaluation. First, we examined the extent to which individual Implementation Fidelity Activities had adequate or inadequate implementation by calculating the percentage of respondents who indicated that the standard related to that activity had been met. Adequate implementation was defined as 80% adherence to the full definition of the indicator.

As a second step, we assessed whether each of the four activity categories was implemented with fidelity or not with fidelity. As discussed above, each category includes a set of related Implementation Fidelity Activities. To conduct this analysis, we determined what percentage of Implementation Fidelity Activities within a given category met the benchmark of 80% compliance. If 80% or more of the individual activities within the category met the benchmark, the category as a whole was found to be implemented with fidelity. While the Implementation Fidelity Matrix distinguishes between indicators that are assessed for Reading Recovery teachers and teacher leaders, all indicators were weighted equally in determining the extent to which a category of activities was implemented with fidelity.

The North American Trainers' Group (NATG), the governing body of Reading Recovery in the U.S., has developed a waiver process by which university trainers can exempt sites from meeting particular standards with proper justification. The waiver process requires that a teacher leader make a formal, written request to deviate from a particular standard. A university trainer determines whether the deviation is permissible and notifies the schools and regional training center of the allocation or refusal of a waiver. If granted, the waiver remains in effect for a single year. CPRE surveys asked respondents to specifically indicate when a waiver had been granted for a given indicator, and non-compliance with standards was not considered a deviation from fidelity when waivers had been approved. We will continue to study the use of waivers in the Reading Recovery system, and their impact on implementation fidelity, going forward with this evaluation.

Findings: Strong fidelity overall

It is clear that Reading Recovery teachers and teacher leaders are implementing the Reading Recovery model with high fidelity. Overall, 85% of the indicators used to assess implementation fidelity had adequate implementation, and all four categories of Implementation Fidelity Activities represented in the streamlined Implementation Fidelity Logic Model (Figure 1) were found to be implemented with fidelity.

Table 14 presents the number and percent of indicators that were adequately implemented by category. Fidelity by category ranged from 81-88%. All categories are above the 80% threshold CPRE established for fidelity.

The overall picture of fidelity presented by the fidelity indices comports with qualitative findings that implementation fidelity is strong. Examination of the individual indicators also reveals consistency between quantitative and qualitative fidelity findings. The percent of teachers who adequately implemented each of the indicators ranged from 60-100%, with the majority of indicators met by more than 90% of respondents. The table in Appendix B details fidelity findings by standard, and includes information about the instrument used to assess fidelity to each standard.

Table 14: Implementation Fidelity of Key Components

Component	# indicators	# adequately implemented	% adequately implemented
Staff Background & Selection	7	6	86
TL & Site Capacity	21	18	86
RRT Training & Ongoing PD	16	13	81
One-to-One RR Lessons	8	7	88

Note. 52 implementation fidelity indicators were assessed in 2013.

Deviations from implementation fidelity

While all four categories of Implementation Fidelity Activities were found to have been implemented with fidelity in Year Two of the evaluation, departures from fidelity were observed for eight individual activities. These are the Implementation Fidelity Activities for which CPRE found less than 80% compliance with the relevant standards:

Commitment to Implementation. A critical factor that can influence school-level implementation is the school system's commitment to Reading Recovery. In the activity category of Teacher Leader and Site Capacity, Standard 3.01 applies to Reading Recovery teachers: Be employed in a school system that has a commitment to implementation. This standard was defined and measured by whether the Reading Recovery teacher reported that

her district intends to train enough teachers in Reading Recovery to reach full implementation. Districts that were at full implementation or had a plan to reach full implementation were considered to meet the standard. Forty percent of the responding Reading Recovery teachers reported that their districts were both below full implementation and did not have a plan to hire additional Reading Recovery teachers.

Instruction by Teacher Leaders. In the activity category of Teacher Leader and Site Capacity, Standard 4.14 applies to teacher leaders in training: Teach four Reading Recovery children per day individually for 30-minute sessions in a school setting throughout the school year. Seventy-eight percent of teacher leaders in training met this standard–nearly enough to achieve overall adequate implementation. However, maintaining the required teaching load was one area in which teacher leaders had relatively lower fidelity.

University Trainer Visits. Also in the category of Teacher Leader and Site Capacity, two standards related to colleague visits had inadequate implementation. Standard 4.17 applies to teacher leaders in training: Receive at least four visits from a university trainer. Only 64% of teacher leaders in training reported having received four visits from a trainer during the school year.

Standard 4.66 applies to teacher leaders in their field year: Receive a minimum of two site visits from a trainer during the teacher leader's first year(s) in the field. Receive at least one site visit during the first year after a teacher leader has changed the site of employment. Seventy-eight percent of teacher leaders in their field year reported having received the required number of site visits.

Achieving Numbers for Training. Reading Recovery teachers reported overall high fidelity to standards for training and ongoing professional development. One exception, however, was related to the size of teacher training classes. In the activity category of Teacher Training and Professional Development, Standard 2.01 applies to sites: Train classes of at least 8 and not more than 12 teachers. Only 66% of Reading Recovery teachers reported that their class size was between 8 and 12 teachers.

Ongoing Professional Development. Also in the activity category of Teacher Training and Professional Development, Standard 3.44 applies to trained Reading Recovery teachers: Participate in a minimum of six professional development sessions each year, including a minimum of four behind-the-glass sessions with two lessons each session. Seventy-six percent of trained Reading Recovery teachers reported participating in at least six professional development sessions each year, including a minimum of four behind-the-glass sessions that had two lessons each.

Standard 3.43 in this category applies to trained Reading Recovery teachers: Consult with the teacher leader about children not making satisfactory progress and other issues. Only 78% of trained Reading Recovery teachers reported consulting their teacher leader in this way.

Student Selection. An important issue in school-level implementation of Reading Recovery is how students are selected to receive the treatment. In the activity category of One-to-One Reading Recovery Lessons, Standard 2.05 applies to sites: Select the lowest-achieving children for service first (based upon Observation Survey tasks) in all decisions. Based on survey data, Reading Recovery teachers were found to have inadequate implementation of this standard; 77% of teachers reported that children with the lowest scores on the OS were selected for service first. According to the Standards and Guidelines, the lowest achieving children—as measured by the OS alone—should always be selected for service first. However, CPRE has observed in each of the first two years of the evaluation that student selection is a variable process across schools and districts. In our estimation, this represents a meaningful deviation from implementation fidelity in terms of its potential implications for program impacts.

In Year One, anecdotal accounts pointed to considerable variability in the ways schools and districts assigned students to Reading Recovery. Many teachers reported that, in recognition of the barriers some students face to learning, their school had chosen to exempt certain students from the Reading Recovery intervention despite their low scores on the OS. Reading Recovery teachers reported that the most common reasons for eliminating students were related to perceptions that the student would not be able to benefit from Reading Recovery for academic or personal reasons, or the fact that the student was already receiving, or was expected to receive, other types of services.

In Year Two, we surveyed Reading Recovery teachers to explore this issue in greater depth. Reading Recovery teachers responding to the survey (N=1,511) were asked to report whether any of their schools' lowest scorers on the OS were not selected to receive Reading Recovery, and then asked to indicate the reasons why any low scorers were not chosen. Table 15 details the reasons teachers provided. The most frequently cited reason for deviation from the standards concerned the exclusion of students with Individualized Education Plans (IEPs) for special education services. Forty three percent of the time that Reading Recovery teachers indicated one or more of their schools' lowest scorers were excluded from receiving Reading Recovery, it was because of having an IEP or receiving other services. Our interviews with teachers, teacher leaders, and principals over both Year One and Year Two suggest, similarly, that this was the most common reason students were excluded from Reading Recovery, and that schools often establish this policy in order to distribute intensive intervention services to more low-achieving students. Those with IEPs in literacy are generally receiving one-on-one support already.

While schools that elect not to serve students with attendance or behavior problems are clearly out of step with Reading Recovery policy, there is less clarity around the exclusion of students with IEPs. The Standards and Guidelines state that all students should be served regardless of disability; however, other Reading Recovery documents recommend excluding those with IEPs for literacy on the grounds that they are already receiving intensive services for reading (RRCNA, 2002). Reading Recovery 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. Where the selection of students with IEPs is concerned, our key finding is that while many schools are not adhering to the Standards and Guidelines, most are making a good-faith effort to comply with their understanding of Reading Recovery 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 deviations from the Standards and Guidelines around student selection do appear to be common, our research suggests that in spite of these differences in the way schools operationalize the selection process, the students selected for Reading Recovery are, across the board, very low-achieving and are routinely among the lowest first-grade readers in their schools.

Table 15: Reasons for exclusion of students from Reading Recovery

Reason for Exclusion	Percent of reasons for exclusion
Student had IEP / receiving other services	43
Student is repeating first grade	20
Student has attendance problems	8
Student is an English Language Learner or has limited language skills	8
School is under-implemented / not enough RR teachers	8
I don't know	5
Student had behavior problems	5
Other reasons	3
Student's parents did not consent to participation	1
Total	100

While excluding students with IEPs represents a deviation from the Standards and Guidelines, there is conflicting guidance on this issue. For example, the Principal's Guide to Reading Recovery (RRCNA, 2012) recommends that students with IEPs in literacy not be served. In conversations with both school and Reading Recovery personnel, CPRE also encountered considerable variation in understanding about this issue. Indeed, while some teacher leaders and UTC directors report the belief that students with IEPs in literacy should be excluded to avoid doubling up on services, others indicate belief more consistent with the Standards and Guidelines' insistence that all students be served.

On a final note, multiple sources of data–both quantitative and qualitative–suggest that despite this variation in school-level policies on student selection, the students receiving Reading Recovery lessons are consistently among the lowest achievers in literacy in their schools.

CPRE will continue to explore this issue, including whether it has implications for school-level effect sizes, in future research.

VI. Conclusion

Consistent with the Year One report, this year's evaluation finds significant positive impacts of Reading Recovery on students' reading achievement, and considerable progress toward the scale-up goals. Many of the Year One findings remained consistent in Year Two, including the importance of ongoing support of both Reading Recovery teachers and teacher leaders, and the importance of school and district leadership buy-in to Reading Recovery.

Recruitment and retention

The i3 award has provided an opportunity for the Reading Recovery community to expand the intervention to schools and districts that are under-implemented or new to the program. Specifically, the objectives for the scale-up include recruiting and training 3,675 new Reading Recovery teachers, with the goal of expanding Reading Recovery lessons to an additional 67,264 students over the life of the grant, and providing other instruction to an additional 302,688 students through classroom or small-group instruction. CPRE finds that, to date, the scale-up is on track to achieve these targets. However, some challenges persisted in Year Two. Teacher leaders and UTC directors again reported that a difficult fiscal climate, high turnover of local decision-makers, and misunderstandings about the program itself each challenged Reading Recovery advocates in recruiting and retaining schools.

Over two years of research, we have developed a relatively clear picture of the issues surrounding recruitment of teachers and schools, including strategies and challenges. Going forward, we are interested in understanding more about why some implementations are more sustainable than others. In Year Two we continued to explore some of the factors that may contribute to attrition and retention of teachers, schools, and sites, and the ways Reading Recovery personnel are thinking about and addressing these issues. Teacher leaders and UTC directors reported that the best way to secure school and district commitment to Reading Recovery is to show evidence of student growth, and that student outcomes data is the essential tool in both recruitment and retention. However, we observe that the reasons some schools fail to adopt or sustain Reading Recovery are rarely related to doubts about the program's effectiveness. Most often, these decisions are based on financial issues that are not easily overcome, even with data showing the significant positive effects of Reading Recovery. This begs the question of how best to address challenges to sustainability. CPRE will continue to explore this question in future research.

Impacts and variation

The Year Two impact analysis reveals statistically significant positive impacts on students' reading achievement. The standardized effect sizes, between 0.36 and 0.42 standard deviations, are large relative to typical effect sizes found in educational evaluations, and

represent an extra 1.4 months of learning. Even when benchmarking the impacts on ITBS scores relative to the full population of first graders in the nation, the standardized effect sizes between 0.32 and 0.33 standard deviations are still large (2.7 times the average effect of Title I programs) and represent a growth rate that is 27% greater than the national average for first graders. These findings are consistent across student subgroups, including students in rural schools and English Language Learners.

Results also showed substantial variation in effect estimates across schools. The vast majority of schools experienced positive impacts, and some schools produced effect estimates that were many times larger than those of typical reading interventions. In addition, this year's analyses revealed that schools with lower than average ITBS scores tended to have larger treatment effects. The cause for this relationship is yet unknown, and because this result did not appear in our Year One analysis, we are hesitant to draw any conclusions at this time. However, this relationship will be of primary interest when we pool data across years for our overall impact analyses for the final report.

Fidelity analysis

Based on our research in Year One, we refined our understanding of fidelity to focus on activities that meet the following criteria:

1. The activities are represented in the Standards and Guidelines.
2. The activities are essential to the core functions of Reading Recovery: the training of teachers and the provision of one-to-one lessons.
3. The activities are performed by core Reading Recovery personnel: university trainers, teacher leaders, and Reading Recovery teachers.

Activities that met all three of these criteria were identified as Implementation Fidelity Activities. Overall, CPRE observed high implementation fidelity across schools participating in the i3 scale-up. Our analysis reveals that Reading Recovery teachers and teacher leaders are implementing the program with high fidelity. Overall, 85% of the indicators used to assess implementation fidelity had adequate implementation.

We observed eight standards with significant deviations from fidelity. The most notable of these concerns the processes by which students are selected to receive Reading Recovery instruction. Our findings from the first two years of the evaluation indicate that student selection, though codified in the Standards and Guidelines as a straightforward process, can vary substantially from site to site. We note that while the students selected are consistently among the lowest achievers in literacy in their schools, many schools exclude students who meet certain criteria–for instance, those receiving special education services. We also note conflicting messages and understandings from within the Reading Recovery community about

whether all students should be served, as the Standards and Guidelines indicate, or whether those with IEPs in literacy should be excluded to avoid doubling up on services.

Ongoing questions

The few deviations from fidelity we observe reveal a tension between strict interpretation of some standards and local realities in schools. The selection of students to receive Reading Recovery is the primary example of this tension. We observe that student selection decisions do not always rely solely on the OS, and that some schools and districts choose to exclude groups of students (e.g., students with IEPs, as noted, or those repeating first grade) from receiving Reading Recovery. While these decisions are departures from the standards, they are adaptations that some schools and districts make with the intention of maximizing the impact of the intervention on their student population. The implications of compliance with, or deviation from, the program model will be a focus of ongoing study.

In addition, we will increasingly focus on the relationship between implementation and program impacts. In order to do this, we have worked to develop a comprehensive understanding, and an approach to measuring, Reading Recovery's implementation. In the past, discussions of implementation in the context of experimental studies have generally been framed in terms of fidelity (Century et al., 2010; Bauman, Stein, & Ireys, 1991; Summerfelt, 2003; Hulleman & Cordray, 2009; Weiss, Bloom, & Brock, 2013); the implicit presumption of the bulk of this work is that fidelity is an adequate measure of implementation. However, we have consistently observed that implementation of Reading Recovery is more complex than can be understood by an exclusive focus on fidelity. Understanding the relationship between fidelity and program quality overall, and the implications of both fidelity and quality for program impacts, will be a focus of our work in the project's final year.

Appendix A: Statistical Model for Impacts of Reading Scores

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

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

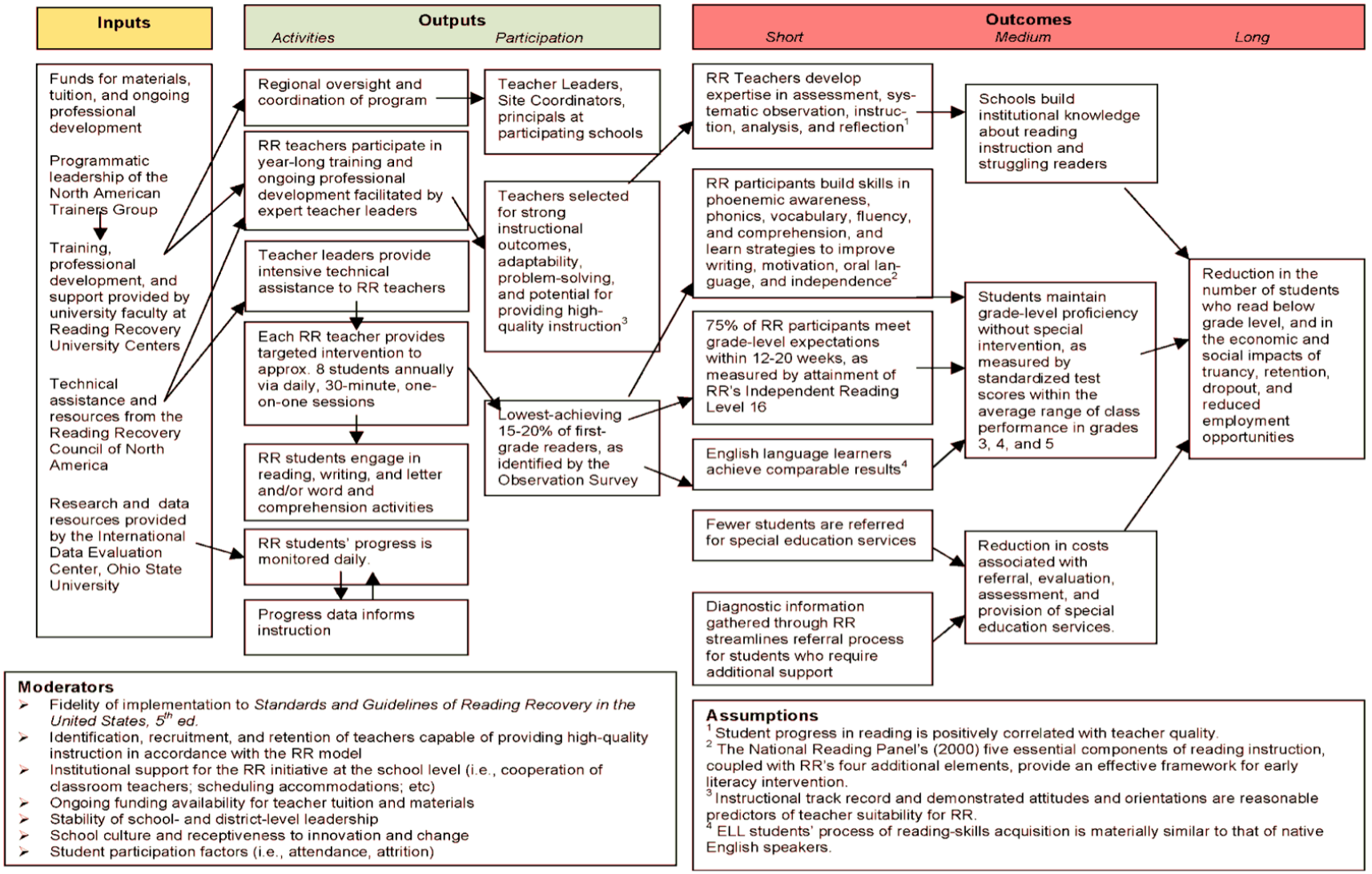
With:

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

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 Observation Survey)
β_2	is the overall treatment effect
Trt	is the treatment assignment indicator, 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
ε_{ijk}	is the student-level residual, with variance σ^2

Appendix B: Reading Recovery Implementation Logic Model



Appendix C: Fidelity Findings by Standard

	Standard	Instrument	% Met	# Applicable
Staff Background & Selection				
Not Met	3.01	RRT Survey	60	689
Met	3.02	RRT Survey	100	1279
Met	3.03	RRT Survey	96	1279
Met	4.01	TL Survey	100	214
Met	4.02	TL Survey	100	214
Met	4.04	TL Survey	96	214
Met	4.05	TL Survey	84	33
Teacher Leader & Site Capacity				
Met	2.11	TL Survey	96	205
Met	2.14	TL Survey	80	204
Met	4.11	TL Survey	100	4
Met	4.13	TL Survey	94	33
Not met	4.14	TL Survey	78	32
Met	4.16	TL Survey	100	18
Not met	4.17	TL Survey	64	17
Met	4.18	TL Survey	100	4
Met	4.24	TL Survey	88	32
Met	4.25	TL Survey	100	4
Met	4.36	TL Survey	100	18
Met	4.37	TL Survey	100	18
Met	4.39	TL Survey	88	25
Met	4.40	TL Survey	94	204
Met	4.55	TL Survey	80	204
Met	4.49	TL Survey	82	197
Met	4.59	TL Survey	96	201
Met	4.65	TL Survey	98	204
Not met	4.66	TL Survey	78	9
Met	4.67	TL Survey	98	204
Met	4.68	TL Survey	98	204

	Standard	Instrument	% Met	# Applicable
Training & Ongoing PD				
Not met	2.01	RRT Survey	66	590
Met	2.19	TL Survey	98	213
Met	3.11	RRT Survey	90	590
Met	3.13	RRT Survey	90	590
Met	3.15	RRT Survey	80	590
Not met	3.43	RRT Survey	78	689
Not met	3.44	RRT Survey	76	689
Met	3.45	RRT Survey	94	689
Met	4.42	TL Survey	100	12
Met	4.43	TL Survey	98	106
Met	4.44	TL Survey	80	105
Met	4.46	TL Survey	98	105
Met	4.47	TL Survey	86	106
Met	4.48	TL Survey	94	105
Met	4.50	TL Survey	100	136
Met	4.51	TL Survey	100	135
One-to-One Reading Recovery Lessons				
Not met	2.05	RRT Survey	77	1373
Met	2.06	RRT Survey	80	689
Met	2.17	RRT Survey	88	1279
Met	3.17	RRT Survey	90	589
Met	3.25	RRT Survey	100	589
Met	3.29	RRT Survey	84	1278
Met	3.32	RRT Survey	84	1278
Met	3.37	RRT Survey	98	689

Appendix D: Pretest & Outcome Measure

The Iowa Test of Basic Skills (ITBS) was the outcome measure for the impact analysis. The ITBS is a well-regarded, group-administered, norm- and criterion-referenced, standardized assessment designed to "assess the extent to which a child is cognitively ready to begin work in the academic aspects of the curriculum" (Hoover et al., 1994, as cited in Tang & Gomez, 2007), and to "measure growth in fundamental areas of school achievement" (Hoover et al., 2003, p.1).

Originally published in 1955, the ITBS is currently available in two forms, A and B, which are broken into multiple parts and subtests that measure achievement for students in kindergarten through the eighth grade. Part 1 of the ITBS is a teacher-administered subtest, while Parts 2-6 are student-administered following a teacher-modeled example. The analysis performed in this evaluation to determine overall program impacts used three scores from the ITBS Reading subtest. These were the Reading Words and Reading Comprehension subtests, used for exploratory analysis, and the "Reading Total" score, used for confirmatory analysis. The parts that comprise each score are as follows:

- Reading Words: Words (Part 1), Pictures (Part 2), and Word Attack (Part 3)
- Reading Comprehension: Sentences (Part 4), Picture Story (Part 5), and Story (Part 6)
- Reading Total: Parts 1-6

These test components were chosen for several reasons. First, the battery of tests utilized during this study (ITBS form A, level 6), is appropriate for students who are six years old and whose level of academic development ranges between K.8 to 1.9. Second, ITBS raw scores can be converted to several other types of scores, including developmental scores (grade equivalents), developmental standard scores, and status scores. Finally, the national standardization of the ITBS was conducted with a normative sample designed to represent the national population of school children, grades kindergarten to eight (Hoover, Dunbar, & Frisbie, 2011).

Information regarding the technical characteristics of the ITBS was obtained through the Guide to Research and Development (GRD Manual), the ITBS technical manual. The GRD Manual contains multiple reliability coefficients (internal consistency, equivalent forms, test-retest), most of which range between the middle .80s to low .90s. Designed to "measure growth in the fundamental areas of school achievement" (Hoover et al., 2003, p.1) - including vocabulary and reading comprehension - the ITBS manual provides sound evidence to support the instruments' content validity and high discriminant ability (item p-values and discrimination indices) (Hoover et al, 2003). Additionally, the ITBS has often been used as an outcome measure for both experimental and quasi-experimental impact studies (Kim & White, 2008; Reis et al., 2008; Jenner & Jenner, 2007). In all, the ITBS is regarded as a well-developed assessment with sound technical qualities established through rigorous processes.

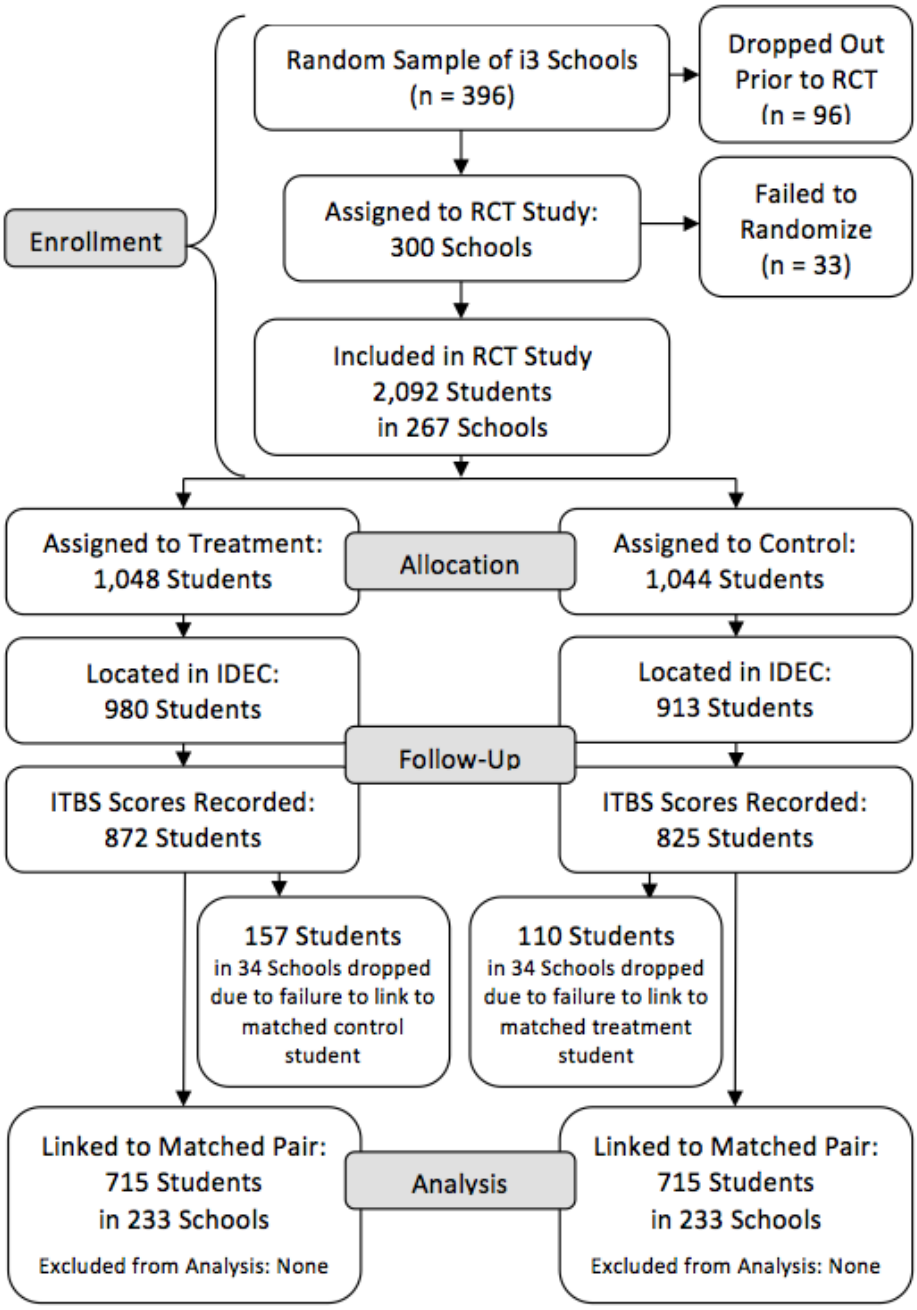
The Pretest Measure: The Observation Survey of Early Literacy Achievement

The Observation Survey of Early Literacy Achievement (OS), developed by Marie Clay, is the primary screening, diagnostic and monitoring instrument for Reading Recovery. It was used as the pretest measure. The OS is a one-to-one, teacher- administered, standardized assessment that includes six sub-scales: Letter Identification,

Concepts about Print, Ohio Word Test, Writing Vocabulary, Hearing and Recording Sounds in Words, and Text Reading Level. The Text Reading Level (TRL) is used to block students during the random assignment process, and later as a pretest covariate in the statistical models of impacts.

Through the administration of the Text Reading OS task, which requires the administration of a running record in which teachers record speed and accuracy with which a student reads a selected text with known difficulty, RR teachers determine a student's text reading level, defined by RR as the level at which a student can read a set of texts, specifically the Scott Foresman Special Practice Books, with at least 90% accuracy (NATG, 2005). Multiple methods have been employed to estimate the reliability of the OS. Reported test-retest and internal consistency reliability estimates range from moderate to high on the individual OS Tasks (Clay, 2002, as cited in Denton, Ciancio, & Fetcher, 2006); measures of the inter-assessor reliability of the Text Reading and Writing Vocabulary tasks yielded coefficients of .92 and .87 (Denton et al., 2006). In addition, evidence of the validity of information yielded through administration of the OS has been provided by several studies that assess the construct and criterion validity of the OS tasks using the sub-tests of various norm-referenced tests, including the Iowa Tests of Basic Skills (ITBS). Across these studies, researchers have found that scores can be validly interpreted for the following purposes: (1) identification of at-risk students (Gomez, Rogers, Wang, & Schultz, 2005), (2) measurement of early reading constructs (Tang & Bellenge, 2007; Gomez, Gibson, Tang, Doyle, & Kelly, 2007), and (3) prediction of the attainment of performance benchmarks (Denton et al., 2006).

Appendix E: Consort Flow Diagram through the Reading Recovery i3 RCT 2012-13



References

Bauman, L. J., Stein, R. E. K, & Ireys, H. T. (1991). Reinventing Fidelity: The transfer of social technology among settings. *American Journal of Community Psychology, 19:* 619-639.

Bloom, H. S., Raudenbush, S. W., & Weiss, M. Under review. Estimating Variation in Program Impacts: Theory, Practice, and Applications. New York: MDRC.

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*(4), 309-326.

Century, J., Rudnick, M., & Freeman, C. (2010). Fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation (31)*2: 199-218.

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

Clay, M. (2005a). *An observation survey of early literacy achievement.* Portsmouth, NH: Heinemann.

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

Clay, M. (2005c). *Literacy lessons designed for individuals*: *Part Two.* Portsmouth, NM: Heinemann.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research, Third edition.* Los Angeles: Sage.

Creswell, J. W. Plano Clark, V. L., Gutmann, M. L, & Hanson, W. E. (2003). Advanced ixed-Methods Research Design. In A. Tashakkori & C. Teddle (Eds.) *Handbook of mixed methods in social & behavioral research.* Sage Publications: Thousand Oaks.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), p. 23-45.

Dusenbury, L.., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research: Theory and Practice, 18*(2): 237-256.

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

Fountas, I. C., & Pinnell, G. S. (2001). *Guiding readers and writers grades 3-6: Teaching comprehension, genre, and content literacy.* Westport, CT: Heinemann Education Books.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* New York: Aldine de Gruyter.

Harris, D. N. & Adams, S. J. (2010). Understanding the level and causes of teacher turnover: A comparison with other professions. *Economics of Education Review, 26* (3): 325-337.

Hassett, D. D. (2008). Teacher flexibility and judgment: A multidynamic literacy theory. *Journal of Early Childhood Literacy, 8*(3), p. 295-327.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1994). *The Iowa Tests: Interpretive guide for school administrators Forms K and L, Levels 5-14.* Itasca, IL: The Riverside Publishing Company.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K.R., Ordman, V. L., Naylor, R. J., et al. (2003). *Iowa Test of Basic Skills guide to research and development.* Itasca, IL: The Riverside Publishing Company.

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

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*(1), 88-110.

Ingersoll, R. (2001a). *Teacher turnover, teacher shortages, and the organization of schools*. Center for the Study of Teaching and Policy. University of Washington, Seattle.

Ingersoll, R. (2001b). Teacher turnover and teacher shortages: An organizational analysis. *American Education Research Journal, 38*(3): 499-534.

Lipsey, M. W., Puzio, K., Yun, C., Herbert, M. A., Steinka-Fry, K., Cole, M. W., Busick, M. D. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms.* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, US Department of Education.

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

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

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78: 33-84.*

Pinnell, G. S., Lyons, C. A., DeFord, D. E., Bryk, A., & Seltzer, N. (1994). Comparing Instructional Models for the Literacy Education of High Risk First Graders. *Reading Research Quarterly, 29, 8-39.*

Raudenbush, S. W., & Bryk, A. S. (2002) *Hierarchical linear modeling: Applications and data analysis methods (2nd edition).* Thousand Oaks, CA: Sage Publications.

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

Schön, D. A. (1991). *The reflective turn: Case studies in and on educational practice*. New York, NY: Teachers College Press.

Schwartz, R. M. (2005*).* Literacy Learning of At-Risk First-Grade Students in the Reading Recovery Early Intervention. *Journal of Educational Psychology*, 97(2), 257-267.

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

U.S. Department of Education Office of Inspector General. (2006). *The Reading First Program's Grant Application Process: Final Inspection Report* (Publication Number I13-F0017). Washington, D.C.

Weiss, M. J., Bloom, H. S., & Brock, T. (2013). *A conceptual framework for studying the sources of variation in program effects*. New York: MDRC.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity and effectiveness. *Journal of Consulting and Clinical Psychology, 49:* 156-167.

A COLLABORATIVE PUBLICATION BETWEEN

CPRE UNIVERSITY OF DELAWARE | Center for Research in Education & Social Policy