NCEE 2008-4016

U. S. DEPARTMENT OF EDUCATION

# Reading First Impact Study: Interim Report

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

# Reading First Impact Study: Interim Report

April 2008

**Beth C. Gamse**, Project Director, Abt Associates
**Howard S. Bloom**, MDRC
**James J. Kemple**, MDRC
**Robin Tepper Jacob**, Abt Associates/University of Michigan

**Beth Boulay**
**Laurie Bozzi**
**Linda Caswell**
**Megan Horst**
**W. Carter Smith**
**Robert G. St.Pierre**
**Fatih Unlu**
Abt Associates

**Corinne Herlihy**
**Pei Zhu**
MDRC

*With the assistance of*
**Diane Greene**
**Jongsun Kim**
**Don LaLiberty**
**Ken Lam**
**Kenyon Maree**
**Rachel McCormick**
**Jesselle Miura**
**Rebecca Unterman**
**Edmond Wong**

NCEE 2008-4016
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This report was prepared for the Institute of Education Sciences under Contract No. ED-01-CO-0093/0004. The project officer was Tracy Rimdzius in the National Center for Education Evaluation and Regional Assistance.



**April 2008**



**To order copies of this report,**

- Write to ED Pubs, Education Publications Center, U.S. Department of Education, P.O. Box 1398, Jessup, MD 20794-1398.
- Call in your request toll free to 1-877-4ED-Pubs. If 877 service is not yet available in your area, call 800-872-5327 (800-USA-LEARN). Those who use a telecommunications device for the deaf (TDD) or a teletypewriter (TTY) should call 800-437-0833.
- Fax your request to 301-470-1244.
- Order online at www.edpubs.org.

This report also is available on the IES website at http://ncee.ed.gov.

**Alternate Formats**
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

## Acknowledgements

The Reading First Impact Study Team would like to express its gratitude to the students, faculty, and staff in the study's participating schools and districts. Their contributions to the study (via assessments, observations, surveys, and more) are deeply appreciated. We are the beneficiaries of their generosity of time and spirit.

The listed authors of this report represent only a small part of the team involved in this project. We would like to acknowledge the support of staff from Computer Technology Services (for the study's data collection website), from DataStar (for data entry), from MDRC (especially Mario Flecha, for his help on the calendar front), from Retail Solutions at Work (and the hundreds of classroom observers who participated in intensive training and data collection activities), from Paladin Pictures (for developing training videos for classroom observations), from RMC Research (especially Chris Dwyer, for help on developing instruments and on training observers), from Rosenblum-Brigham Associates (for district site visits), from Westat (especially Sherry Sanborne and Alex Ratnofsky, for managing the student assessment, and the many Student Assessment Coordinators and even more test administrators), and from Westover (especially Wanda Camper, LaKisha Dyson, and Pamela Wallace for helping with meeting logistics).

The study has also benefited from both external and internal technical advisors, including:

**External Advisors**
Josh Angrist
David Card
Robert Brennan
Thomas Cook*
Jack Fletcher*
David Francis
Larry Hedges*
Robinson Hollister*
Guido Imbens
Brian Jacob
David Lee
Tim Shanahan*
Judy Singer
Jeff Smith
Faith Stevens*
Petra Todd
Wilbert Van der Klaauw
Sharon Vaughn*

**Internal advisors**
Steve Bell (A)
Gordon Berlin (M)
Nancy Burstein (A)
Fred Doolittle (M)
Barbara Goodson (A)
John Hutchins (M)
Marc Moss (A)
Chuck Michalopoulous (M)
Larry Orr (A)
Cris Price (A)
Janet Quint (M)
Howard Rolston (A)

(A—Abt Associates)
(M—MDRC)

* Individuals who have served on the study's Technical Work Group

Finally, we want to recognize the steady contributions of Abt staff, including Brenda Rodriguez, Fran Coffey, Lynn Reneau, Davyd Roskilly, Jon Schmalz, and Estella Sena, who were instrumental in completing multiple data collections, and Eileen Fahey, Katheleen Linton, and Jan Nicholson for countless hours of production support.

## Disclosure of Potential Conflicts of Interests[1]

The research team for this evaluation consists of a prime contractor, Abt Associates, and two major subcontractors, MDRC and Westat. None of these organizations or their key staff has financial interests that could be affected by findings from the Reading First Impact Study. No one on the Technical Work Group, convened to provide advice and guidance, has financial interests that could be affected by findings from the evaluation.

[1] Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the particular tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

# Contents

# List of Exhibits

# Executive Summary

This report presents preliminary findings from the Reading First Impact Study, a congressionally mandated evaluation of the federal government's $1.0 billion-per-year initiative to help all children read at or above grade level by the end of third grade. The No Child Left Behind Act of 2001 (P.L. 107-110) established Reading First (Title I, Part B, Subpart 1) and mandated its evaluation. This evaluation is being conducted by Abt Associates and MDRC with RMC Research, Rosenblum-Brigham Associates, Westat, Computer Technology Services, DataStar, Field Marketing Incorporated, and Westover Consulting under the oversight of the U.S. Department of Education, Institute of Education Sciences (IES).

The present report is the first of two; it examines the impact of Reading First funding in 2004-05 and 2005-06 in 17 school districts across 12 states and one statewide program (18 sites). The report examines program impacts on students' reading comprehension and teachers' use of scientifically based reading instruction. Key findings are that:

- On average, across the 18 participating sites, estimated impacts on student reading comprehension test scores were not statistically significant.
- On average, Reading First increased instructional time spent on the five essential components of reading instruction promoted by the program (phonemic awareness, phonics, vocabulary, fluency, and comprehension).
- Average impacts on reading comprehension and classroom instruction did not change systematically over time as sites gained experience with Reading First.
- Study sites that received their Reading First grants later in the federal funding process (between January and August 2004) experienced positive and statistically significant impacts both on the time first and second grade teachers spent on the five essential components of reading instruction and on first and second grade reading comprehension. Time spent on the five essential components was not assessed for third grade, and impacts on third grade reading comprehension were not statistically significant. In contrast, there were no statistically significant impacts on either time spent on the five components of reading instruction or on reading comprehension scores at any grade level among study sites that received their Reading First grants earlier in the federal funding process (between April and December 2003).

The study's final report, which is due early 2009, will provide an additional year of follow-up data, and will examine whether the magnitude of impacts on the use of scientifically based reading instruction is associated with improvements in reading comprehension.

## The Reading First Program

Reading First promotes instructional practices that have been validated by scientific research (No Child Left Behind Act, 2001). The legislation explicitly defines scientifically based reading research and outlines the specific activities state, district, and school grantees are to carry out based upon such research (No Child Left Behind Act, 2001). The Guidance for the Reading First Program provides further detail to states about the application of research-based approaches in reading (U.S. Department of Education, 2002). Reading First funding can be used for:

- *Reading curricula and materials* that focus on the five essential components of reading instruction as defined in the Reading First legislation: 1) phonemic awareness, 2) phonics, 3) vocabulary, 4) fluency, and 5) comprehension;
- *Professional development and coaching* for teachers on how to use scientifically based reading practices and how to work with struggling readers;
- *Diagnosis and prevention* of early reading difficulties through student screening, interventions for struggling readers, and monitoring of student progress.

Reading First grants were made to states between July 2002 and September 2003. By April 2007, states had awarded subgrants to 1,809 school districts, which had provided funds to 5,880 schools. Districts and schools with the greatest demonstrated need, in terms of student reading proficiency and poverty status, were intended to have the highest funding priority (U.S. Department of Education, 2002). In addition to grants for individual schools, states and districts could reserve up to 20 percent of their Reading First funds to support staff development and reading assessments, among other activities, for all high-need schools (U.S. Department of Education, 2002).

## The Reading First Impact Study

The Reading First Impact Study (RFIS) was commissioned to address the following questions:

1) What is the impact of Reading First on student reading achievement?
2) What is the impact of Reading First on classroom instruction?
3) What is the relationship between the degree of implementation of scientifically based reading instruction and student reading achievement?

The current report presents preliminary answers to the first two questions. The study's final report will address all three questions.

### Research Design

The Reading First Impact Study employs a regression discontinuity design that capitalizes on the systematic process used by a number of school districts to allocate their Reading First funds. A regression discontinuity design is the strongest quasi-experimental method that exists for estimating program impacts. Under certain conditions, outlined below, all of which are met by the present study, this method can produce unbiased estimates of program impacts:

1) Schools eligible for Reading First grants were rank-ordered for funding based on a quantitative rating, such as an indicator of past student reading performance or poverty.
2) A cut-point in the rank-ordered priority list separated schools that did or did not receive Reading First grants, and this cut-point was set without knowing which schools would then receive funding.
3) Funding decisions were based only on whether a school's rating was above or below its local cut-point; nothing superseded these decisions.
4) The shape of the relationship between schools' ratings and outcomes is correctly modeled.

Under these conditions, there should be no systematic differences between eligible schools that did and did not receive Reading First grants (Reading First and non-Reading First schools respectively), except for the characteristics associated with the school rating used to determine the funding decision. By controlling for differences in schools' ratings, one can then control statistically for all systematic pre-existing differences between the two groups. This makes it possible to estimate the impact of Reading First by comparing the outcomes for Reading First schools and non-Reading First schools in the study sample, controlling for differences in their ratings. Non-Reading First schools in a regression discontinuity analysis thereby play the same role as do control schools in a randomized experiment—they represent the best indications of what outcomes would have been for the treatment group (Reading First schools) in the absence of the program being evaluated.

**Study Sample**

Twenty-eight school districts plus one state Reading First program that met the preceding criteria were identified. Sixteen districts plus the state program were chosen from this pool to participate in the regression discontinuity design; the final selection reflected wide variation in district characteristics and provided enough schools to meet the study's sample size requirements. One other school district agreed to randomly assign some of its eligible schools to Reading First or a control group. The 17 school districts and one state Reading First program are referred to as study sites. The regression discontinuity sites provide 238 schools for the analysis and the randomized experimental site provides 10 schools. Half of these schools at each site are Reading First schools and half are non-Reading First schools; the study schools comprise some, not all, of the RF schools in study sites.

Exhibit ES.1 compares background characteristics of Reading First schools in the study sample to those of all Reading First schools in the 18 study sites, all Reading First schools in the 13 study states, and all Reading First schools in the nation. Visual inspection of the data displayed in this exhibit suggests that, overall, the present sample is similar to the other three groups of Reading First schools. Almost all are eligible for Title I support, they enroll high percentages of students eligible for free or reduced price lunch, and their past third grade reading scores are near their state averages for Reading First schools. The RFIS sample, on average, has proportionally lower percentages of Hispanic students and higher percentages of Black students than Reading First schools in the study states or in the nation; at the same time, RFIS sample schools, on average, have a lower percentage of Black students and a higher percentage of White students than Reading First schools in study districts. A greater proportion of Reading First schools in the study sample are in large or mid-size cities, and not other locales, than are Reading First schools in the study states or in the nation. Also, the sizes of Reading First schools in the study sample, on average, are somewhat smaller than those in the three other groups. Further, these data cannot provide conclusive evidence that the study sample fully represents the experience of the entire national Reading First program, as the study sample might differ from the Reading First population in other ways that were not observed.

**Exhibit ES.1: Baseline Characteristics of Relevant Groups of Reading First Schools for 2002-2003**

| Characteristic | RF Schools in Study Sample | RF Schools in Study Districts | RF Schools in Study States | RF Schools in U.S. |
|---|---|---|---|---|
| **Students** | | | | |
| Male (%) | 52.3 | 52.0 | 51.7 | 51.5 |
| Race (%) | | | | |
| Asian | 3.1 | 2.5 | 1.5 | 3.5 |
| Black | 35.6 | 41.1 | 26.4 | 30.5 |
| Hispanic | 26.7 | 28.6 | 37.1 | 34.8 |
| White | 34.2 | 27.4 | 34.3 | 28.6 |
| American Indian/Alaskan | 0.5 | 0.4 | 0.6 | 2.5 |
| Free Lunch and Reduced Lunch (%) | 74.4 | 75.0 | 67.8 | 73.2 |
| **Schools** | | | | |
| Eligible for Title 1(%) | 97.6 | 97.4 | 96.4 | 94.8 |
| Locale (%) | | | | |
| Large City | 39.2 | 39.8 | 26.7 | 26.8 |
| Mid-size City | 36.8 | 36.5 | 21.0 | 19.5 |
| Other[a] | 24.0 | 23.7 | 52.3 | 53.6 |
| Size | | | | |
| Total Number of Students | 474.8 | 487.4 | 502.4 | 531.4 |
| Number of Students in Grade 3 | 71.6 | 75.1 | 80.2 | 84.9 |
| Student/Teacher Ratio | 15.1 | 14.8 | 15.1 | 16.5 |
| **Third Grade Reading Performance** | | | | |
| Deviation from State RF Mean Proficiency Rate (%)[b] | -1.3 | -3.3 | 0.0 | 0.0 |
| **Number of Schools[c]** | 125 | 274 | 1,728 | 4,793 |

**Notes:**

The RF study sample includes 128 schools from 18 sites (17 districts and 1 state) located in 13 states. The RF schools in Study Districts include all RF schools ranked and/or rated on the RF grant application for each of the 18 sites in the study. All RF schools in Study States include all RF schools in the 13 states included in the study. All RF schools nationally include all schools that received RF grants.

[a] Other Locale includes urban fringe of a large city, urban fringe of a mid-sized city, large town, small town, and rural.

[b] A school's proficiency score is defined as the percentage of third grade students (or fourth or fifth grade when third grade is unavailable) in the school that score at or above the state-defined proficiency threshold on the state's reading assessment. The values in this row represent the average percentage point deviation from the mean proficiency score for the Reading First schools in the state. By definition, for a given state the mean proficiency score for all Reading First schools in the state is the benchmark for comparison. Therefore, in the final two columns, the deviation from the benchmark within each state is zero and the average deviation across states is zero.

[c] Due to missing values for some variables, the number of schools included varies by characteristic.

*Sources:* *Baseline characteristic data are from the Common Core of Data. RF school samples are defined based on information from the Southwest Educational Development Laboratory.*

### Data Collection and Outcome Measures

Exhibit ES.2 summarizes the study's three-year, multi-source data collection plan. The present report reflects data for 2004-05 and 2005-06. Key outcome measures include student reading comprehension, teacher reading instructional practices, and student engagement with print.

**Exhibit ES.2: Data Collection Schedule for the Reading First Impact Study**

| Data Collection Elements | 2004-2005 Fall | 2004-2005 Spring | 2005-2006 Fall | 2005-2006 Spring | 2006-2007 Fall | 2006-2007 Spring |
|---|---|---|---|---|---|---|
| Student Testing | ✓ | ✓ | | ✓ | | ✓ |
| Classroom Observations | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Teacher, Principal, Reading Coach Surveys | | ✓ | | | | ✓ |
| District Staff Interviews | | ✓ | | | | ✓ |

***Student reading comprehension*** was assessed with the Stanford Achievement Test, 10th Edition (SAT 10, Harcourt Assessment, Inc., 2004). Its comprehension subtests are well documented, broadly accepted, and widely used.[2] Test scores are analyzed in two forms: scaled scores and the percentage of students who read at or above grade level, based upon national SAT 10 norms. The SAT 10 was administered to students in grades one, two, and three during spring 2005 and spring 2006, with completion rates of 80 percent or higher for both waves.

***Classroom instruction*** was assessed in first grade and second grade reading classes through an observation system developed by the study team called the Instructional Practice in Reading Inventory (IPRI). Observations were conducted in each study school on two consecutive days in spring 2005, fall 2005, and spring 2006, with completion rates over 96 percent.

Measures of classroom instruction were created from IPRI data to represent the components of reading instruction emphasized by the Reading First legislation:[3]

- *Total daily minutes of instruction in all five dimensions:* This measure equals the total number of minutes of instruction in phonemic awareness, phonics, vocabulary, fluency, and comprehension during the daily reading block, which is the time period designated for reading instruction.
- *Minutes of instruction per day in each of the five dimensions:* These five measures correspond to the number of minutes of instruction in each of the five dimensions per daily reading block.
- *Percentage of three-minute observational intervals with instruction in the five dimensions that involve highly explicit instruction:* This measure records instances of "highly explicit instruction" that occur during instruction in any of the five dimensions. Highly explicit instruction means active teaching, modeling or explaining concepts, or helping children use reading strategies.

[2] In spring 2007, the study added the Test of Silent Word Reading Fluency (TOSWRF) for grade 1; findings based on this test will be presented in the final report.

[3] For ease of explication, the measures created from IPRI data are referred to as the five dimensions of reading instruction (or "the five dimensions") throughout the report. References to the programmatic emphases as required by legislation are labeled as the five essential components of reading instruction.

- *Percentage of three-minute observational intervals with instruction in the five dimensions that involve high quality student practice:* This measure records instances of "high quality student practice" that occur during instruction in any of the five dimensions. High quality student practice involves dimension-specific opportunities for students to practice their skills.

***Student engagement with print*** was assessed beginning in fall 2005 through classroom observations using the Student Time-on-Task and Engagement with Print (STEP) instrument to measure the percentage of students engaged in academic work who are reading or writing print. The STEP, which was developed by the study team, was used to observe classrooms in both fall 2005 and spring 2006, with a completion rate of over 97 percent.

## Average Impacts Across All Sites

Exhibit ES.3 reports average impacts for school years 2004-05 and 2005-06.[4] All impact estimates are regression-adjusted to control for a linear specification of the rating variable each site used to select its Reading First schools as well as selected teacher and/or student background characteristics used in the analysis. The impacts have been estimated using multi-level models to account for the clustering of students within classrooms, classrooms within schools, and schools within sites. In Exhibit ES.3, values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in Reading First schools absent Reading First funding and are calculated by subtracting the impact estimates from the Reading First schools' actual mean values. Impacts were estimated for each study site and averaged across sites in proportion to their number of Reading First schools in the sample. Average impacts thus represent the average study school. On average:

- **Reading First did not improve students' reading comprehension**. The program did not increase the percentages of students in grades one, two, or three, whose reading comprehension scores were at or above grade level. In each of the three grades, fewer than half of the students in the Reading First schools were reading at or above grade level.
- **Reading First increased total class time spent on the five essential components of reading instruction promoted by the program.** The program increased average class time spent on the five essential components of reading instruction by 8.56 minutes per daily reading block in grade one, and by 12.09 minutes per daily reading block in grade two. This implies a weekly increase of three quarters of an hour for grade one and one hour for grade two.
- **Reading First increased highly explicit instruction in grades one and two and increased high quality student practice in grade two**. The program increased the percentage of class observational intervals spent on the five dimensions of reading instruction that involve highly explicit instruction by 3.65 percentage points in grade one and by 6.98 percentage points in grade two. The program also increased the percentage of class observational intervals spent on the five dimensions of reading instruction that involve high quality student practice by 3.67 percentage points in grade two. There was virtually no observed change in grade one.

[4] Exhibit ES.3 and all other tables indicate whether findings are based on the full study sample or specific subgroups. Where appropriate, each exhibit also includes an "Exhibit Reads" section that walks readers through the exhibit by highlighting the first row or line of information presented.

**Exhibit ES.3: Estimated Impacts on Reading Comprehension, Instruction, and Percentage of Students Engaged with Print: Spring 2005, Fall 2005, and Spring 2006**

| | Actual Mean with Reading First | Estimated Mean without Reading First | Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|
| **Reading Comprehension** | | | | |
| *Percent Reading At or Above Grade Level* | | | | |
| Grade 1 | 45.4 | 42.2 | 3.15 | (0.260) |
| Grade 2 | 38.9 | 38.8 | 0.12 | (0.965) |
| Grade 3 | 37.9 | 40.1 | -2.22 | (0.383) |
| **Instruction** | | | | |
| *Number of minutes of instruction in the five dimensions combined* | | | | |
| Grade 1 | 59.41 | 50.85 | 8.56* | (0.003) |
| Grade 2 | 59.53 | 47.44 | 12.09* | (<0.001) |
| *Percentage of intervals in five dimensions with Highly Explicit Instruction* | | | | |
| Grade 1 | 29.78 | 26.13 | 3.65* | (0.023) |
| Grade 2 | 31.55 | 24.57 | 6.98* | (<0.001) |
| *High Quality Student Practice* | | | | |
| Grade 1 | 19.21 | 18.35 | 0.86 | (0.559) |
| Grade 2 | 18.78 | 15.11 | 3.67* | (0.012) |
| **Percentage of Students Engaged with Print** | | | | |
| Grade 1 | 46.92 | 42.29 | 4.63 | (0.216) |
| Grade 2 | 49.72 | 58.14 | -8.42* | (0.030) |

**Notes:**

The complete Reading First Impact Study (RFIS) sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

**EXHIBIT READS: The observed average percent of first-graders reading at or above grade level with Reading First was 45.4 percentage points. The estimated average percent without Reading First was 42.2 percentage points. The impact of Reading First on the percent of first grade students reading at or above grade level was 3.2 percentage points, which was not statistically significant at the p<.05 level (p=.260).**

*Sources: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already used the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006; RFIS Student Time-on-Task and Engagement with Print, fall 2005 and spring 2006.*

- **Reading First had mixed effects on student engagement with print**. The program reduced the percentage of students engaged with print by a statistically significant 8.42 percentage points in grade two. The impact on student engagement with print in grade one (4.63 percentage points) was not statistically significant.

## Impact Differences

Study sites differ from each other in ways that could potentially influence the effectiveness of Reading First. For example, sites differ in terms of the length of time since date of Reading First grant award, levels of Reading First funding per student, and prior levels of reading performance. Consequently, average impacts for the full study sample might mask important differences that exist over time and/or across sites. The study explored this possibility by examining the pattern of impacts over time for two groups of study sites. The first group consists of the eight "late award" sites that received Reading First grants between January and August 2004. As of May 2006, these sites had been receiving Reading First funds for an average of approximately two years. The second group consists of the 10 "early award" sites that received Reading First grants between April and December 2003. As of May 2006, these sites had been receiving Reading First funds for an average of approximately three years, although data from the study are available only for the last two years. Study findings indicate that:

- **The impacts of Reading First on classroom instruction and student reading comprehension have not changed consistently over time**. Exhibit ES.4 shows estimated impacts for the two years that data are available for late award and early award sites, respectively. For both groups of sites, estimates of program impacts on reading comprehension and classroom instruction vary from year to year (across columns). However, this variation exhibits no consistent pattern and is not statistically significant. These findings do not suggest that program impacts increased or decreased with program maturity.
- **The estimated impacts of Reading First were consistently positive for late award sites and mixed for early award sites.** Exhibit ES.5 presents estimated impacts for the two groups of sites that are averaged over the two years for which data are available. It indicates that, for grades one and two in late award sites, Reading First produced positive and statistically significant increases both in teachers' instruction in the five dimensions and in students' reading comprehension. Impacts on third grade reading comprehension were not statistically significant for late award sites, though the direction of the (not significant) estimated impact was positive. None of the impact estimates presented in Exhibit ES.5 are statistically significant for early award sites. The (not significant) estimated impacts on teachers' instruction were positive, and the (not significant) estimated impacts on student reading comprehension were negative. Differences in impacts on reading comprehension test scores between early and late award sites were statistically significant for grades two and three, and not statistically significant for grade one. Differences in impacts on instruction in the five dimensions between early and late award sites were not statistically significant.
- **It is not possible to determine which of numerous differences between early award sites and late award sites may have caused observed differences in Reading First impacts, only some of which were statistically significant.** The average per K-3 student Reading First funding was higher in late award sites than early award sites ($574 versus $432 per student). Although the study did not begin to collect data until after early award sites began to implement Reading First, it appears that the benchmarks of comparison for student reading comprehension were lower for late award sites. Thus, late award sites may have had more room to increase reading comprehension skills. Any or all of these differences, plus others not measured, could have produced the impact differences observed.

**Exhibit ES.4: Estimated Impacts on Reading Comprehension and Minutes in the Five Dimensions, by Implementation Year, Calendar Year, and Award Status**

| | Implementation Year | | | | | |
|---|---|---|---|---|---|---|
| | Year 1 | | Year 2 | | Year 3 | |
| | Impact | (p-value) | Impact | (p-value) | Impact | (p-value) |
| ***Panel 1*** | | | | | | |
| **Late Award Sites** | **2005** | | **2006** | | **2007** | |
| Grade 1 | | | | | | |
| Percent reading at or above grade level (%) | 6.3 | (0.077) | 9.4* | (0.024) | N/A | N/A |
| Instruction in five dimensions (minutes) | 11.51* | (0.001) | 12.03* | (0.004) | N/A | N/A |
| Grade 2 | | | | | | |
| Percent reading at or above grade level (%) | 6.3* | (0.028) | 5.7 | (0.155) | N/A | N/A |
| Instruction in five dimensions (minutes) | 14.84* | (<0.001) | 16.11* | (<0.001) | N/A | N/A |
| Grade 3 | | | | | | |
| Percent reading at or above grade level (%) | 1.7 | (0.537) | 4.2 | (0.269) | N/A | N/A |
| Instruction in five dimensions (minutes) | N/A | N/A | N/A | N/A | N/A | N/A |
| ***Panel 2*** | | | | | | |
| **Early Award Sites** | **2004** | | **2005** | | **2006** | |
| Grade 1 | | | | | | |
| Percent reading at or above grade level (%) | N/A | N/A | -2.6 | (0.708) | -1.9 | (0.751) |
| Instruction in five dimensions (minutes) | N/A | N/A | 5.49 | (0.376) | 4.16 | (0.457) |
| Grade 2 | | | | | | |
| Percent reading at or above grade level (%) | N/A | N/A | -8.2 | (0.163) | -6.8 | (0.303) |
| Instruction in five dimensions (minutes) | N/A | N/A | 10.93 | (0.083) | 4.56 | (0.410) |
| Grade 3 | | | | | | |
| Percent reading at or above grade level (%) | N/A | N/A | -9.9 | (0.110) | -7.7 | (0.225) |
| Instruction in five dimensions (minutes) | N/A | N/A | N/A | N/A | N/A | N/A |

**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

Implementation year represents the number of years since sites received notice of their Reading First grants. For early award sites, this occurred in 2003, and Years 1, 2, and 3 refer to the 2003-2004, 2004-2005, and 2005-2006 school years, respectively. For late award sites, notification of funding occurred in 2004, and Years 1 and 2 refer to the 2004-2005 and 2005-2006 school years, respectively (data are available for the 2004-2005 and 2005-2006 school years only).

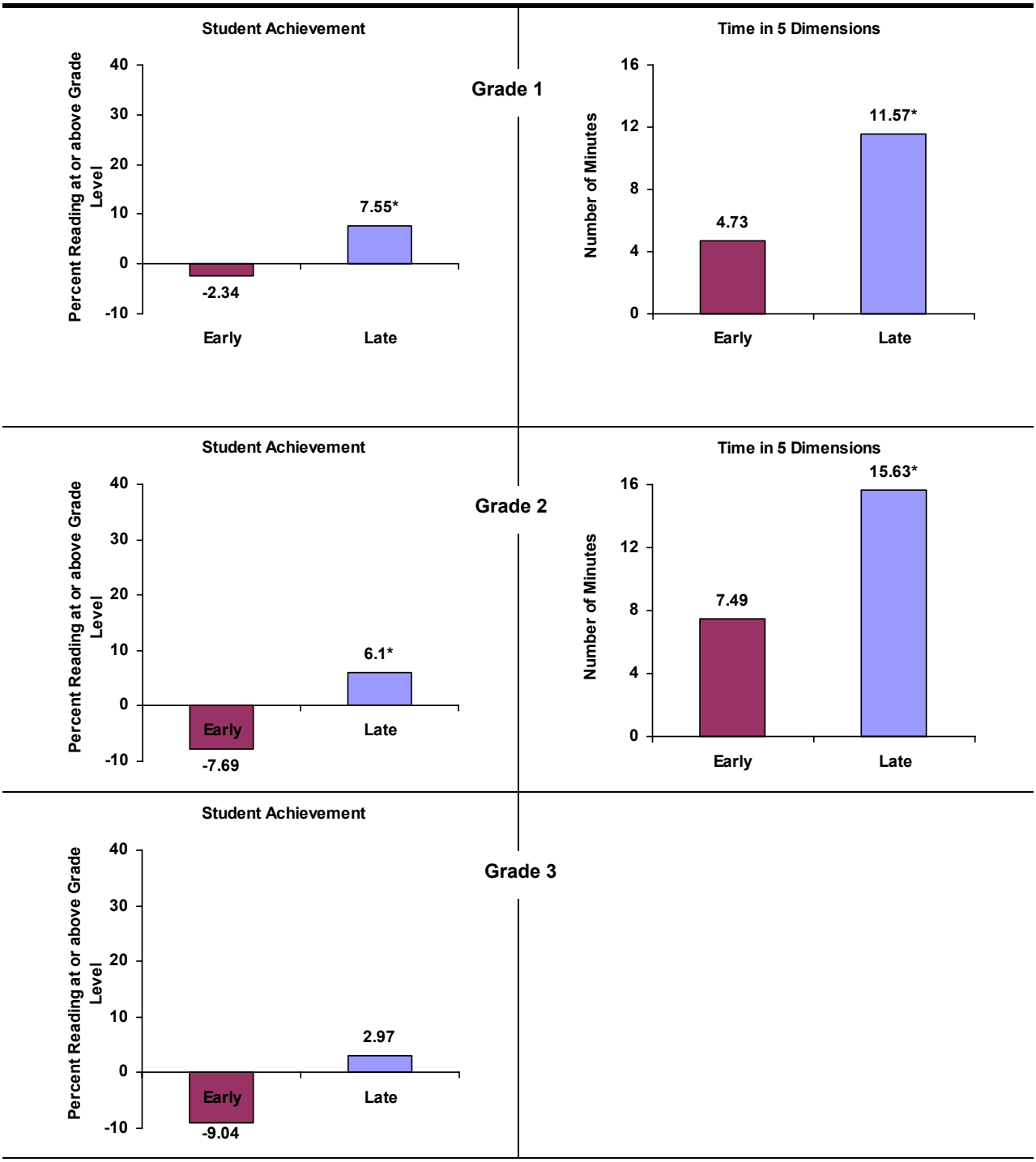
Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

**EXHIBIT READS: The impact of Reading First on the percent of students reading at or above grade level in grade one, for late award sites, in implementation Year 1 and Calendar Year 2005, was 6.3 percentage points, which was not statistically significant at the $p \leq .05$ level (p=.077).**

*Sources: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); and RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006.*

## Exhibit ES.5: Estimated Impacts on Key Outcomes for Early and Late Award Sites, by Grade




**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. There are 8 late award sites, with 137 schools, and 10 early award sites, with 111 schools.

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

**EXHIBIT READS: For grade one, the impact of Reading First on the percent of students reading at or above grade level was 7.55 percentage points for late award sites, which was statistically significant ($p \leq .05$). The corresponding impact for grade one in early award sites was –2.34 percentage points, which was not statistically significant.**

*Sources: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); and RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006.*

## Further Research

Data for the study's final report will include three years of follow-up on students' reading comprehension for grades one, two and three and three years of follow-up on teachers' classroom instruction for grades one and two. These data will enable the study to examine program impacts on comprehension and instruction for an additional school year and on one year of follow-up on first grade students' decoding skills. Finally, the study's final report will explore whether the observed Reading First impacts on instructional practices are associated with observed impacts on student reading comprehension.

# Chapter One: Study Overview

The No Child Left Behind Act of 2001 (NCLB) established the Reading First Program, a major federal initiative designed to help ensure that all children can read at or above grade level by the end of third grade. The RF legislation requires the U.S. Department of Education to contract with an outside entity to evaluate the impact of the Reading First Program. To meet this requirement, the Department contracted with Abt Associates in September 2003 to design and conduct the Reading First Impact Study (RFIS). The partner organizations included MDRC, RMC Research, Rosenblum-Brigham Associates, and Westat.[5] The RFIS is a multi-year study that encompasses data collection over the course of three school years: 2004-05, 2005-06, and 2006-07.

This interim report presents major findings based on data collected during the 2004-05 and 2005-06 school years. This chapter begins with an overview of the Reading First Program, briefly describes the conceptual framework underlying the program and this evaluation as a whole, and then outlines the study's guiding evaluation questions and data collection activities.

## Overview of Reading First Program

The No Child Left Behind Act (P.L. 107-110), signed into law in January 2002, established the Reading First Program (Title I, Part B, Subpart 1). The Reading First legislation requires programs and instruction to be based on scientific research in reading, and aims to ensure that all children can read at or above grade level by the end of third grade, thereby significantly reducing the number of students who experience difficulties in later years. The overarching goal of Reading First is to improve students' reading achievement. The program targets low-income, low-performing schools whose districts and states prepared articulated plans for increasing the use of teachers' research-based instruction through intensive professional development for teachers, reading coaches, and administrators, with the explicit aim of reaching out to all eligible schools over time (No Child Left Behind Act, 2001).

To qualify for Reading First funding, state and district professional development plans must include training on reading instructional methods and materials that incorporate the five essential components of reading instruction (phonemic awareness, phonics, vocabulary development, reading fluency, and reading comprehension strategies), and on the use of assessments that effectively screen, diagnose, and monitor student progress in reading (No Child Left Behind Act, 2001).

The Reading First legislation outlines the general components and activities to be included in state and local plans, and the Reading First Guidance describes several strategies that states and local educational agencies should use to improve students' reading skills (No Child Left Behind Act, 2001; U.S. Department of Education, 2002). First, the guidance specifies that *curricula* used in classrooms must reflect scientifically based reading research that includes the essential components of reading instruction, and further, that students should have sufficient opportunity to practice the development of their skills in these essential components. Second, it addresses teacher *professional development* on the implementation of scientifically based reading practices; states must offer comprehensive professional development on how teachers should work with academically struggling students, as

[5] Other subcontractor organizations included: Computer Technology Services, Inc.; DataStar, Inc.; Field Marketing Inc.; Paladin Pictures, Inc.; and Westover Consultants, Inc.

well as how teachers can implement research-based reading instruction. Third, state and local plans must include procedures for *diagnosis and prevention* of early reading difficulties through a) using valid, reliable measures to screen students; b) using empirically validated intensive interventions to help struggling students; and c) monitoring the progress of students experiencing difficulties to ensure that the early interventions are indeed effective.

Reading First is an ambitious federal program, yet it is also a funding stream that combines local flexibility and national commonalities. The commonalities are reflected in the guidelines to states and districts and schools about allowable uses of resources. The flexibility is reflected in two ways: one, states (and districts) could allocate resources to various categories within target ranges rather than on a strictly formulaic basis. Two, states could make local decisions about the specific choices within given categories (e.g., which materials, reading programs, assessments, professional development providers, etc.). The activities, programs, and resources that were likely to be implemented across states and districts would therefore reflect both national priorities and local interpretations.

All states received RF grants after their applications were subjected to an expert review process, and all states received funds for a six-year period. States then awarded sub-grants to local school districts and/or directly to schools based on a competitive process. As of April 2007, all states, territories, and the District of Columbia reported that over 5,880 sub-grants had been awarded to schools in over 1,809 school districts (Southwest Educational Development Laboratory, 2007).

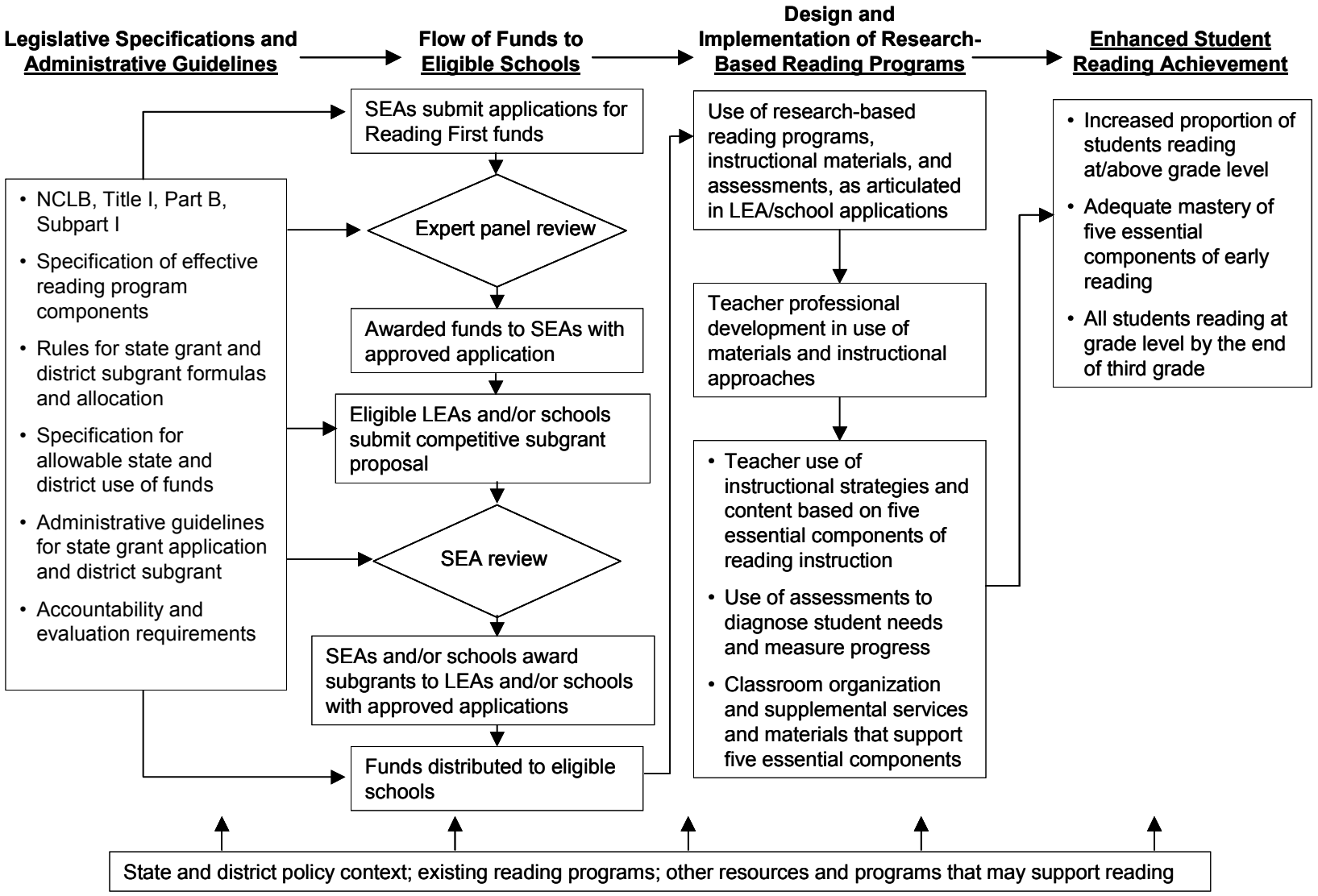
## A Conceptual Framework for the Reading First Impact Study

To understand the implementation and desired effects of Reading First, the conceptual framework presented below identifies the program's central goals and specifies the pathways through which its principles and components are hypothesized to improve reading instruction, and subsequently student reading achievement. The conceptual framework provides a substantive backdrop for the Reading First Impact Study.

Exhibit 1.1 shows the pathways through which Reading First is hypothesized to influence reading achievement: (1) the Reading First legislation provides programmatic specifications and administrative guidelines; (2) Reading First funds flow to states, districts, and ultimately to eligible schools; (3) districts and schools design and implement research-based reading programs and provide school personnel with training on research-based instructional strategies; and (4) student reading achievement is enhanced. Each of these steps is influenced by contextual variables, especially state and district funding for other reading programs.[6] The general focus of the Reading First Impact Study is on elements within the third and, ultimately, the fourth steps specified above (columns 3 and 4 in Exhibit 1.1). Each column is described below.

[6] Schools and districts could have sought and obtained other (non-RF) funding to support reading-related programs and instruction.

**Exhibit 1.1: Conceptual Framework for the Reading First Program: From Legislation and Funding to Program Implementation and Impact**



### Legislative Specifications and Administrative Guidelines

The first column of Exhibit 1.1 shows Reading First's major legislative specifications and administrative guidelines (No Child Left Behind Act, 2001). The Reading First legislation defines five essential components of reading instruction: (1) phonemic awareness; (2) phonics; (3) vocabulary development; (4) reading fluency, including oral reading skills; and (5) reading comprehension strategies (No Child Left Behind Act, 2001). The legislation also specifies state and district grant formulas, based primarily upon the proportion or number of children from low-income families who are reading below grade level in K–3, reflecting each district's percentage of the state's total Title I, Part A funds (No Child Left Behind Act, 2001). Sub-grants to eligible districts and schools must be of sufficient size and scope to enable full implementation of the selected research-based reading programs. Consequently, as indicated by states' Reading First applications and subsequent subgrant announcements, states did not fund all eligible entities, in order to concentrate resources and maximize the quality of implementation.[7]

The Reading First legislation and guidance indicate that states must allocate at least 80 percent of their funding to school districts, with the remainder allocated to state-level activities, including: (1) teacher professional development (not more than 13 percent of the state grant); (2) technical assistance for districts and schools (not more than five percent of the state grant); and (3) planning, administration and reporting (not more than two percent of the state grant). It is important to note that the residual funds (up to 20 percent) were to be used by states to disseminate Reading First-like information and resources to all schools (including those not awarded RF grants), in order to broaden the potential reach of the program beyond the RF-funded districts and schools awarded sub-grants (No Child Left Behind Act, 2001).[8] Local districts could spend up to 3.5 percent of their grants on administrative and technical assistance (U.S. Department of Education, 2002).

### The Flow of Reading First Funds

The second column of Exhibit 1.1 shows that RF funds flow from the federal government through the states to eligible districts and schools, as specified in the Reading First legislation (No Child Left Behind Act, 2001). First, the U.S. Department of Education convened expert panels to evaluate the State Education Agency (SEA) applications and make recommendations to the Department. Second, state departments of education scrutinized Local Education Agency (LEA) and/or school applications to determine which LEAs and/or schools were most likely to be able to meet the state's goals and specifications for Reading First.[9]

---

[7] For examples of state applications, see "Making Reading First in Michigan," (Michigan Department of Education, 2002, p. 64, 68) and "The State of Wisconsin Reading First Grant Proposal" (Wisconsin Department of Education, 2003, p. 47). For a list of award announcements, see "Reading First: Awards" (Southwest Educational Development Laboratory, 2007).

[8] The study did not collect data on other funding sources districts or schools obtained to support reading instruction.

[9] In some states, subgrants were made directly to schools (e.g., Hawaii, Kentucky).

### Design and Implementation of Research-Based Reading Programs

The activities listed in the third column of Exhibit 1.1 represent short-term or mediating outcomes for the Reading First program as well as the hypothesized precursors to the longer-term outcomes identified in the fourth column. Implementing research-based reading programs includes the following: use of reading programs deemed effective through scientifically based reading research; aligned materials and assessments for diagnosing student needs and measuring progress; well-designed professional development activities that train teachers explicitly in the essential components of reading instruction; strategies for adapting these practices to the varying skill levels of their students; and appropriate use of materials and assessments that support the chosen reading program (No Child Left Behind Act, 2001).

According to the Reading First guidelines, a well-implemented, high quality reading program sets high expectations for reading achievement and includes explicit strategies for monitoring student progress (U.S. Department of Education, 2002). Effective classroom reading instruction should also include differentiated small group instruction with flexible placement and movement based on ongoing assessment. Teachers should be using effective classroom management strategies to maximize time on reading-based tasks and activities. Most importantly, teachers and students should be continuously engaged in activities related to the five essential components of reading instruction.

### Enhanced Student Reading Achievement

The final column of Exhibit 1.1 identifies longer-term Reading First outcomes, all of which are focused on student reading achievement, including increased proportion of students reading at/above grade level in grades 1, 2, and 3; adequate mastery of the five essential components; and all students reading at or above grade level by the end of the third grade. The hypothesis underlying Reading First is that these outcomes will be achieved only through successful implementation of appropriate research-based reading programs, teacher professional development, use of diagnostic assessments, and appropriate classroom organization and provision of supplemental services.

## Reading First Impact Study Evaluation Questions

There are three major evaluation questions for the Reading First Impact Study:

1. **What is the impact of Reading First on student reading achievement?**
2. **What is the impact of Reading First on classroom instruction?**
3. **What is the relationship between the degree of implementation of scientifically based reading instruction and student reading achievement?**

The question about **impact on student reading achievement** focuses on the some of the elements represented in the final column of Exhibit 1.1. The Reading First Impact Study focuses primarily on student reading comprehension skills, by comparing student reading performance in Reading First schools to students' reading performance that would have been observed without Reading First funding. Students in the study schools are assessed with the Stanford Achievement Test, reading comprehension subtest, 10th Edition (Harcourt Assessment, Inc., 2004).

The question about **impact on classroom instruction** focuses primarily on the elements represented in the third box of the third column of Exhibit 1.1. Impacts on classroom instruction are assessed by

comparing characteristics of classroom instruction in Reading First schools to estimates of what those same characteristics of classroom instruction would have been had the schools not received Reading First funding.

The third question, about **relationships between implementation and students' reading achievement**, focuses on the connections between elements represented in the third and fourth columns of Exhibit 1.1. Results of analyses addressing these relationships will be presented in the final report.

The evaluation design (described in more detail in Chapter 2) calls for three years of data collection. This report presents findings based upon two years of data collection. While there is no prior research on the amount of time necessary for schools to have fully implemented the Reading First program, prior research on implementation of programs designed to improve student achievement through changing teachers' instructional practices suggests that while changes in instruction may be evident sooner, changes in student achievement can take several years to appear (e.g., Aladjem et al., 2006; Bloom, 2001; Borman et al., 2003). This holds particular salience for the Reading First program, which attempts to promote a comprehensive approach to reading instruction that persists from kindergarten through grade three. Some aspects of Reading First may be easy to implement quickly (i.e., purchase of new core reading programs and assessments, providing research-based professional development). Yet other aspects may require several years to implement effectively and consistently across the entire K-3 grade span (i.e., aligning curricula, instructional practices, and support services with the underlying principles of Reading First) to yield sustained improvement in student reading performance. Further, it will take four years of implementation before any students will have been able to experience Reading First funded activities as they progress from kindergarten through third grade.

The next chapter presents a discussion of the study design, estimation methods, and sample.

# Chapter Two: Study Design, Methods, and Sample

This chapter describes the study design and sample. It begins with a description of regression discontinuity design, a type of quasi-experimental study design that lends itself to a study of Reading First, in particular. The discussion of the regression discontinuity design (RDD) outlines the criteria that must be met to use this design, the requirements of sample size, and the outcome measures to be used, and it also presents a brief description of the estimation models and other key technical features of the analytic approach. The chapter then describes the study's sample of schools.

## Study Design

### Approach

The Reading First Impact Study is based on a regression discontinuity design that capitalizes on the systematic process used by a number of school districts to allocate their Reading First funds.[10] A regression discontinuity design is the strongest quasi-experimental method that exists for estimating program impacts. Under certain conditions (which are met by the present study) this method can approach the rigor of a randomized experiment.[11] The conditions include:

1) Eligible schools were rank-ordered for funding based on a quantitative rating, such as an indicator of past student reading performance or poverty.
2) A cut-point in the rank-ordered priority list separated schools that did or did not receive Reading First grants, and this cut-point was set without knowing which schools would then receive funding.
3) Funding decisions were based only on whether a school's rating was above or below its local cut-point; nothing superseded these decisions.
4) The shape of the relationship between schools' ratings and outcomes is correctly modeled.

To see how the method works, consider a hypothetical school district that allocates its $2 million annual Reading First grant to 10 schools in equivalent allotments of $200,000, per year, per school. The district also has prioritized the schools with the highest rates of poverty, as measured by the percentage of students eligible for free or reduced priced meals. The district therefore awards grants first to the school with the highest poverty rate, then to the school with the next-highest poverty rate, and so on, until ten schools receive grants and all of the Reading First funding has been allocated.

Exhibit 2.1 illustrates how the dividing line, or "cut-point," between the last funded school and the first school *not funded* on the district's priority list (or between the 10th and 11th schools on this

[10] The Reading First Impact Study was originally planned as a randomized control study, in which eligible schools from a sample of districts were to receive Reading First funds or become members of a non-Reading First control group. The approach was not feasible, however, in the 38 states that had already begun to allocate their Reading First grants before the study began. Furthermore, in the remaining states, randomization was counter to the spirit of the Reading First Program, which strongly emphasizes serving the schools most in need. It was possible, however, to randomize schools in one site.

[11] Regression discontinuity analysis was introduced by Thistlethwaite and Campbell (1960) and has more recently experienced a resurgence of interest (e.g., Cappelleri et al., 1991; Goldberger, 1972; Hahn, Todd and Van Der Klaauw, 2001; Mohr, 1995; and Reichardt, Trochim, and Cappelleri, 1995).

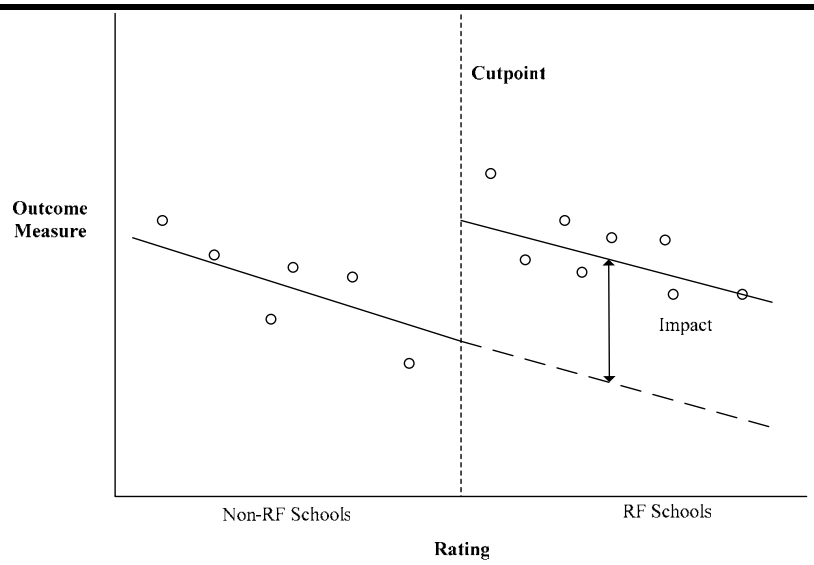
hypothetical district's list) creates a "discontinuity" that makes it possible to estimate program impacts on future outcomes. The vertical axis of the exhibit represents a future outcome measure for each school, such as its average student reading score in a subsequent year. The horizontal axis represents the rating used to determine each school's priority for Reading First (in this example, the percentage of past students eligible for free or reduced price meals). Schools to the left of the cut-point do not receive Reading First funding and serve as a "comparison group" for the impact analysis; these schools are referred to as non-Reading First schools. Schools to the right of the cut-point receive Reading First funding; these schools represent the "treatment group" for the impact analysis, and are referred to as Reading First schools.

The exhibit illustrates a downward-sloping relationship between schools' ratings and their future outcomes. This implies that schools with a higher proportion of past (and thus future) students who live in poverty will tend to have lower levels of future student achievement. In the absence of Reading First, average student achievement at non-Reading First schools would therefore tend to be higher than at Reading First schools. Consequently, the average outcome for non-Reading First schools most likely over-states what this average would have been for Reading First schools without the program (their "counterfactual"). Because of this, a simple comparison of average outcomes for Reading First schools and non-Reading First schools would understate the impact of Reading First.

Given the way that schools were selected for Reading First, however, it is possible to obtain unbiased estimates of the program's impacts on future outcomes by controlling statistically for the relationships that exist between school outcomes and ratings. (These relationships comprise the "regression" part of regression discontinuity analysis.) Intuitively, this analysis would proceed as follows. The first step is to fit a regression line through the data points for non-Reading First schools, as indicated by the solid line to the left of the cut-point in Exhibit 2.1. The second step is to extrapolate the fitted line across the cut-point to predict what student achievement would have been for Reading First schools—in the absence of the program. This is indicated by the dashed line in the exhibit. The third step is to fit a regression line through the data points for Reading First schools, as indicated by the solid line to the right of the cut-point. (For the purpose of this hypothetical example, the two fitted lines are assumed to have the same slope and are thus parallel, which simplifies the analysis but is not necessary.) The impact of Reading First thus can be measured by the vertical distance between the solid fitted line for Reading First schools (what actually happened in Reading First schools after the program was launched) and the dashed extrapolated line for Reading First schools (the counterfactual prediction of what would have happened in Reading First schools without the program). This distance is indicated by a two-sided arrow.

In short, the analysis uses the observable discontinuity in the regression relationship to identify the impact of Reading First. The magnitude of the discontinuity indicates the magnitude of the impact. If the regression model has the correct shape for the data being modeled (for example, two parallel straight lines for Reading First and non-Reading First schools), the discontinuity provides an unbiased impact estimate.

**Exhibit 2.1: Regression Discontinuity Analysis for a Hypothetical School District**




The approach works properly, if schools' ratings are the only thing that determines their selection for Reading First. Consequently, only background characteristics that are correlated with ratings can be correlated with selection for the program. In other words, the only characteristics that can differ systematically between Reading First schools and non-Reading First schools are those correlated with their ratings. Controlling statistically for the ratings thereby controls for any systematic pre-existing differences between the two groups of schools.[12] It is this control that makes unbiased impact estimates possible, yet it (regression discontinuity design) requires a much larger sample size than a randomized control trial to provide the same precision, because one must include the rating variable in any models to account for the design effect (Bloom, Kemple and Gamse, 2004).

Seventeen of the 18 sites in the Reading First Impact Study (16 school districts and one state program) allocated their Reading First grants in ways that meet the requirements of a regression discontinuity design (see Appendix B for a more detailed discussion). Each site prioritized its eligible schools according to a specified quantitative indicator, in most cases, an indicator based on a measure of student poverty, student performance, or both.[13] Each site then allocated its Reading First funds according to the prioritized list, funding the top priority school first, the second priority school next, and so on through the list, until all available resources were allocated. In the context of this study, these sites are referred to as regression discontinuity design (RDD) sites.

As explained later in this chapter in the section entitled "The Study Sample", the study sample was drawn from Reading First schools and non-Reading First schools whose ratings were as close as possible to their sites' local cut-point. Half of the schools in the study sample are Reading First

[12] It is because regression discontinuity analysis utilizes "selection on observables" (i.e., values of the rating) that it can produce unbiased impact estimates (Cain, 1975). This feature is what distinguishes the approach from other quasi-experimental designs.

[13] Exhibit 2.5 reports the criteria used by each site to rate its schools for Reading First. A separate rating coefficient (in the impact estimation model) was specified for each site to account for differences in rating variables and cut-points. These differences enhance the generalizability of the present study because it comprises 17 regression discontinuity analyses from different parts of the United States.

schools and half are non-Reading First schools.[14] Only 9 of the 248 sample schools from study sites had their rating-based Reading First funding status changed. Consequently, the study's sites support what is called a "sharp" regression discontinuity analysis, which is the strongest form of the design.[15]

In the 18th study site (a school district), it was possible to randomly assign a subset of its Reading First-eligible schools to receive or not receive Reading First funds. In this site, five candidate schools were assigned to Reading First and five were assigned to a control group. Hence, this site provides a group-randomized experiment. This site is referred to as the experimental site.

### Measures

The Reading First Impact Study focuses on three categories of outcome measures: student reading comprehension, classroom reading instruction, and student engagement with print during reading instruction. These three categories represent the outcome *domains* for the study. The outcome for student reading comprehension is represented by scores on the Stanford Achievement Test, 10th Edition (Harcourt Assessment, Inc., 2004). Classroom reading instruction and student engagement with print were measured through classroom observations made by trained observers. The outcome measures for instruction are represented by amount of instructional time on the five essential components of reading instruction, and the outcome for student engagement with print is the average percentage of students engaged with print during the reading block. Chapter Three describes what these measures mean and how they were obtained.

### Estimation

For each measure from the preceding outcome domains, an extension of the statistical model in Equation 1 was used to estimate the impacts of Reading First in the 17 RDD sites.[16] This equation is referred to as a linear regression discontinuity model.

$$Y_k = \beta_0 + \beta_1 T_k + \beta_2 R_k + \mu_k \qquad (1)$$

where:

- $Y_k$ = the outcome measure for school k,
- $T_k$ = one if school k is a Reading First school and zero otherwise,
- $R_k$ = the value of the rating for school k,
- $\mu_k$ = a random error for school k that is assumed to be independently and identically distributed.

---

[14] These proportions were exact for the original study sample of 258 schools. With the subsequent loss of 10 schools, they remain almost exact.

[15] A sharp regression discontinuity analysis has very few cases where assignment to treatment or comparison status based on ratings is changed due to other considerations. A "fuzzy" regression discontinuity design has more such aberrant cases. A fuzzy regression discontinuity analysis is more complex and requires further assumptions (Shadish, Cook and Campbell, 2002).

[16] Full statistical models for estimating impacts on all study outcomes for all 17 RDD sites are presented in Appendix B. The models include an indicator for the site where schools were randomized (for which impacts were estimated using a standard regression-adjusted difference of mean outcomes for the treatment group and control group).

The coefficient, $B_2$, for the rating, $R_k$, represents the slope of the two fitted regression lines in Exhibit 2.1. This summarizes the continuous relationship between outcomes and ratings that exists on either side of the cut-point. As noted, controlling for this relationship controls for all systematic pre-existing differences between Reading First schools and non-Reading First schools. The coefficient, $B_1$, for the treatment indicator, $T_k$, represents the discontinuity in the regression line produced by Reading First. The estimated value of $B_1$ therefore provides an estimate of the impact of Reading First.

The Reading First Impact Study is composed of separate regression discontinuity designs for each of the 17 RDD sites, plus a group-randomized experiment for the experimental site; as a result, the impact estimates presented are averaged across the study's 18 sites. The average is weighted in proportion to the number of Reading First schools in the study sample from each site. Findings presented in this report therefore represent average impacts for the average Reading First school in the sample.

To increase the precision of impact estimates a limited number of covariates (student background characteristics, teacher background characteristics, and/or school baseline test scores) were added to the estimation model. In addition, because students are clustered within classrooms, and classrooms are clustered within schools, multi-level models were used to estimate impacts on student outcomes. Appendix B describes the statistical models used to estimate impacts for outcomes in each of the study's three domains.

***Specification Tests***

As noted earlier, in developing the study sample, Reading First schools and non-Reading First schools were selected to be as close as possible to their local cut-points for receipt of Reading First funding. This was done to yield two groups of schools that were as similar as possible.[17] In addition, program impacts were estimated using a linear regression discontinuity model that controls for values of the ratings used to choose schools for program funding. Furthermore, as discussed earlier, estimates of impacts on measures of student reading comprehension control explicitly for school-level baseline measures of reading achievement. This *combination* of sample design and statistical analysis was expected to provide internally valid estimates of program impacts.

Three sets of specification tests were conducted to assess whether this expectation was met.[18] Although none of these tests by itself can *prove* that internal validity was achieved, in combination they provide evidence that this is most likely the case. The most important such test used a linear regression discontinuity model (as represented in Equation 1) to compare baseline characteristics of Reading First schools and non-Reading First schools. If a linear regression discontinuity model is an appropriate way to control for all pre-existing differences between the two groups, observable or not, then it should eliminate their observed baseline differences.

---

[17] See Appendix B, Part 2, Exhibit B.1 for unadjusted baseline characteristics of schools in the study sample.

[18] See Appendix B for a detailed presentation of the specification tests conducted to assess the study's internal validity.

Results of the baseline specification tests are presented in Exhibit 2.2. These findings were obtained using aggregate school-level baseline characteristics.[19] The first column presents adjusted residual differences between Reading First schools and non-Reading First schools for the selected baseline characteristics. The second column presents *p*-values for each of these residual differences. *None* of the residual differences in the exhibit are statistically significant. Hence, there is little evidence of residual differences in these school-level baseline characteristics. Results shown in the exhibit do not provide statistical evidence of substantial bias in impact estimates for the present report. Also, because impact estimates for student reading comprehension control explicitly for observed differences in school-level mean baseline test scores (typically the strongest predictor of future test scores), they provide further protection against bias.

***Statistical Significance***

Two-tailed t-tests are used to assess the statistical significance of impact estimates, and an asterisk (*) denotes statistically significant estimates at the conventional 0.05 probability level. The 0.05 standard for statistical significance implies that if a true impact is zero, there is only a one-in-twenty chance that its estimate will be statistically significant. Statistical significance does not represent the size, meaning, or importance of an impact estimate. It only indicates the probability that it occurred by chance. For example, a statistically significant impact estimate is not necessarily policy relevant; it is large enough that it is likely not due entirely to chance. This could occur for a small impact estimate from a large sample, for which the actual size of the estimated impact might not be deemed substantively meaningful, even though it was statistically significant. Lack of statistical significance for an impact estimate does not mean that the impact being estimated equals zero, only that that estimate cannot be distinguished from zero reliably. This could occur for a large impact estimate from a small sample, for which the actual size of the estimated impact might be substantively meaningful, although there is uncertainty about the estimate.

The Reading First Impact Study focuses on several different outcomes and subgroups, and therefore estimates numerous impacts. Each individual estimate has only a 5 percent chance of falsely indicating an impact's statistical significance when there is no impact. However, the group of estimates together has a much greater chance of falsely indicating that some impacts are statistically significant, even if none are.

[19] Baseline data were available at the school level only.

**Exhibit 2.2: Estimated Residual Differences in Baseline Characteristics of Schools in the Study Sample: 2002-2003**

| Characteristic | Estimated Residual Difference | Statistical Significance of Difference (p-value) |
|---|---|---|
| **Students** | | |
| Male (%) | 0.9 | (0.246) |
| Race (%) | | |
| Asian | 0.9 | (0.363) |
| Black | -7.2 | (0.199) |
| Hispanic | 3.3 | (0.345) |
| White | 2.8 | (0.503) |
| American Indian/Alaskan | 0.2 | (0.182) |
| Free Lunch and Reduced Lunch (%) | -6.0 | (0.073) |
| **Schools** | | |
| Eligible for Title I (%) | -1.4 | (0.802) |
| Locale (%) | | |
| Large City | 4.3 | (0.419) |
| Mid-size City | 9.1 | (0.108) |
| Other[a] | -13.4 | (0.083) |
| Size | | |
| Total Number of Students | -0.9 | (0.982) |
| Number of Students in Grade 3 | -3.8 | (0.558) |
| Student/Teacher Ratio | 0.1 | (0.861) |
| **Third Grade Reading Performance** | | |
| Deviation from State RF Mean Proficiency Rate (%)[b] | 4.3 | (0.085) |

**Notes:**

**The complete RF study sample includes 248 schools from 18 sites (17 districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools.**

The "Estimated Residual Difference" is the adjusted residual differences between Reading First schools and non-Reading First schools estimated using the regression discontinuity model, which controls for each school's rating.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

[a] Other Locale includes urban fringe of a large city, urban fringe of a mid-sized city, large town, small town, and rural.

[b] A school's proficiency score is defined as the percentage of third grade students (or fourth or fifth grade when third grade is unavailable) in the school that score at or above the state-defined proficiency threshold on the state's reading assessment. The values in this row represent the average percentage point deviation from the mean proficiency score for the Reading First schools in the state.

**EXHIBIT READS: The estimated residual difference on the percent of male students between Reading First and non-Reading First schools was 0.9 percentage points. The difference was not statistically significant at the $p \leq .05$ level (p=.246).**

*Sources: Data on baseline characteristics are from the Common Core of Data.*

Given the study's broad research questions, the number of impacts estimated was limited to the minimum possible to reduce the problem of "multiple hypotheses testing."[20] As a further safeguard, composite hypothesis tests were used to assess the overall statistical significance for groups of impact estimates within outcome domains. These composite tests measure the statistical significance of impact estimates that are pooled across outcome measures, subgroups, or both. A statistically significant composite test would suggest that some of its components are statistically significant. If the composite test is not statistically significant, the statistically significant findings for its

[20] Researchers disagree about whether and how to account for multiple hypothesis testing (e.g., Gelman and Stern, 2006; Shaffer, 1995).

components might be due to chance. The composite tests therefore help to "qualify" or call into question statements that are based on individual findings.[21]

***Statistical Precision***

The statistical precision of an impact estimator reflects its ability to detect true intervention effects when they exist. A common way to represent precision is a minimum detectable effect (MDE), which is the smallest true effect that an estimator has a "good chance" of detecting (Bloom, 1995). The current analysis uses the standard convention of defining a minimum detectable effect as the smallest true impact that has an 80 percent chance of being found to be statistically significant (it has 80 percent statistical power) at the 0.05 level of statistical significance for a two-tailed test of the null hypothesis of no effect. When a minimum detectable effect is expressed as a standardized effect size (in standard deviation units), it is referred to as a minimum detectable effect size (MDES).

Exhibit 2.3 reports minimum detectable effects and effect sizes for estimates of program impacts on the two most central study outcomes for the full study sample (e.g., student reading comprehension and amount of time on instruction in the five dimensions of reading instruction). These findings are based on data from the two follow-up years for which information is now available, rather than the initial assumptions that guided the study design. As such, they represent the actual precision of the design. The top panel in Exhibit 2.3 presents MDEs for the study's two measures of student achievement, average scores in reading comprehension on the SAT 10 and percent at or above grade level, while the bottom panel presents information on the study's primary measure of classroom instruction, average time per daily reading block spent on the five essential components of reading instruction (phonemic awareness, phonics, fluency, vocabulary, and comprehension). Columns in the table provide findings for each grade.

Minimum detectable effects for reading comprehension range from about 6 to 8 scaled-score points, corresponding to standardized effect sizes of roughly 0.15 to 0.16 standard deviations, even smaller than the 0.20 standard deviations that the study was initially designed to detect.[22] The minimum detectable effect is about 7 to 8 percentage points with respect to the percentage of students who read at or above grade level. The minimum detectable effect for "time in the five dimensions" is about 8 minutes, or roughly 0.38 standard deviations when expressed as an effect size. Because the study conducted some analyses at the subgroup level, MDEs were also calculated for a subgroup comprising about half of the schools in the sample for which a minimum detectable effect equals about $\sqrt{2}$ or 1.4 times the minimum detectable effect for the full sample.[23]

[21] See Appendix B for a detailed discussion of the study's approach to multiple hypothesis testing.

[22] See Gamse et al. (2004).

[23] See Appendix B, Exhibit B.16, for a table of MDEs for the study's key outcome measures by grade.

**Exhibit 2.3: Minimal Detectable Effects for Full Sample Impact Estimates**

| | Grade Level | | |
|---|---|---|---|
| | **Grade 1** | **Grade 2** | **Grade 3** |
| ***Panel 1*** | | | |
| **Student Reading Comprehension** | | | |
| Mean Scaled Score | 8.04 | 6.75 | 6.08 |
| Effect Size | 0.16 | 0.16 | 0.15 |
| Percent at or above grade level | 7.81 | 7.28 | 7.11 |
| ***Panel 2*** | | | |
| **Instructional Outcomes** | | | |
| Instruction in the five dimensions combined | | | |
| Minutes | 7.87 | 7.98 | N/A |
| Effect Size | 0.38 | 0.38 | N/A |

**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

Minimal detectable effects are based on the standard errors and standard deviations of the impact estimates for the full sample pooled across two school years of follow-up.

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

**EXHIBIT READS: The minimal detectable effect of the Reading First program on reading comprehension for a mean scaled score in grade 1 is 8.04 scaled score points. The minimal detectable effect of the Reading First program on reading comprehension for a mean scaled score in grade 2 is 6.75 scaled score points. The minimal detectable effect of the Reading First program on reading comprehension for a mean scaled score in grade 3 is 6.08 scaled score points.**

*Sources: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); and RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006.*

## The Study Sample

The initial sample for the Reading First Impact Study contained 258 schools (half in the Reading First program and half in a comparison group) from 17 school districts plus a statewide program. For reasons discussed below, 10 schools dropped out of the study. The 18 study sites are located in 13 states, which received their Reading First grants over a 16-month period, from June 2002 to September 2003. Sites received their sub-grants between April 2003 and August 2004 (Appendix A provides information on award dates by site).

The following criteria determined sites' eligibility for a regression discontinuity analysis of the impacts of Reading First:

- Sites used quantifiable criteria to rate schools eligible for RF funds **and** had at least three more schools than could be funded, thereby providing a minimum of three comparison schools. Any quantifiable criteria could be used to rate schools.
- Sites' decisions about school ratings and the determination of their local funding cut-points were made independently of one another.

- Sites' funding decisions about schools were based only on their ratings and their site's cut-point. These decisions were not overridden by other considerations.

Exhibit 2.4 indicates that 29 sites met the above criteria. From this pool of 29 candidate sites, a final sample of 18 sites was chosen. The final site selection attempted to balance such factors as: geographic diversity, inclusion of both larger and moderate-size districts (small districts would not contribute adequately to overall sample size), and a desire to avoid districts that were participating in other major evaluation studies. As noted previously, the study team selected 17 regression discontinuity sites and one additional site agreed to conduct a group-randomized experiment.

Once sites were identified, local schools were chosen as follows:

- From each site, a sample of schools located as close as possible to just above and just below the local cut-point were selected. This was done to minimize pre-existing differences among schools. Half of the schools chosen were Reading First schools and half were non-Reading First schools.
- Reading First and non-Reading First schools from a given site were chosen to have as similar a range of ratings as possible above and below the local cut-point. This was done in order to avoid asymmetries between the treatment group and comparison group. In addition, schools were chosen to avoid large gaps in the rating distribution, which could mask non-linearities.

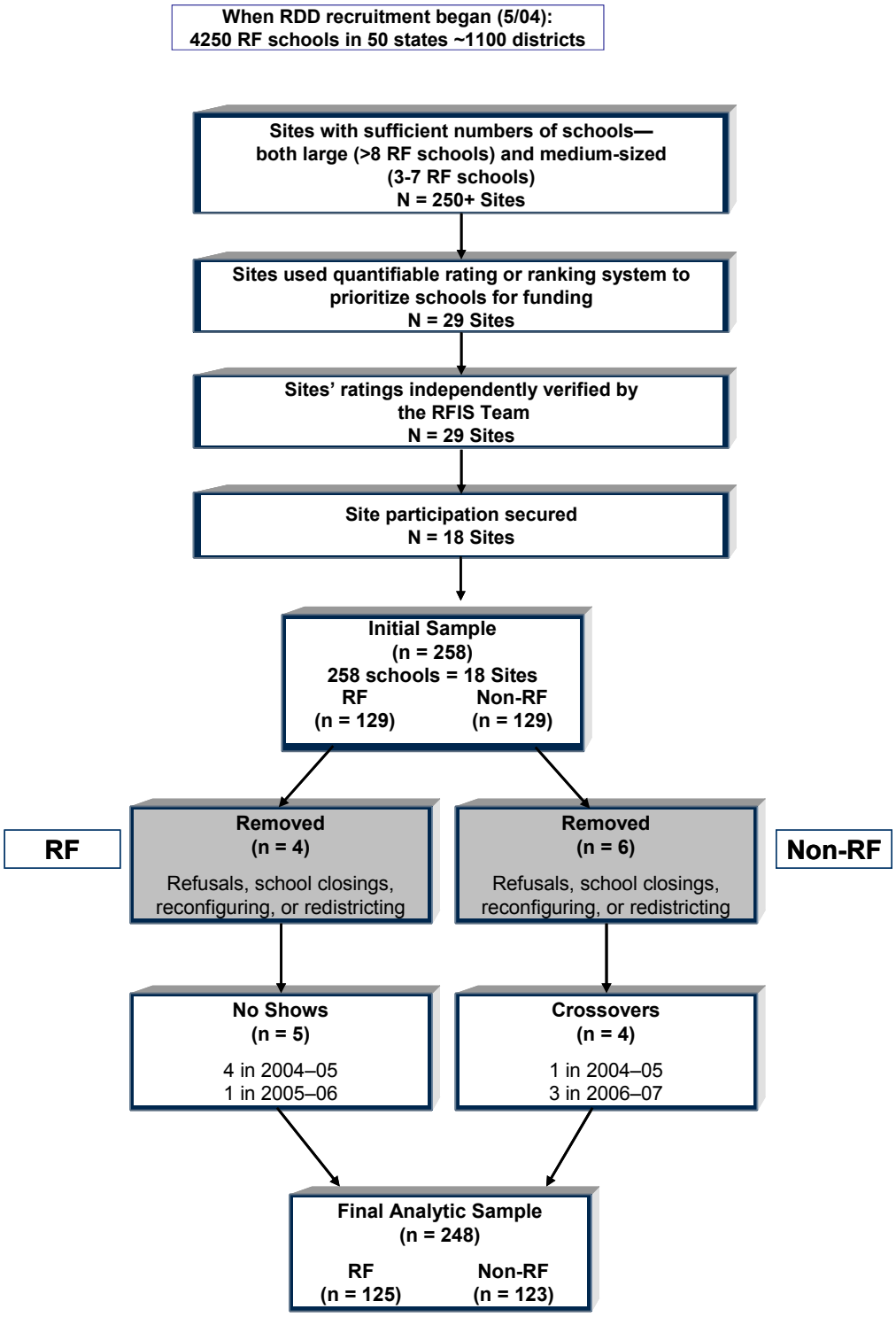
Information about the ratings, cut-points, and numbers of schools rated and funded from each of the 17 RDD sites is presented in Exhibit 2.5.[24] Ratings were based mainly on measures of student reading performance (standardized test scores) and/or poverty (eligibility for free or reduced price lunch) (Gamse et al., 2004). In two sites, eligible schools submitted proposals for funding that were rated according to locally determined criteria. The criteria that sites used to rate and fund Reading First schools reflect Reading First programmatic emphases, i.e., to serve the lowest performing and/or the neediest schools.[25] The exhibit's first column indicates, for each site, which criteria were used. The second column presents the number of schools that were rated, and, in parentheses, funded. The cut-point score for each site is presented in the box in the center of each shaded bar. The numbers in the shaded bars represent the numbers of RF and non-RF schools for each site (non-RF in the left, RF in the right side of each shaded bar). The numbers to the far left of each shaded bar represent the lowest rating for all non-funded (i.e., non-Reading First) schools, and then the rating for the lowest-rated school in the study sample. The numbers to the far right of each shaded bar represent the highest rating for all funded schools, and the highest rating for funded schools in the study sample is immediately to the right of each shaded bar.

The 248 schools in the initial sample represent about 37 percent of all rated schools in the 17 RDD sites. (This number does not include the 10 schools in the experimental site.) The 129 Reading First

[24] Exhibit 2.5 does not include the one site that agreed to random assignment for its 10 RFIS schools.

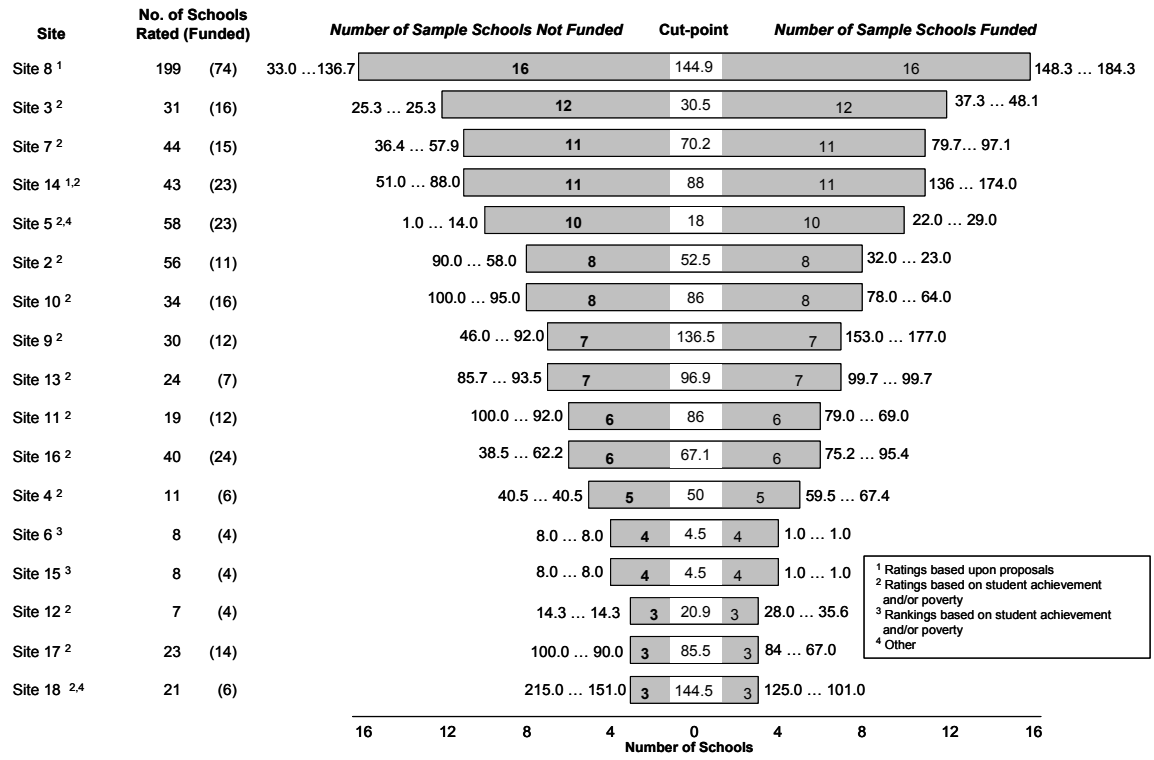
[25] The site that agreed to random assignment to RF or non-RF status also determined its schools' eligibility on the basis of prior student achievement and poverty. In that site, 12 of 17 eligible schools were funded; 5 schools were funded via random assignment, and 7 schools were selected by the site.

**Exhibit 2.4: RFIS Sample Selection: From Regression Discontinuity Design Target Sample to Analytic Sample**




*The final analytic sample includes 146 schools from 7 sites that have 8 or more RF schools (74 RF, 72 non-RF schools) and 102 schools from 6 sites that have between 3 and 7 RF schools (51 RF, 51 non-RF schools).

**Exhibit 2.5: Numbers, Ratings, and Cut-points for Selection of Reading First and Reading First Impact Study Schools, by Site (Initial Sample for 17 Sites, Excluding Random Assignment Site)**




**Notes:**

Ratings varied in directionality and metrics; in some sites, higher scores indicated greater needs; in other sites, lower scores indicated greater needs.

**EXHIBIT READS: Site 8 rated 199 schools, and funded 74 schools. The RFIS sample in Site 8 included 32 schools—16 non-Reading First schools and 16 Reading First schools—that were rated from 136.7 to 148.3, shown at the left and right sides of the shaded bar, respectively. The cut-point was at 144.9. The lowest school rating was 33, and the highest school rating was 184.3.**

*Sources: Interviews with sites' Reading First coordinators in 2004.*

schools in the initial study sample represent 46 percent of all Reading First schools at the study sites.[26, 27] Because schools in the RFIS sample are broadly distributed across sites, study findings are unlikely to be dominated by one or two sites. The final analytic sample contains 248 schools (125 Reading First schools and 123 non-Reading First schools). Ten of the original study schools dropped

[26] Many states and districts have subsequently held additional grant competitions, and the number of funded Reading First schools within districts may have since changed as a result.

[27] The number of Reading First and non-Reading First schools was initially equivalent; in three sites, the number is no longer equivalent, reflecting the closing or reconfiguration of several schools after they had been chosen for the study sample.

out because of subsequent closures, reconfigurations, or refusals.[28] Only nine of 248 schools in the analytic sample for the present report changed program status after it was determined by their ratings; five schools whose ratings qualified them for funding did not receive it; four schools with ratings that did not qualify them for funding subsequently received funding. No-shows and cross-overs were included, however, in the study's data collection and analyses. In the analysis, schools are assigned to the group (with or without Reading First) defined by their rating even though their program status may have subsequently changed.

The discussion above indicates that the stability of the study sample satisfies the requirements for an internally valid regression discontinuity analysis. The discussion below assesses the sample's generalizability or external validity.

### Representativeness of the Sample

Although the RFIS sample is not a national probability sample, it shares many important characteristics with the national Reading First population. One way to examine these characteristics is to compare baseline characteristics of the sample to those of: (1) all Reading First schools in the 18 study sites; (2) all Reading First schools in the sample's 13 states; and (3) all Reading First schools in the U.S.

Exhibit 2.6 illustrates how these groups are related. At the center are the 125 Reading First schools in the final study sample, which is a subset of the 274 Reading First schools in the study sites, and that is a subset of all 1,728 Reading First schools in the 13 states with a study site. The outermost level of the figure represents all 4,793 Reading First schools nationally (as of June 2005).[29]

Exhibit 2.7 compares baseline characteristics and student reading achievement for the RFIS Reading First schools and the three other groups of Reading First schools. These comparisons are based on information from the national Common Core of Data (CCD) database as well as from a national assessment database maintained by the U.S. Department of Education. The information presented is for the most recent year before Reading First was funded at any RFIS site (2002-03). The exhibit compares student characteristics, school characteristics, and prior third grade reading performance. Visual inspection of the data displayed in this exhibit suggests that, overall, the present sample is similar to the other three groups of Reading First schools. Almost all are eligible for Title I support, they enroll high percentages of students eligible for free or reduced price lunch, and their past third grade reading scores are near their state averages for Reading First schools. The RFIS sample, on average, has proportionally lower percentages of Hispanic students and higher percentages of Black students than Reading First schools in the study states or in the nation; at the same time, RFIS sample schools, on average, have a lower percentage of Black students and a higher percentage of White

[28] Ten schools were removed from the initial sample. Three comparison schools refused to participate; all were in districts (in the same state) that had received *no* Reading First funding, and the districts asserted that absent any RF funding, they were not obligated to participate in the study. Two RF schools and two comparison schools were subsequently closed; two RF schools were substantially reconfigured (entirely new faculty and staff); and one comparison school merged with a Reading First school.

[29] See http://www.sedl.org/readingfirst/ for further information about all Reading First schools nationwide.

**Exhibit 2.6: Relevant Groups of Reading First Schools**

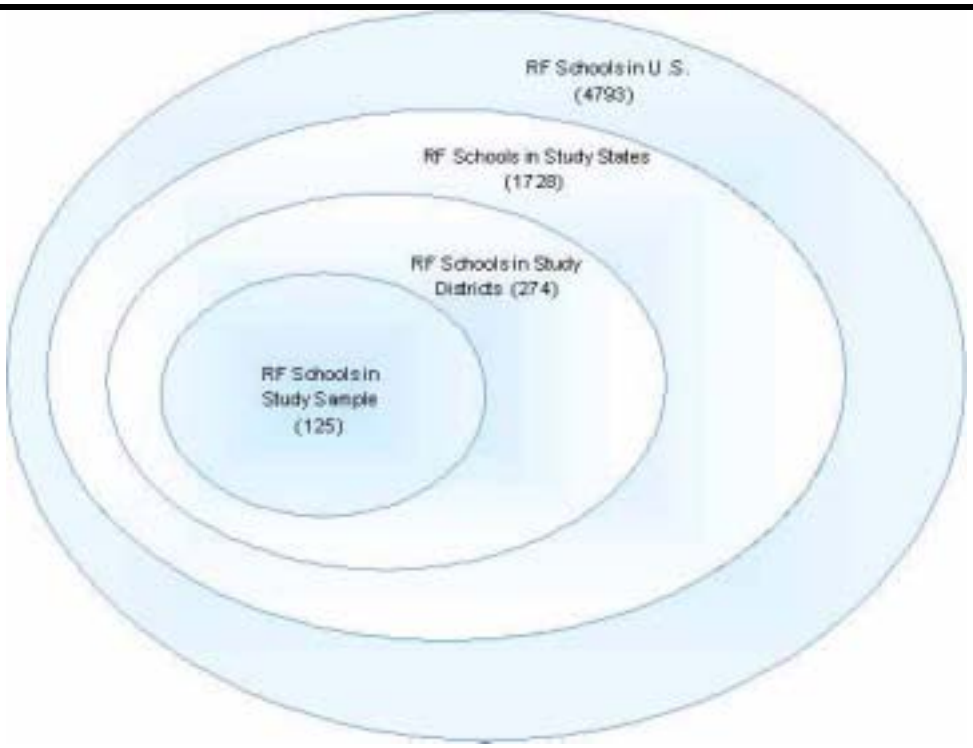


**Exhibit 2.7: Baseline Characteristics of Relevant Groups of Reading First Schools for 2002-2003**

| Characteristic | RF Schools in Study Sample | RF Schools in Study Districts | RF Schools in Study States | RF Schools in U.S. |
|---|---|---|---|---|
| **Students** | | | | |
| Male (%) | 52.3 | 52.0 | 51.7 | 51.5 |
| Race (%) | | | | |
| Asian | 3.1 | 2.5 | 1.5 | 3.5 |
| Black | 35.6 | 41.1 | 26.4 | 30.5 |
| Hispanic | 26.7 | 28.6 | 37.1 | 34.8 |
| White | 34.2 | 27.4 | 34.3 | 28.6 |
| American Indian/Alaskan | 0.5 | 0.4 | 0.6 | 2.5 |
| Free Lunch and Reduced Lunch (%) | 74.4 | 75.0 | 67.8 | 73.2 |
| **Schools** | | | | |
| Eligible for Title 1(%) | 97.6 | 97.4 | 96.4 | 94.8 |
| Locale (%) | | | | |
| Large City | 39.2 | 39.8 | 26.7 | 26.8 |
| Mid-size City | 36.8 | 36.5 | 21.0 | 19.5 |
| Other[a] | 24.0 | 23.7 | 52.3 | 53.6 |
| Size | | | | |
| Total Number of Students | 474.8 | 487.4 | 502.4 | 531.4 |
| Number of Students in Grade 3 | 71.6 | 75.1 | 80.2 | 84.9 |
| Student/Teacher Ratio | 15.1 | 14.8 | 15.1 | 16.5 |
| **Third Grade Reading Performance** | | | | |
| Deviation from State RF Mean Proficiency Rate (%)[b] | -1.3 | -3.3 | 0.0 | 0.0 |
| **Number of Schools[c]** | 125 | 274 | 1,728 | 4,793 |

**Notes:**

The RF study sample includes 128 schools from 18 sites (17 districts and 1 state) located in 13 states. The RF schools in Study Districts include all RF schools ranked and/or rated on the RF grant application for each of the 18 sites in the study. All RF schools in Study States include all RF schools in the 13 states included in the study. All RF schools nationally include all schools that received RF grants.

[a] Other Locale includes urban fringe of a large city, urban fringe of a mid-sized city, large town, small town, and rural.

[b] A school's proficiency score is defined as the percentage of third grade students (or fourth or fifth grade when third grade is unavailable) in the school that score at or above the state-defined proficiency threshold on the state's reading assessment. The values in this row represent the average percentage point deviation from the mean proficiency score for the Reading First schools in the state. By definition, for a given state, the mean proficiency score for all Reading First schools in the state is the benchmark for comparison. Therefore, in the final two columns, the deviation from the benchmark within each state is zero and the average deviation across states is zero.

[c] Due to missing values for some variables, the number of schools included varies by characteristic.

*Sources:* *Baseline characteristic data are from the Common Core of Data. RF school samples are defined based on information from the Southwest Educational Development Laboratory.*

students than Reading First schools in study districts. A greater proportion of Reading First schools in the study sample are in large or mid-size cities and not other locales than are Reading First schools in the study states or in the nation. Also, the sizes of Reading First schools in the study sample, on average, are somewhat smaller than those in the three other groups. Further, these data cannot provide conclusive evidence that the study sample fully represents the experience of the entire national Reading First program, as the study sample might differ from the Reading First population in other ways that were not observed.

Exhibit 2.7 also presents information on the proportion of third grade students' test scores at or above their state proficiency threshold for reading, using an index that accounts for differences in states' reading tests' difficulty and established proficiency standards. The index reflects the mean percentage of third-grade students who performed at or above their state proficiency threshold for the 2002-03 year; a positive value would indicate a school's proficiency rate is above its statewide average, and a negative value would mean a school's performance is below the statewide average.

The mean for the study's sample of Reading First schools was -1.3 percentage points, which is below the statewide Reading First mean just before the program began. The RFIS schools' average proficiency is closer to the statewide Reading First mean than that of other Reading First schools in study districts, because the RFIS sample includes schools that are closer to their respective cut-points, and therefore had somewhat stronger academic performance than did all of the RF schools in study sites. (Recall that RFIS schools were rated on the basis of past student performance, and schools closest to the cut-point had higher academic performance, on average, than schools further away from the cut-point).

Exhibit 2.8 provides another way to examine the study sample's similarity to other Reading First schools nationally. It summarizes responses from surveys administered in spring 2005 to principals, reading coaches, and teachers from RFIS Reading First schools and from the Reading First Implementation Study, which surveyed a large, nationally representative sample of Reading First schools. Visual inspection of the data presented in this exhibit suggests that, overall, survey respondents' reports from the two studies are similar. The differences include:

(1) a smaller percentage of students for whom English is a second language (10.8 percent versus 20.3 percent) for the RFIS sample;
(2) a higher percentage of students who read at or above grade level (50.2 percent versus 46.9 percent) for the RFIS sample; and
(3) a higher percentage of schools making Adequate Yearly Progress (75.3 percent versus 69.9 percent) for the RFIS sample.

The discussion above indicates three important features of the study design and sample. One, the RFIS regression discontinuity design will yield unbiased impact estimates. Two, the RFIS sample size is adequate to detect impacts of less than 0.20 standard deviation units on student reading achievement. Three, although the sample of RF schools in the study was selected opportunistically, it is generally similar to other RF schools in the national program as of September 2004.

The next chapter reviews the study data collection activities and describes the measures used to assess the impacts of Reading First.

**Exhibit 2.8: School-Level Characteristics of Reading First Schools in the Reading First Impact Study and the Reading First Implementation Study for 2004-2005**

| Characteristic | Reading First Schools in RFIS Sample (n=125) Mean | Reading First Schools in National Sample (n=1,092) Mean |
|---|---|---|
| **Principals** | | |
| Years in This School | 5.7 | 4.8 |
| **Reading Coaches** | | |
| Years of Experience | 16.0 | 18.0 |
| Years as Reading Coach in This School | 1.8 | 1.8 |
| **Teachers**[1] | | |
| Years of Experience | 11.9 | 12.9 |
| Full Certification (%) | 93.1 | 93.0 |
| Highly Qualified[2] | 89.0 | 87.8 |
| **Students** | | |
| Reading At or Above Grade Level (%) | 50.2 | 46.9 |
| Participating in Interventions for Struggling Readers (%) | 35.1 | 34.3 |
| Special Education Services (%) | 9.3 | 7.6 |
| English as a Second Language Instruction (%) | 10.8 | 20.3 |
| Instruction in a Language Other than English (%) | 3.1 | 6.5 |
| **School Performance** | | |
| Adequate Yearly Progress (AYP)[3] | 75.3 | 69.9 |

**Notes:**

Missing values were imputed from district- or state-level means.

[1] Data, with the exception of that for "highly qualified teachers," were taken from the teacher surveys and aggregated to the school level for the purposes of these comparisons. Thus, mean teacher experience for each school was compared.

[2] A "highly-qualified teacher" is one who meets three criteria: 1) full state certification; 2) at least a bachelor's degree; and 3) proven knowledge of the subject taught. These data are taken from the principal surveys.

[3] "Adequate Yearly Progress" (AYP) is the amount of yearly improvement each school is expected to make. Each state is responsible for defining and measuring AYP. This figure is the percent of schools in the sample that met AYP in the previous school year.

*Sources:* *The Reading First Impact Study Principal, Reading Coach, and Teacher Surveys; and the Reading First Implementation Study Principal, Reading Coach, and Teacher Surveys.*

# Chapter Three: Measures and Data Collection

The Reading First program provides resources to states, districts, and schools to improve the effectiveness of reading instruction, and ultimately, to improve students' reading performance such that by the end of third grade, students will be able to read at or above grade level. The programmatic focus on improved classroom instruction and a clearly articulated reading comprehension goal led the study team to concentrate its data collection activities on teachers' instructional practices and students' reading comprehension skills.

The study design draws upon a variety of data sources to address its evaluation questions. Exhibit 3.1 summarizes the data collection schedule for the study as a whole: student reading comprehension data (standardized test scores), observations during classroom instruction, surveys of school personnel, and district staff interviews. This interim report is based on data collected during the 2004-05 and 2005-06 school years. Exhibit 3.2 provides information about the number of respondents for each type of data collection activity. Exhibit 3.3 provides a description of the measures utilized in the study.

The RFIS draws upon student achievement data from the Stanford Achievement Test, 10th Edition (SAT 10, Harcourt Assessment, Inc., 2004) reading comprehension subtest, administered in the fall of 2004, the spring of 2005 and the spring of 2006.[30] The RFIS has tested over 10,000 students in each grade level (grades 1, 2, and 3) in each round of testing.

**Exhibit 3.1: Data Collection Schedule for the Reading First Impact Study**

| | 2004-2005 | | 2005-2006 | | 2006-2007 | |
|---|---|---|---|---|---|---|
| **Data Collection Elements** | **Fall** | **Spring** | **Fall** | **Spring** | **Fall** | **Spring** |
| Student Testing | ✓ | ✓ | | ✓ | | ✓ |
| Classroom Observations | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Teacher, Principal, Reading Coach Surveys | | ✓ | | | | ✓ |
| District Staff Interviews | | ✓ | | | | ✓ |

[30] Students in two of the RFIS's 18 sites were excluded from fall 2004 testing as a result of hurricane-related school closures. Those students were tested in subsequent data collections.

**Exhibit 3.2: Summary of RFIS Data Collection Activities and Respective Response Rates, by Grade**

| | Fall 2004 | | | | Spring 2005 | | | | Fall 2005 | | | | Spring 2006 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RF | | Non-RF | | RF | | Non-RF | | RF | | Non-RF | | RF | | Non-RF | |
| | N | (%) | N | (%) | N | (%) | N | (%) | N | (%) | N | (%) | N | (%) | N | (%) |
| *Student assessments* | | | | | | | | | | | | | | | | |
| *Grade 1* | 7,563 | (72) | 7,492 | (69) | 9,225 | (84) | 8,786 | (80) | | | | | 7,552 | (86) | 6,576 | (85) |
| *Grade 2* | 7,289 | (71) | 7,160 | (70) | 8,867 | (85) | 8,611 | (82) | | | | | 7,514 | (86) | 6,582 | (85) |
| *Grade 3* | 7,208 | (73) | 7,063 | (69) | 8,748 | (84) | 8,399 | (84) | | | | | 7,220 | (87) | 6,953 | (87) |
| *Classroom observations (reading instruction)* | | | | | | | | | | | | | | | | |
| *Grade 1* | | | | | 809 | (97) | 820 | (96) | 720 | (98) | 704 | (98) | 718 | (99) | 707 | (99) |
| *Grade 2* | | | | | 766 | (96) | 760 | (95) | 664 | (97) | 668 | (98) | 666 | (100) | 668 | (100) |
| *Classroom observations (student engagement)* | | | | | | | | | | | | | | | | |
| *Grade 1 + 2* | | | | | | | | | 683 | (98) | 678 | (99) | 677 | (97) | 677 | (98) |
| *Surveys: Teacher* | | | | | | | | | | | | | | | | |
| *Grade 1* | | | | | 396 | (73) | 363 | (67) | | | | | | | | |
| *Grade 2* | | | | | 362 | (73) | 319 | (65) | | | | | | | | |
| *Grade 3* | | | | | 318 | (71) | 279 | (64) | | | | | | | | |
| *Reading Coach* | | | | | 118 | (95) | 79 | (72) | | | | | | | | |
| *Principal* | | | | | 98 | (78) | 89 | (72) | | | | | | | | |
| *Site/District interviews* | | | | | 18 | (100) | 18 | (100) | | | | | | | | |

**Notes:**

Blank cells indicate no data collection for that component at that time period. Response rates shown are for the analytic sample of 248 schools.

Active consent (i.e., only students whose parents had signed and returned consent forms) was used in fall 2004. Passive consent (i.e., all eligible students were tested unless their parents submitted forms refusing to allow their children to be tested) was used in subsequent test administrations.

Reading instruction in each classroom was observed on two consecutive days in each wave of data collection. Observations of student engagement were scheduled for the same classrooms as observations of teachers' reading instruction. Observations of student engagement occurred on one of the two days during which reading instruction was observed (see Appendix C for a complete discussion of the observation protocols).

**EXHIBIT READS: In fall 2004, 7563 student assessments were completed in Reading First grade 1 classrooms, corresponding to 72 percent of all eligible grade 1 classrooms.**

**Exhibit 3.3: Description of Measures Utilized in the Reading First Impact Study**

| Domain | Outcome Measure and Description |
|---|---|
| **Student reading comprehension** | **Two outcome variables**<br>**Mean scaled scores** on the reading comprehension subtest of the Stanford Achievement Test, 10th Edition (SAT 10), represented as a continuous measure of student reading comprehension. Because scaled scores are continuous across grade levels, values for all three grade levels can be shown on a single set of axes.<br>**Percentage of students at or above grade level** on the SAT 10, based upon established test norms that correspond to grade level performance, by grade and month. The on or above grade level performance percentages were based on the start of the school year, date of the test and the scaled score, as well as the related grade equivalent. |
| **Classroom reading instruction** | **Eight outcome variables**<br>**Minutes of instruction in phonemic awareness,** or how much instructional time teachers spent on phonemic awareness, from the Instructional Practice in Reading Inventory (IPRI) observational data.<br>**Minutes of instruction in decoding,** or how much instructional time teachers spent on decoding, from IPRI observational data.<br>**Minutes of instruction in fluency building**, or how much instructional time teachers spent on fluency building, from IPRI observational data.<br>**Minutes of instruction in vocabulary development**, or how much instructional time teachers spent on vocabulary development, from IPRI observational data.<br>**Minutes of instruction in comprehension**, or how much instructional time teachers spent on comprehension of connected text, from IPRI observational data.<br>**Minutes of instruction in all five dimensions combined**, or how much instructional time teachers spent on all five dimensions combined, from IPRI observational data.<br>**Proportion of each observation with highly explicit instruction**, or the proportion of time spent within the five dimensions when teachers used highly explicit instruction, from IPRI observational data (e.g., instruction included teacher modeling, clear explanations, and the use of examples).<br>**Proportion of each observation with high quality student practice**, or the proportion of time spent within the five dimensions when teachers provided students with high quality student practice opportunities, from IPRI observational data (e.g., teachers asked students to practice such word learning strategies as context, word structure, and meanings). |
| **Student engagement with print** | **One outcome variable**<br>**Percentage of students engaged with print**, from the Student Time-on-Task and Engagement with Print (STEP) observational data, represented as the per-classroom average of the percentage of students engaged with print across three sweeps in each classroom during observed reading instruction. |

Note: For more information on measures, see Appendix C.

In spring 2005, the RFIS conducted classroom observations of reading instruction in first and second grade classrooms. The RFIS observed reading instruction for two consecutive days in designated classrooms; the response rate for these observations was 96 percent, on average. The RFIS observed in first and second grade classrooms in both fall 2005 and spring 2006, for two consecutive days each. Response rates were 97 percent and above for both Reading First and comparison classrooms. Over 1,400 classrooms were observed at each time point (see Exhibit 3.2).

A measure of the percentage of students on-task and engaged with print was added to the RFIS data collection in fall 2005. Over 1,300 classrooms were observed using this measure in both fall 2005 and spring 2006, with a response rate of over 97 percent at each time point.

Surveys of teachers, reading coaches, or reading specialists (non-Reading First schools do not universally have reading coaches), and school principals were fielded in spring 2005 with combined RF and non-RF response rates of 69 percent for teachers, 84 percent for reading coaches, and 75 percent for principals.[31]

## Student Reading Comprehension

At the heart of this evaluation is a question about the impact of Reading First on the reading achievement of students. To answer this question, the study must obtain valid and reliable measures of reading performance from students in RF and non-RF schools. The RFIS had initially planned to use a battery of individually-administered tests to assess students across the specific components of reading instruction targeted by the legislation: phonemic awareness, phonics, fluency, vocabulary and comprehension (No Child Left Behind Act, 2001). When the study's design shifted to a RDD, with a quadrupled number of schools and students in the study sample, the individualized student assessment data collection was no longer practical.

The RFIS Team, working with its Technical Work Group and staff from the National Center for Educational Evaluation, Institute of Education Sciences, at the U.S. Department of Education, focused on identifying a single test (or subtests) to measure students' ability to comprehend text, such as subtests of word reading, vocabulary, listening comprehension, and/or reading comprehension. Reading comprehension was selected, rather than other dimensions of early reading skill, because comprehension is perceived as "the essence of reading" that sets the stage for children's later academic success (Durkin, 1993, p.12; National Institute of Child Health and Human Development, 2000; Stevens, Slavin, and Farnish, 1991).

The priorities in selecting a test for the RFIS included the following:

- direct measurement of skills related to text comprehension;

[31] As a condition of approval to collect survey data for this study, the Office of Management and Budget required the RFIS to conduct a study of the effect of incentives on survey response rates for teachers. Schools within districts were randomly assigned to one of three incentive conditions: $30, $15, and $0. Six sites, representing 39 schools, refused to participate in the incentive sub-study because their labor contracts mandate that district employees be compensated equally for completion of identical tasks, thereby reducing the number of schools in the sub-study from 251 to 215. The results of the incentive sub-study indicated that incentives significantly increased response rates; as a result, in future waves of survey administration, all respondents will be eligible for incentives.

- ease and appropriateness of administration to groups or entire classrooms of students—including modest time demands;
- appropriateness for first, second, and third grade students—including those students just beginning first grade in the fall;
- use of a norm-referenced test—which would provide a national norming sample, thereby allowing the study to ascertain the absolute level of reading comprehension for Reading First and other students;
- consistent reliability and validity data from norming samples and prior research;
- evidence of prior use on a large scale, which therefore would render its use more credible in the research community; and
- selecting an assessment already used by some localities/states, either for statewide testing or for Reading First assessment purposes, in order to minimize additional testing.

The RFIS selected the Stanford Achievement Test, 10th Edition, because it best met the criteria outlined above.[32]

The outcome measures used to assess reading comprehension are student test scores on the SAT 10, reported in terms of continuous scaled scores as well as in terms of the percentage of students who scored at or above grade level, according to established test norms that correspond to grade level performance, by grade and month.[33] The latter metric is also salient for this study, as one of the program's explicit objectives is to increase the number and percent of students who read at or above grade level (No Child Left Behind Act, 2001).

The RFIS administered the following subtests of the SAT 10 to first, second, and third grade students in the spring of 2005 and spring of 2006, respectively: the Primary 1 Reading Comprehension Subtest (40 items), the Primary 2 Reading Comprehension Subtest (40 items), and the Primary 3 Reading Comprehension Subtest (54 items). Where already administered, the RFIS obtained SAT 10 reading comprehension test score data from schools/districts for the grades of interest, which reduced the testing burden for those students and schools.[34]

Items from the SAT 10 reading comprehension subtests (different versions for first, second, and third grades) are multiple choice. Students must read sentences, paragraphs, or longer passages and select the response that either correctly completes a sentence describing a picture or answers a question

---

[32] See Appendix C for more information on SAT 10 selection, data collection, and response rates.

[33] The study team converted mean SAT 10 scaled scores to grade equivalents by using the scaled score to grade equivalents table in the *Stanford Achievement Test Series Fall Multilevel Norms Books* (based on 2002 normative data) to match the average scaled score to a grade equivalent (Harcourt Assessment, Inc., 2003, p. 24-26). In order to calculate at or above grade level, a dummy variable was created for each student based on the start of the school year, date of the test, and their scaled score (as well as the related grade equivalent). For more information on the construction and interpretation of grade level performance (percentile ranks and grade equivalents), please see pages 24-26 of the *Stanford Achievement Test Series Fall Multilevel Norms Book*.

[34] In Spring 2007, another assessment, the Test of Silent Word Reading Fluency (TOSWRF), was added for Grade 1 (Mather et al., 2004). Results from this test will be included in the final report.

about the passage. As the grade level increases, the length of passages increases and test items require higher levels of inference. A test proctor first reads aloud standardized instructions and guides students through one or two sample items for practice, giving feedback on the correct answers to sample items to ensure that students understand test directions. Then students complete test items on their own.[35]

## Reading Instruction

A key part of the evaluation is to determine the impact of Reading First on instruction in the targeted grades. Therefore, classroom observations of instructional practices in reading were needed from both RF and non-RF classrooms. Because the Reading First legislation calls for reading instruction to be based on scientifically based reading research findings, the RFIS observational instrument built upon findings describing evidence-based instructional practices such as those in the National Research Council's (1998) report (Snow, Burns, and Griffin, 1998) and the National Reading Panel report (National Institute of Child Health and Human Development, 2000). The Reading First legislation highlights five essential components of reading instruction. These five components, or dimensions, of reading instruction formed the basis for the development of the RFIS observation instrument.[36] Each dimension is described below.[37]

***Phonemic Awareness***

Phonemic awareness instruction teaches students to distinguish and manipulate the sounds in words.[38] A phoneme is the smallest unit of sound that affects the meaning of a spoken word. Before learning to read print, children must first understand that words are made up of component sounds. For example, changing the first phoneme in the word ***hat*** from /h/ to /p/ changes the word from ***hat*** to ***pat***. Phonemic awareness instruction improves children's word reading and helps children learn to spell (e.g., Ball and Blachman, 1991; Bus and van Ijzendoorn, 1999; see also NICHD, 2000).

***Decoding***

Decoding (also known as phonics) instruction helps children learn and understand the relationships between the letters of written language and the sounds (phonemes) of spoken language. Instruction in decoding helps children understand that there are predictable relationships between letters and sounds, helps them recognize familiar words, and allows children to "decode" unfamiliar printed words (see Chapter 2, Part II, NICHD, 2000).

---

[35] In Fall 2004, first-graders completed the SESAT2 version of the SAT 10. In this version, students listen to a test proctor who reads aloud each item (because students cannot necessarily read the printed test item at the start of first grade) and then select the correct response (a picture or word) in the test booklet (Harcourt Assessment, Inc., 2004).

[36] For ease of explication, the measures created from IPRI data are referred to as the five dimensions of reading instruction (or "the five dimensions") throughout the report. References to the programmatic emphases as required by legislation are labeled as the five essential components of reading instruction.

[37] See Appendix C, Exhibit C.3 for specific examples of instructional activities associated with each of the five dimensions.

[38] Phonemic awareness is a subcategory of phonological awareness. Phonological awareness includes phonemic awareness, but also refers to the ability to recognize and work with larger parts of spoken language, such as syllables and onsets and rimes.

***Fluency Building***

Fluency is the ability to read text accurately and smoothly. The more automatically students can read individual words, the more they can focus on understanding the meaning of whole sentences and passages (NICHD, 2000). Fluency instruction helps students who are learning to read by building a bridge between recognizing words more efficiently and comprehending the meaning of text (e.g., Reutzel and Hollingsworth, 1993; also see Chapter 3, NICHD, 2000).

***Vocabulary Development***

Oral vocabulary refers to words used in speaking or recognized in listening. Reading vocabulary refers to words that are recognized or used in print. Instruction for beginning readers uses oral vocabulary to help them make sense of the words they see, and instruction that develops their reading vocabulary allows them to progress to more complex texts (e.g., Beck, Perfetti and McKeown, 1982; McKeown et al., 1983; also see NICHD, 2000). Readers must know what words mean before they can understand what they are reading.

***Comprehension of Connected Text***

Comprehension is understanding what is being or has been read. Students will not understand text if they can read individual words, but do not understand what sentences, paragraphs, and longer passages mean. Proficient readers elicit meaning from—or comprehend—text, rather than simply identifying a series of words. Instruction in comprehension strategies provides specific tools for readers to use to make sense of the text they read (see NICHD, 2000). Comprehension strategies are vital to the development of competent readers because they aid in understanding the collective significance of words, sentences, and passages.

### Development of Classroom Observational Measures

To address the question about the impact of Reading First on classroom instruction, the study team needed to adopt or design a measure of classroom instruction that would allow comparison of RF and non-RF schools. It is important to note that the Reading First program is neither a specific intervention, nor a uniformly implemented program. Rather, Reading First is, at its core, a funding stream. Although the Reading First Program Guidance required schools and districts to implement scientifically based reading instruction, it did not require states or districts or schools to use the same core reading program (U.S. Department of Education, 2002).

Consequently, the RFIS team had to identify or develop an instrument sensitive enough to capture observed differences between Reading First and non-Reading First schools, while simultaneously flexible enough to accommodate a variety of instructional programs likely to be used by Reading First as well as by comparison schools. Preliminary reviews of existing instruments began shortly after the

study's September 2003 start. The study team soon determined that it would need to develop its own instrument, customized to assess the specific components of Reading First.[39]

The RFIS Team developed an instrument called the Instructional Practice in Reading Inventory (IPRI). The IPRI was designed to capture both pedagogical strategies and content across the five dimensions of reading instruction described above.[40] The instrument focuses specifically on teachers, reflecting the Reading First program's emphasis on changing teachers' instruction, specifically having teachers incorporate explicit instructional strategies and ample student practice opportunities (U.S. Department of Education, 2002, p. 6) within each of the five dimensions of reading instruction.

The RFIS team defined behaviors associated with those pedagogical objectives for each of the five dimensions of reading instruction. *Explicitness* includes modeling by the teacher as well as clear explanations of strategies, principles, or rules, with sufficient numbers of examples. *Explicit teaching* includes making relationships overt, emphasizing distinctive features of new concepts, and providing prompts. *Adequate practice* ensures that all students have multiple opportunities to practice new skills and review recently learned skills and concepts. Teachers need to assess *skill mastery* and provide ample *corrective* feedback both to assist students when they encounter difficulty as well as to ensure mastery of skills and strategies. This includes working towards a high level of response accuracy, monitoring student understanding and performance on an ongoing basis, eliciting responses from all students, and providing extra instruction, practice, and review. [41]

The instrument can be used for observations of varying lengths, reflecting the fact that schools' defined reading blocks can vary; most reading blocks are 90 minutes or more. Observers use a booklet containing a series of individual IPRI forms, each of which corresponds to a three-minute interval of observation. The observer watches the teacher for three minutes and records those target instructional behaviors that occur during the three-minute interval. Then, the observer turns the page to a new form and starts another three-minute observation, again recording the presence of targeted behaviors. Therefore, an observation of a reading block yields multiple and sequentially ordered IPRI forms. Observers wear a special wristwatch that vibrates every three minutes, which signals when to turn to a new form for the next three-minute interval. Over the course of the designated reading block, observers record approximately 30-35 separate three-minute intervals, on average, for each day of observation.

---

[39] Among the instruments reviewed were the following: *The Instructional Content Emphasis (ICE)* (Edmonds and Briggs, 2003); *Foorman and Schatschneider direct observation system and instruments from the Center for Academic and Reading Skills (CARS)* (Foorman and Schatschneider, 2003); *English Language Learner Classroom Observation Instrument (ELLCOI)* (Haager et al., 2003); *Teachers' Instructional Practice (TIP)* (Carlisle and Scott, 2003); *Utah's Profile of Scientifically-based Reading Research* (Dole et al., 2001); *The Classroom Observation Record* (Abt Associates and RMC Research, 2002); and *Observation Measure of Language and Literacy Instruction (OMLIT)*, developed by Abt Associates as part of the Even Start Classroom Literacy Interventions and Outcomes (CLIO) Study (Goodson et al., 2004).

[40] See Appendix C for a copy of the IPRI as well as a more comprehensive description of its development. See also *The Development of the Instructional Practice in Reading Inventory* (Dwyer et al., 2007).

[41] See, for example, Graves, Gerston, and Haager, 2004; Gunn et al., 2002 for specific examples of highly explicit instruction in phonemic awareness, and Foorman and Torgesen, 2001, Graves, Gerston, and Haager, 2004, for specific examples of highly explicit instruction in phonics. Exhibit C.6 in Appendix C identifies the specific sources of instructional strategies for each of the five dimensions of reading instruction.

The IPRI was designed to be used by field observers with a range of reading-related expertise, and therefore it was deliberately constructed with lower-inference and more behaviorally specific items. The lower-inference items represent discrete behaviors that the research cited in the National Reading Panel report (National Institute of Child Health and Human Development Report, 2000) suggests are important for improving elements of reading instruction, and the behaviors that are hypothesized to differ between RF and non-RF classrooms. Classroom observers had to demonstrate mastery of the IPRI over the course of an intensive week-long training session before they were hired to conduct observations. Mastery was measured by comparing observers' and master trainers' ratings of classroom instruction.[42]

The RFIS team created eight measures of classroom instruction from the IPRI data, which, taken together, represent the essential components of reading instruction emphasized by the Reading First program. This number is deliberately constrained to focus only on the most pivotal aspects of the program and to limit the number of statistical tests required. The instructional measures include:

- ***Minutes of Instruction in the Five Dimensions Combined.*** This reflects the number of minutes of instruction summed across the five dimensions of reading instruction: phonemic awareness, decoding, vocabulary, fluency, and comprehension.
- ***Minutes of Instruction in Each of the Five Dimensions***. These five measures reflect the number of minutes of instruction in each of the five dimensions separately: phonemic awareness, decoding, vocabulary, fluency, and comprehension.
- ***Percentage of Instructional Intervals in the Five Dimensions with Highly Explicit Instruction.*** This measures "highly explicit instruction" during lessons in the five dimensions. Instruction was considered "highly explicit" if teachers actively taught, modeled, explained, or assisted children in using specific reading strategies. The specific instructional activities comprising "highly explicit instruction" vary across the five dimensions, based on current research on reading instruction. Note that (1) this measure is based only on instruction in four of five dimensions (all except fluency building), and (2) the observations do not record highly explicit instruction in other literacy activities, such as spelling or writing.
- ***Percentage of Instructional Intervals in the Five Dimensions with High Quality Student Practice.*** This measures "high quality student practice," which reflects teachers' provision of dimension-specific practice opportunities, as based on current research. Note that this measure is based only on instruction in the five dimensions; the observations do not record high quality student practice in other literacy activities, such as spelling or writing.

To create the six analytic variables about time spent in the dimensions of reading instruction, data from classroom observations of instruction were transformed from intervals into minutes. In cases where one instructional behavior/activity was observed, that interval was designated accordingly. In

[42] The inter-rater reliability for the IPRI has been calculated using overall percent agreement, occurrence percent agreement, non occurrence agreement, and generalizability coefficients, all of which yield consistent results. In fall 2005, raters demonstrated overall agreement and non-occurrence agreement of 90 percent or above, and occurrence agreement of approximately 70 percent on average. The least reliable items were those that occurred infrequently. For a more complete discussion of inter-rater reliability, see Appendix C.

cases where multiple instructional behaviors were observed during one three-minute interval, the minutes were distributed across the specific instructional behaviors that had been observed. (See Appendix C for a more detailed discussion of the transformation of intervals into minutes.) To create the last two analytic variables, the data from classroom observations were summed across all the individual three-minute intervals within an observation. The total number of intervals (within each observation) with highly explicit instruction and high quality student practice was then divided by the total number of intervals (within each observation) with instruction in the five dimensions of reading.

The IPRI data are used to describe the content of instruction as well as the use of pedagogical strategies hypothesized to improve students' reading skills. The eight specific outcome measures used in analysis correspond to the amount of instructional time allocated to the each of the five dimensions of reading instruction described above, as well as one outcome representing all five dimensions combined, and one outcome each for the proportion of instructional time allocated to highly explicit instruction and provision of high quality student practice.

## Student Time-on-Task and Engagement with Print

The Reading First program legislation explicitly articulates a number of goals related to professional development, use of research-based materials and assessments, classroom reading instruction, and students' reading performance (No Child Left Behind Act, 2001). The Guidance for the Reading First Program (U.S. Department of Education, 2002) indicates that Reading First classrooms should also be characterized by "active student engagement in a variety of reading-based activities" and "high levels of time on task." There is research indicating that students benefit from more time on reading-related tasks and from instruction that is structured to provide more time on task (see for example, Snow, Burns, and Griffin 1998; Taylor et al, 1999). The RFIS observational instrument, the IPRI, focuses primarily on teacher behaviors, and in order to ensure that the study also collected some data on student behavior during observed reading instruction, the RFIS developed a measure that captures information about students' time on task and attention to printed material.

Student behavior during reading instruction was assessed through structured observations using the Student Time-on-Task and Engagement with Print (STEP) instrument.[43] The STEP is designed to record student engagement in instruction and students' exposure to print materials. Specifically, it is designed to capture the percentage of students in a classroom engaged in productive academic work (i.e., "on task"), and, of those, the percentage who are engaged in either reading or writing print.

The STEP is completed by a separate STEP observer during ongoing observations of reading instruction by an IPRI observer. The STEP observer records a time-sampled "snapshot" of student engagement three times in each classroom, e.g., three "sweeps" during the designated reading block in each classroom. Six minutes after entering the classroom during ongoing reading instruction, the STEP observer begins collecting the first of these sweeps. During each sweep, which lasts for approximately three minutes, the observer classifies every student in the classroom as either on- or off-task, and, if on-task, whether the student is: 1) reading connected text (a story or passage); 2) reading isolated text (letters, words, or isolated sentences); and/or 3) writing. The STEP observer waits until six minutes have elapsed between the end of one sweep and the start of the next. After the

[43] See Appendix C for a copy of the STEP, as well as more information on data collection, response rates, and inter-rater reliability.

third and final sweep, the STEP observer leaves the classroom. The STEP observer typically completes STEP observations in three classrooms spending about 25-30 minutes in each classroom. Data collected with the STEP measure are used to create one outcome representing the average percentage of students engaged with print during the designated reading block.

It is important to note that the theory of action for Reading First does not specify whether students' time on task and engagement with print during regular reading instruction would increase or decrease as a result of Reading First. One could hypothesize that well-implemented Reading First classrooms would increase both students' time-on-task and engagement with print, because teachers would manage time effectively and ensure that students' assignments are matched to their reading skills, whether those tasks are carried out in whole class, small group, or other grouping arrangements. One could also hypothesize that younger students would spend more time attending to the teacher than focusing directly on print, because they are not yet proficient enough readers to read independently, in which case Reading First could lead to decreases in the percentage of students engaged with print.

Chapter Four presents findings on all three of the outcome domains.

# Chapter Four: Impact Findings

This chapter presents findings on Reading First's impact on students' reading comprehension, teachers' reading instruction, and student engagement with print during reading instruction. It begins with a discussion of the program's overall impacts across the 18 study sites, and then explores variation in impacts among the 18 sites. It also explores alternative approaches to weighting that might influence study findings because of potential site-by-site variation. Finally, it assesses the impacts for the two groups of sites the study team had hypothesized would differ based on the length of time they had access to Reading First funding during the study's follow-up period. The findings in the chapter are based on data collected during the 2004-2005 and 2005-2006 school years, which represent between one and three years of Reading First funding across the sites.

The key findings include the following:

- On average, across the study sites, estimated impacts on student reading test scores were not statistically significant.
- For teachers in grades one and two, Reading First produced positive and statistically significant increases in the total time spent on the five dimensions of reading instruction. For first grade teachers, these impacts were concentrated in phonemic awareness and phonics. For second grade teachers, these impacts were concentrated in phonics, vocabulary, and comprehension.
- Impacts on the percentage of students engaged with print were mixed. For second grade classrooms, Reading First produced a statistically significant reduction in the percentage of students engaged with print. For first grade classrooms, the estimated impact on the percentage of students engaged with print was not statistically significant.
- The overall variation in impacts among the 18 sites was not statistically significant. Estimated impacts varied by more than one standard deviation on reading comprehension test scores, and by more than two standard deviations on the instructional time teachers spent in the five dimensions of reading instruction.
- Study sites that received their Reading First grants later in the federal funding process (between January and August 2004) experienced positive and statistically significant impacts both on the time first and second grade teachers spent on the five essential components of reading instruction, and on first and second grade student reading comprehension. Time spent on the five essential components was not assessed for third grade, and impacts on third grade reading comprehension were not statistically significant. In contrast, there were no statistically significant impacts on either time spent on the five components of reading instruction or on reading comprehension scores at any grade level among study sites that received their Reading First grants earlier in the federal funding process (between April and December 2003).
- Although there are multiple differences between the sites that received awards earlier and later, there is no way to distinguish which mix of these or other unmeasured factors explains the differences in the observed patterns of estimated impacts.

As described in Chapter 2, all impact estimates are regression-adjusted to control for a linear specification of the rating variable each site used to select its Reading First schools as well as selected

teacher and /or student background characteristics used in the analysis.[44] The impacts have been estimated using multi-level models to account for the clustering of students within classrooms, classrooms within schools, and schools within sites. In the exhibits that follow, values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

## Average Impacts for the Study Sites

This section presents estimates of the average impacts of Reading First on student reading comprehension, classroom reading instruction, and student engagement with print for the 18 study sites. The impact estimates are based on two school years: 2004-2005 and 2005-2006, and they pool results from students, teachers, and classrooms across the two school years. The study pools estimates both to improve statistical power and to be more parsimonious with respect to findings. The differences in impacts between the two years are not statistically significant for data collected in both years.[45,46] (Appendix D presents impact estimates separately for each follow-up year.)

### Reading Comprehension

Impacts on reading comprehension are based on student scores on the Stanford Achievement Test, 10th Edition (SAT 10). The analysis used both a continuous measure and a dichotomous measure of student scores. The continuous measure was mean student scaled score. To facilitate interpretation of the average scaled score, Exhibit 4.1 also includes the grade equivalent and national percentile, which corresponds to the averages for schools with Reading First and estimated averages of how these schools would have performed in the absence of Reading First, respectively. Impacts were not estimated for grade equivalents or national percentile ranks because these metrics are not equal-interval measures and should not be used in arithmetic calculations. The dichotomous measure was the percentage of students who scored at or above grade level.

Exhibit 4.1, Panel 1, presents estimates of the overall impacts of Reading First on mean reading comprehension scores for all study sites during spring 2005 and spring 2006, separately for students in grades one, two, and three. Specifically:

- The impact on reading comprehension in first grade was not statistically significant. The average scaled score for first grade students in schools with Reading First was estimated to be 3.6 points higher than their scores would have been without Reading First. This is equivalent to a mean effect size of 0.07 standard deviations.

---

[44] See Appendix B for a description of the background characteristics used in the estimation of impacts.

[45] P-values for reading comprehension outcomes range across grades from 0.472 to 0.910, and range from 0.669 to 0.940 for outcomes in the instruction domain.

[46] To account for possible modeling differences associated with the year of data collection, impact estimation models include indicator variables for each data collection period and interactions between these and all other covariates. The indicator variables account for year-to-year variation in the levels of the outcome measures as well as in the relationship between covariates and outcome measures.

## Exhibit 4.1: Estimated Impacts on Student Achievement: Spring 2005 and 2006[1]

| | Actual Mean with Reading First | Estimated Mean without Reading First | Impact | Effect Size of Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|---|
| ***Panel 1*** | | | | | |
| **All Sites** | | | | | |
| Reading Comprehension | | | | | |
| Grade 1 | | | | | |
| Scaled Score | 543.1 | 539.6 | 3.57 | 0.07 | (0.215) |
| *Corresponding Grade Equivalent* | *1.7* | *1.7* | | | |
| *Corresponding Percentile* | *44* | *41* | | | |
| Grade 2 | | | | | |
| Scaled Score | 584.3 | 582.9 | 1.41[a] | 0.03 | (0.559) |
| *Corresponding Grade Equivalent* | *2.5* | *2.4* | | | |
| *Corresponding Percentile* | *39* | *38* | | | |
| Grade 3 | | | | | |
| Scaled Score | 608.4 | 610.0 | -1.63 | -0.04 | (0.455) |
| *Corresponding Grade Equivalent* | *3.3* | *3.3* | | | |
| *Corresponding Percentile* | *39* | *39* | | | |
| ***Panel 2*** | | | | | |
| **All Sites** | | | | | |
| Percent Reading At or Above Grade Level | | | | | |
| Grade 1 | 45.4 | 42.2 | 3.15 | N/A[2] | (0.260) |
| Grade 2 | 38.9 | 38.8 | 0.12 | N/A | (0.965) |
| Grade 3 | 37.9 | 40.1 | -2.22 | N/A | (0.383) |

**Notes:**

The complete Reading First Impact Study (RFIS) sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test scores were not available.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005 and 2006 SAT-10 test scores (by grade).

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

[1] 95 percent confidence intervals for main impacts can be found in Appendix E.

[2] The "at or above grade level" variable is dichotomous; therefore effect sizes are not appropriate.

[a] Due to estimation variation and rounding, the estimated pooled sample impact can be slightly larger than for 2005 and 2006 separately.

**EXHIBIT READS: The observed mean reading comprehension score for first-graders with Reading First was 543.1 scaled score points. The estimated mean without Reading First was 539.6 scaled score points. The impact of Reading First was 3.6 scaled score points (or 0.07 standard deviations), which was not statistically significant at the $p \leq .05$ level (p=0.215). The observed average percent of first-graders reading at or above grade level with Reading First was 45.4 percentage points. The estimated average percent without Reading First was 42.2 percentage points. The impact of Reading First on the percent of first grade students reading at or above grade level was 3.2 percentage points, which was not statistically significant at the $p \leq .05$ level (p=.260).**

*Sources: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already used the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR).*

The average first grade score with or without Reading First was equivalent to the seventh month (of a nine-month school year) of first grade, based on national norms. The corresponding national percentile ranks for the scaled score means were 44 and 41, respectively.

- The impact on reading comprehension in second grade was not statistically significant. The average scaled score for second grade students in schools with Reading First was estimated to be 1.4 points higher than their scores would have been without Reading First, which is equivalent to an effect size of 0.03 standard deviations.

Average second grade scores with or without Reading First were equivalent to the fifth month and fourth month of second grade, respectively. The corresponding percentile ranks were 39 with Reading First and 38 in the absence of the program.

- The impact on reading comprehension in third grade was not statistically significant. The average scaled score for third grade students in schools with Reading First was estimated to be 1.6 points below what their scores would have been without Reading First. This is equivalent to an effect size of -0.04 standard deviations.

The average score with or without Reading First was equivalent to the third month of third grade, and the percentile rank was 39 in both cases.

Exhibit 4.1, Panel 2, reports estimates of the impacts of Reading First on the percentage of students who scored at or above grade level in reading comprehension. Grade level was defined as the grade equivalent score that matches the grade and month in which a student was tested. Thus, for example, students tested in the seventh month of second grade were judged to read at or above grade level if the grade equivalent of their scaled score was 2.7 or higher. Findings indicate that:

- Estimated impacts on the percentage of students reading at or above grade level for grades one, two, and three were not statistically significant.

Panel 2 in Exhibit 4.1 indicates that, on average, across all three grade levels, fewer than half of the students in schools with Reading First scored at or above grade level.

Exhibit 4.1 includes six statistical tests of program impacts on reading comprehension—one for each combination of grade and reading comprehension measure. A composite test of these estimates (using an index that combines measures and pools the sample across grades) was not statistically significant. The estimated effect size of the impact of Reading First on the composite reading comprehension index was 0.02 standard deviations and its p-value was 0.668.

Exhibit 4.2 presents these findings in visual terms, using effect sizes to display the impact estimates as well as the 95 percent confidence intervals.[47] The exhibit displays reading comprehension impact estimates as well as instructional outcome impact estimates. Because instructional data were collected in grades one and two only, the bottom panel includes only an impact estimate for reading comprehension. The exhibit presents separate graphs for each of the three grades. Each graph plots the estimated mean impact, represented by a small square, and the 95 percent confidence interval for

[47] See Appendix E for 95 percent confidence intervals for main impact estimates in relevant metrics.

each estimate, represented by a line extending outward from the mean. The confidence intervals indicate the margin of error for each estimate; the wider the confidence interval, the broader the margin of error, and the more uncertainty about the estimate. If a 95 percent confident interval does not include zero, the estimated impact was statistically significant (p-value less than or equal to 0.05). The display indicates that the impact estimates and associated confidence intervals for reading comprehension are close to or cover zero.

### Reading Instruction

Measures of reading instructional practice for grades one and two are based on classroom observations conducted by trained observers. Limitations of resources precluded such observations for grade three. The impacts on classroom instruction are based upon continuous measures of the amount of instructional time teachers spent on the five dimensions of reading instruction (for all five dimensions combined and separately) as well as measures of the proportion of observational intervals that included highly explicit instruction and high quality student practice.

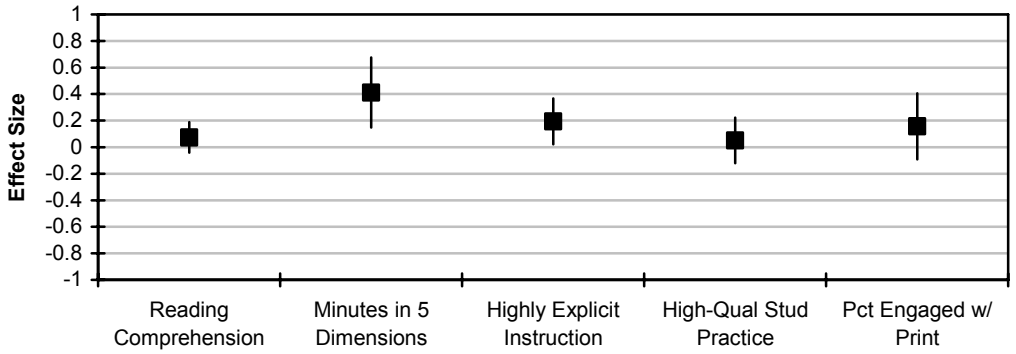
Exhibit 4.3 summarizes resulting estimates of the impacts of Reading First on instructional practices. The top panel in the exhibit presents estimates of program impacts on the average number of minutes per day spent on the five dimensions of reading instruction combined (phonemic awareness, phonics, vocabulary, fluency, and comprehension).

- For first grade classrooms, Reading First produced an increase of 8.6 minutes per daily reading block, which is statistically significant; this is equivalent to an effect size of 0.41 standard deviations. This impact represents roughly 45 minutes of additional instruction in the five dimensions per week.
- For second grade classrooms, Reading First produced an additional 12.1 minutes per daily reading block. This impact estimate is equivalent to an effect size of 0.57 standard deviations, and it is statistically significant. This represents about 60 minutes of additional instruction in the five dimensions per week.

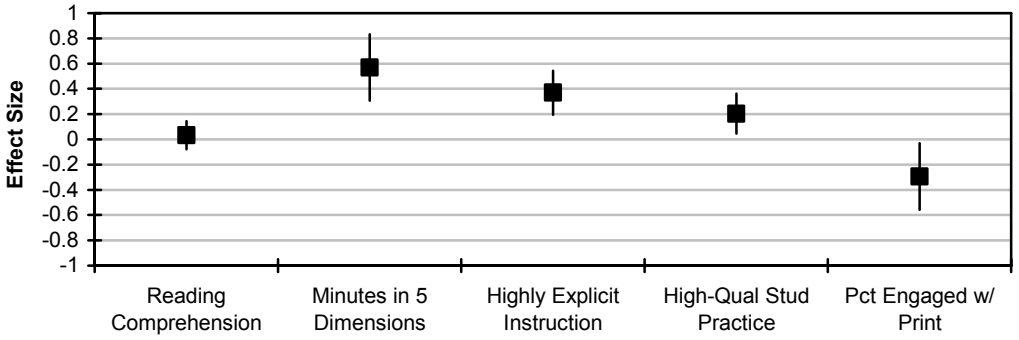
The bottom panel of Exhibit 4.3 presents estimates of Reading First impacts on two other instructional outcomes. One represents the percentage of three-minute classroom observation intervals in which teachers used highly explicit instructional strategies associated with the five dimensions. The second outcome captures the percentage of three-minute classroom observation

## Exhibit 4.2: Estimated Impacts and Confidence Intervals for Key Outcomes, in Effect Size, by Grade

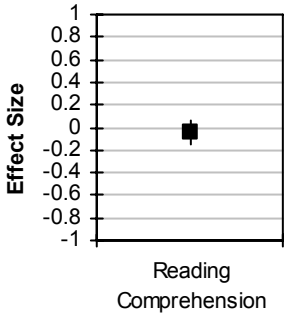
**Grade 1**




**Grade 2**




**Grade 3**




**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

The outcome measure depicted for reading comprehension is the SAT 10 scaled score.

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005 and 2006 SAT-10 test scores; spring 2005, fall 2005, and spring 2006 IPRI data and fall 2005 and spring 2006 STEP data (by grade).

**For each outcome and grade level, impact estimates and 95 percent confidence intervals are presented in effect size terms. (See Exhibits 4.1, 4.4, and 4.6 for actual impact estimates.)**

*Source: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006; and RFIS Student Time-on-Task and Engagement with Print, fall 2005 and spring 2006.*

**Exhibit 4.3: Estimated Impacts on Instructional Outcomes: Spring 2005, Fall 2005, and Spring 2006[1]**

| | Actual Mean with Reading First | Estimated Mean without Reading First | Impact | Effect Size of Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|---|
| ***Panel 1*** | | | | | |
| **Number of minutes of instruction in the five dimensions combined** | | | | | |
| Grade 1 | 59.41 | 50.85 | 8.56* | 0.41* | (0.003) |
| Grade 2 | 59.53 | 47.44 | 12.09* | 0.57* | (<0.001) |
| ***Panel 2*** | | | | | |
| **Percentage of intervals in five dimensions with Highly Explicit Instruction** | | | | | |
| Grade 1 | 29.78 | 26.13 | 3.65* | 0.20* | (0.023) |
| Grade 2 | 31.55 | 24.57 | 6.98* | 0.36* | (<0.001) |
| **High Quality Student Practice** | | | | | |
| Grade 1 | 19.21 | 18.35 | 0.86 | 0.05 | (0.559) |
| Grade 2 | 18.78 | 15.11 | 3.67* | 0.20* | (0.012) |

**Notes:**

The complete RFIS sample includes 248 schools from 18 sites (17 districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005, fall 2005, and spring 2006 IPRI data (by grade).

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

[1] 95 percent confidence intervals for main impacts can be found in Appendix E.

**EXHIBIT READS: The observed mean amount of time spent in instruction in the five dimensions (phonemic awareness, phonics, vocabulary, fluency, and comprehension) in first grade classrooms with Reading First was 59.41 minutes. The estimated mean amount of time without Reading First was 50.85 minutes. The impact of Reading First on the amount of time spent in instruction in the five dimensions was 8.56 minutes (or 0.41 standard deviations), which was statistically significant at the $p \leq .05$ level (p=.003).**

*Sources: RFIS, Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006*

intervals in which students were provided with high quality practice opportunities focused on skills within the five dimensions. These findings include the following:

- Reading First increased the incidence of highly explicit instruction by 3.65 percentage points for grade one, and by 6.98 percentage points for grade two, corresponding to effect sizes of 0.20 and 0.36, respectively. Both estimates are statistically significant.
- The impact of Reading First on high quality student practice is statistically significant in grade two but not in grade one. In grade two, Reading First increased the incidence of high quality student practice by 3.67 percentage points, corresponding to an effect size of 0.20. In grade one, Reading First increased the incidence of high quality student practice by 0.86 percentage points, corresponding to an effect size of 0.05.

Exhibit 4.2 (above) graphs the impact estimates and 95 percent confidence intervals for the instructional outcomes, by grade. The exhibit indicates positive and statistically significant impacts for two of the three instructional outcomes in Grade 1 and all three instructional outcomes in Grade 2.

As was the case for the reading comprehension impact estimates, a composite test of the six impact estimates in Exhibit 4.3 was conducted using an index consisting of the average of the three instructional outcomes and pooling the sample across grades. The composite test indicates a statistically significant overall impact of Reading First on instructional practice. The estimated effect size of the impact of Reading First on the composite index of reading instruction is 0.44 standard deviations and its p-value was less than 0.0001. This composite test also holds for the findings presented next in Exhibit 4.4, because they represent subdivisions of the preceding results.

Exhibit 4.4 presents separate estimates for each of the five dimensions of reading instruction. These findings illustrate the relative emphasis placed by Reading First schools on each dimension, how this emphasis differs by grade, and how the impacts of Reading First are distributed across the five dimensions. The majority of instructional time spent by Reading First teachers was focused on comprehension and phonics.

- In first grade classrooms, the impact on phonics was statistically significant, while the impact on comprehension was not statistically significant. First grade classroom instruction in schools with Reading First included about 21.4 minutes on phonics and about 23.6 minutes on comprehension per daily reading block. This reflects an estimated daily impact of 3.9 additional minutes for phonics and 2.3 more minutes for comprehension.
- Second grade classroom instruction in schools with Reading First included about 29.2 minutes per daily reading block on comprehension, and about 14.0 minutes on phonics. This reflects an estimated daily impact of 5.3 extra minutes for comprehension and 3.9 extra minutes for phonics, both of which were statistically significant.

Classroom instruction in both first and second grade in schools with Reading First included less time per daily reading block on other dimensions of reading than on comprehension and phonics, as follows: vocabulary (7.8 and 11.6 minutes, respectively), fluency (4.5 and 4.3 minutes, respectively), and phonemic awareness (2.1 and 0.4 minutes, respectively). Impacts on phonemic awareness in grade one and on vocabulary in grade two were statistically significant.

**Exhibit 4.4: Estimated Impacts on Instructional Outcomes: Spring 2005, Fall 2005, and Spring 2006[1]**

| | Actual Mean with Reading First | Estimated Mean without Reading First | Impact (minutes) | Effect Size of Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|---|
| ***Number of minutes of instruction in:*** | | | | | |
| **Phonemic Awareness** | | | | | |
| Grade 1 | 2.07 | 1.35 | 0.72* | 0.27* | (0.016) |
| Grade 2 | 0.42 | 0.28 | 0.15 | 0.12 | (0.167) |
| **Phonics** | | | | | |
| Grade 1 | 21.36 | 17.46 | 3.90* | 0.29* | (0.015) |
| Grade 2 | 14.01 | 10.16 | 3.85* | 0.36* | (0.004) |
| **Vocabulary** | | | | | |
| Grade 1 | 7.80 | 7.16 | 0.65 | 0.10 | (0.378) |
| Grade 2 | 11.63 | 9.49 | 2.14* | 0.25* | (0.031) |
| **Fluency** | | | | | |
| Grade 1 | 4.54 | 3.46 | 1.09 | 0.18 | (0.112) |
| Grade 2 | 4.25 | 3.60 | 0.65 | 0.12 | (0.287) |
| **Comprehension** | | | | | |
| Grade 1 | 23.63 | 21.35 | 2.29 | 0.16 | (0.204) |
| Grade 2 | 29.22 | 23.96 | 5.26* | 0.32* | (0.008) |

**Notes:**

The complete RFIS sample includes 248 schools from 18 sites (17 districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005, fall 2005, and spring 2006 IPRI data (by grade).

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

[1] 95 percent confidence intervals for main impacts can be found in Appendix E.

**EXHIBIT READS: The observed mean amount of time spent per daily reading block in instruction in phonemic awareness for first grade classrooms with Reading First was 2.07 minutes. The estimated mean amount of time without Reading First was 1.35 minutes. The impact of Reading First on the amount of time spent in instruction in phonemic awareness was 0.72 minutes (or 0.27 standard deviations), which was statistically significant at the $p \leq .05$ level (p=.016).**

*Sources: RFIS, Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006*

### Student Engagement with Print

Measures of student engagement with print were obtained from direct observation of classrooms by trained observers. The measure of student engagement used in impact analyses is the per-classroom average of the percentage of students engaged with print across three sweeps in each classroom.

Estimates of the impacts of Reading First on this outcome are presented in Exhibit 4.5. Findings in the exhibit indicate that at particular points in time during the observation, about half of the first and second grade students in schools with Reading First were engaged with print (46.9 percent of first-graders and 49.7 percent of second-graders, on average). For second grade in schools with Reading First, this represents a statistically significant decrease of 8.4 percentage points, relative to what is estimated to occur without Reading First. For first grade, this represents an impact that was not statistically significant. The percentage of students engaged with print was 4.6 points greater for schools with Reading First relative to what was estimated to occur without Reading First. Referring back to Exhibit 4.2, the impact estimates and their 95 percent confidence intervals are displayed visually, in effect size terms, for student engagement with print.

As with other outcomes, a composite test was conducted that pools findings across grades; it was not statistically significant. The estimated effect size of the impact of Reading First on the index of percentage of students engaged with print is 0.07 standard deviations and its p-value is 0.710. The statistically significant impact for second grade classrooms should therefore be interpreted with caution.

**Exhibit 4.5: Estimated Impacts on the Percentage of Students Engaged with Print: Fall 2005 and Spring 2006[1]**

| Construct | Actual Mean with Reading First | Estimated Mean without Reading First | Impact (percentage points) | Effect Size of Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|---|
| **Percentage of students engaged with print** | | | | | |
| Grade 1 | 46.92 | 42.29 | 4.63 | 0.16 | (0.216) |
| Grade 2 | 49.72 | 58.14 | -8.42* | -0.29* | (0.030) |

**Notes:**

The complete RFIS sample includes 248 schools from 18 sites (17 districts and one state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the fall 2005 and spring 2006 STEP data (by grade).

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

[1] 95 percent confidence intervals for main impacts can be found in Appendix E.

**EXHIBIT READS: The observed average percentage of students engaged with print in first grade classrooms with Reading First was 46.92 percent. The estimated average percentage without Reading First was 42.29 percent. The impact of Reading First on the average percentage of student engagement with print was 4.63 percentage points (or 0.16 standard deviations), which was not statistically significant at the $p \leq .05$ level (p=.216).**

*Source: RFIS, Student Time-on-Task and Engagement with Print, fall 2005 and spring 2006.*

## Variation in Impacts Across Sites

As discussed in Chapter 2, the impacts presented above reflect an average aggregated across the 18 study sites. To the degree that there is variation in impacts across the sites, the overall average may be masking important differences in the effectiveness (or lack of effectiveness) of Reading First under some conditions. For example, the participating sites differ in terms of the amount of Reading First program funds allocated per school or student as well as when they could first access Reading First grant funding. While the study is not designed to establish causal relationships between differences across sites in Reading First impacts and differences in site characteristics, an assessment of variation provides a context for interpreting the overall average impacts.

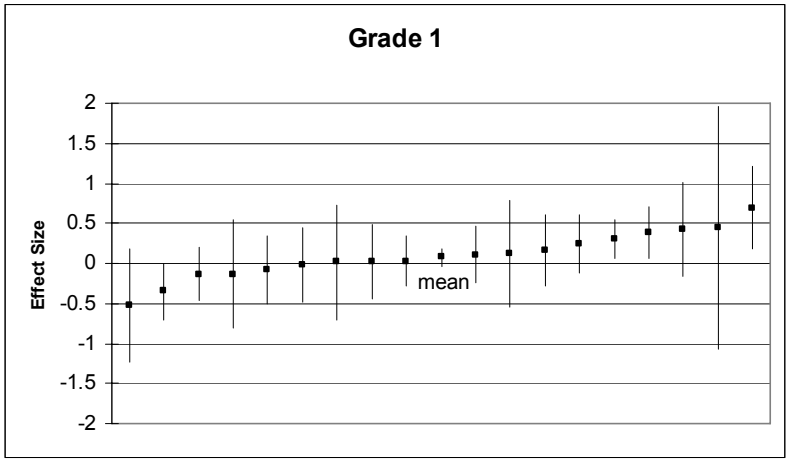
### Variation in Impacts on Reading Comprehension

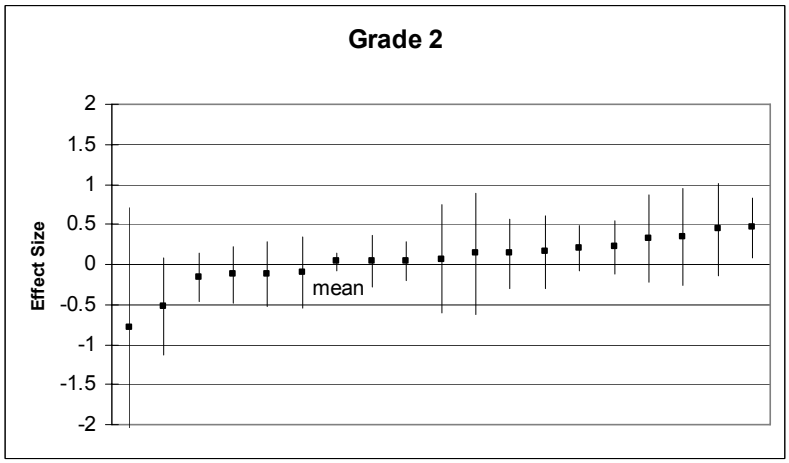
Exhibit 4.6 illustrates the variation across sites of estimated program impacts on reading comprehension scaled scores.[48] This exhibit presents separate graphs for each grade, and it displays mean impact estimates and 95 percent confidence intervals for each site. Here, too, the wider the confidence interval, the broader the margin of error and the greater the uncertainty about the estimate. For first grade, for example, the site-by-site estimates range from a decrease of 0.5 standard deviations to an increase of 0.7 standard deviations; 13 estimates are positive and five are negative. On balance, for grade one, confidence intervals for all negative impact estimates and all but three positive impact estimates include zero. Note that the RFIS was not designed to be able to detect differences at the site level.

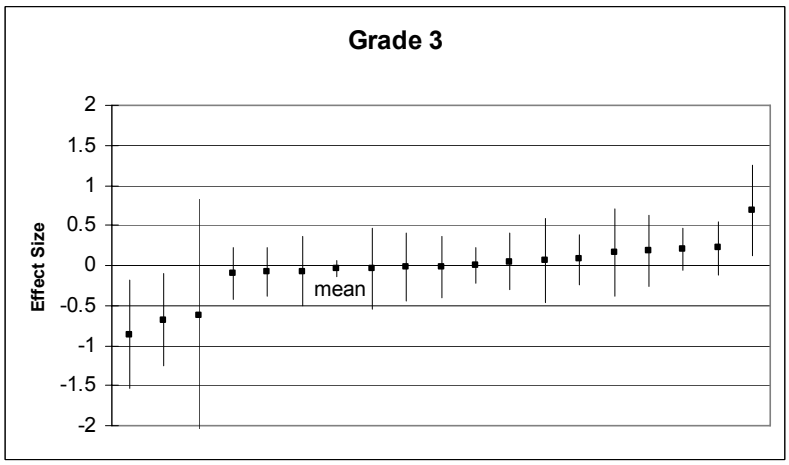
To examine cross-site variability in impacts more systematically, a composite F-test was used to assess the null hypothesis that there were no statistically significant differences across the site-level impacts on reading comprehension test scores. This test was conducted for each grade separately and then with all grades pooled together (see Exhibit 4.7). The exhibit shows that the p-value for the grade one F-test was 0.06; the null hypothesis cannot be rejected and thus, site-to-site variation was not statistically significant, and cannot be distinguished from zero reliably. The statistical tests of site-to-site variation in impacts on reading comprehension test scores for grades two and three follow a similar pattern. For all three grades, the estimated variation in impacts on test scores across sites was not systematically different from the variations that could occur by chance. Even though the observed variation encompasses more than one full standard deviation of the reading test score measure, the variation was not statistically significant. The lack of significance is not surprising, given the limited statistical power for estimating variation across sites, due to the small number of sites (18) and to the weak precision of impact estimates by site (an average of 14 schools per site). As a result, it is not possible to determine the true extent to which program impacts vary across study sites with confidence.

[48] Each grade-specific graph presents impact estimates in numerical (ascending) order; therefore each graph (by grade and by outcome) presents sites in a different order.

## Exhibit 4.6: Fixed Effect Impact Estimates on Reading Comprehension, by Site, by Grade




**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005 and 2006 SAT-10 test scores (by grade).

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

*Source: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR).*

**Exhibit 4.7: Results of Composite F-Test for Variation in Site Level Impacts**

| Outcome | p-value |
|---|---|
| Reading Comprehension Scaled Score | |
| Grade 1 | (0.063) |
| Grade 2 | (0.294) |
| Grade 3 | (0.102) |
| All Grades | (0.096) |
| Minutes in Five Dimensions | |
| Grade 1 | (0.518) |
| Grade 2 | (0.129) |
| All Grades | (0.244) |
| Percentage of Student Engagement with Print | |
| Grade 1 | (0.007)* |
| Grade 2 | (0.212) |
| All Grades | (0.009)* |

**Notes:**

The complete RFIS sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

**EXHIBIT READS: The p-value for the joint F-test that tests whether the program impact is the same across all sites for first grade reading comprehension is 0.063, which is not statistically significant at the $p \leq .05$ level.**

*Sources: RFIS SAT 10 administrations in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already used the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006; and RFIS Student Time-on-Task and Engagement with Print, fall 2005 and spring 2006.*

### Variation in Impacts on Reading Instruction

The site-by-site variation in estimated program impacts on minutes of instruction in the five dimensions is illustrated in graphs similar to those shown in Exhibit 4.6 (see Appendix F). For first grade, for example, the site-by-site estimates range from a decrease of 17 minutes (an effect size of -0.81 standard deviations) to an increase of 27 minutes (an effect size of 1.29 standard deviations) per daily reading block; three estimates were negative and 15 were positive.

The middle rows in Exhibit 4.7 show the results of statistical tests of the site-to-site variation in impacts on instructional time in the five dimensions of reading instruction. The p-values of the F-tests (0.52 for grade one and 0.13 for grade two) indicate that the variation in estimated impacts for grade one and grade two was not statistically significant, even though the observed differences among impact estimates across sites covers more than two standard deviations of the instructional time measure. The lack of statistical significance is due to lack of statistical power for estimating cross-site variation in impacts on instructional behaviors, as was the case for estimating cross-site variation in impacts on reading comprehension, noted earlier.

### Variation in Impacts on Student Engagement with Print

Appendix F also presents graphs that are similar to Exhibit 4.6 to illustrate the site-by-site variation of estimated program impacts on percentage of students engaged with print. For second grade, for

example, the estimates range from a decrease of 71 percentage points to an increase of nearly 28 percentage points; 12 estimates are negative and six are positive.

Corresponding findings in the third set of numbers in Exhibit 4.7 show that the p-values for these F-tests of cross-site variation in impacts on student engagement with print were 0.01 for grade one and 0.21 for grade two. This suggests that the variation for grade one was statistically significant while it was not for grade two. When samples were combined across grades one and two, the test for site-to-site variation in impacts was also statistically significant.

## Alternative Approaches to Weighting: Implications of Variation in Impacts Across Sites

To the extent that overall average impacts vary across sites, alternative approaches to weighting can yield different results. Recall that this study is using a weighting strategy that weights each site estimate in proportion to the number of RF schools in that site; this approach yields impact estimates for the average RF school in the study sample. To gauge the sensitivity of the impacts to weighting, the average impacts were re-estimated using two weighting strategies that had initially been considered for the study. One alternative is to weight site-specific impact estimates in proportion to each site's number of Reading First students (rather than its number of Reading First schools), which produces impact estimates for the average Reading First student in the study sample.

The second alternative is to specify one treatment indicator for all sites, instead of specifying site-specific treatment indicators and then averaging their coefficients. This is called a *pooled* estimator rather than a weighted estimator, because it pools data for the full sample directly into a single average impact estimate. It should be noted, however, that the pooled estimator, like any other, represents a weighting of impact estimates across sites. The implicit weights for this strategy were approximately proportional to the precision of impact estimates for each site, which in turn reflect the site's sample size and study design.[49]

Appendix B compares estimates of the average impacts of Reading First produced by the weighting strategy used in this study and two other alternative approaches to weighting. Results are presented for estimates of impacts on reading comprehension, instruction in the five dimensions, and percentage of students engaged with print.

For reading comprehension (in effect size terms), the alternative estimates range from 0.03 to 0.11 standard deviations for grade one, from 0.00 to 0.07 standard deviations for grade two, and from -0.04 to 0.02 standard deviations for grade three. Estimates using the weighting strategy chosen for this study were generally between those for the other two strategies. Only the pooled estimate for grade one was statistically significant.

[49] This alternative strategy weights each site's impact estimate in proportion to its total amount of "free" (non-collinear) variation in treatment status across schools, which is the major factor that determines the precision of these estimates. For detailed explanation and an application of this approach for an experiment, see Cullen, Jacob and Levitt, 2006).

For instruction in the five dimensions, alternative estimates range from estimated increases of 8.52 minutes to 8.79 minutes per daily reading block for grade one and 11.75 minutes to 12.38 minutes per daily reading block for grade two. All of these estimates were statistically significant.

For student engagement with print, alternative estimates ranged from 3.39 to 4.63 percentage points for grade one (none of which were statistically significant) and from -5.82 to -8.42 percentage points for grade two (the larger two of which were statistically significant).

In summary, there is some fluctuation due to weighting approaches used to average findings across sites. The overall conclusions of the findings reported here would not change, however, as a function of the approach to weighting.

## Differences in Impacts by Length of Time That Reading First Funding Was Available

Study sites received their Reading First grants between April, 2003 and August, 2004[50] and the follow-up data available for this report encompass the 2004-2005 and 2005-2006 school years. Hence, the follow-up periods for this report represent different lengths of time during which sites (and schools within sites) had access to Reading First funds, and therefore had different amounts of time to use those funds to work with teachers and students. Prior research suggests that complex educational initiatives take time to implement fully and that program effectiveness may improve as the program matures (Aladjem et al., 2006; Bloom, 2001; Borman et al., 2003). Consequently, the study team hypothesized that Reading First implementation would mature over time, and that the impact of Reading First on teachers' classroom instruction and students' reading comprehension would increase. In addition, the longer Reading First funds were used within schools, the more likely it is that individual students would experience cumulative exposure to Reading First-funded activities across grades.

The study team recognized that an overall average impact estimate might mask differences in impacts, if the findings suggested that the amount of time Reading First funds had been available was related to differences in impacts. Schools that received Reading First funds in 2003, for example, could have had up to three full school years to implement Reading First activities by the end of 2005-2006, whereas schools funded in 2004 could have had up to two full years to implement Reading First-funded activities.

The study team sought to account for the variation in the length of time that sites had access to Reading First funds by designating two groups of sites: those for which funding was first made available between April and December 2003 (early award sites) and those whose funding became available between January and August 2004 (late award sites). There are 10 sites and 111 schools in

[50] Information about the public announcement of the grant awards was compiled by SEDL (2004). Information about when funds were available to sites was confirmed by telephone with state and district Reading First Coordinators. The study relies on the dates when sites first had access to their Reading First funding grants, because those signify when sites could access Reading First funds from their grants to purchase materials and to support professional development activities associated with the implementation of their reading programs. In some cases, the public announcement of the grant awards came several months earlier.

the early award group, and 8 sites and 137 schools in the late award group. As of May 2005 (the end of the first wave of data collection for the RFIS), early and late award sites had had Reading First funding available for an average of 22 and 13 months, respectively. As of May 2006 (the end of the second data collection period for the study), early and late award sites had had access to Reading First grants for an average of 34 and 25 months, respectively.

Analyses were conducted to examine the relationship between how long sites had had access to Reading First funding and observed impacts on instructional and achievement outcomes. Observed changes in impacts from the first to second year of RF funding can be reported for late award sites only, given the study's data collection schedule, which began in school year 2004-05 and continued through 2006-07. The early award sites received their first year of funding in the 2003-04 school year, when the study had not yet begun to collect data. Therefore, for the early award sites the study can observe changes in impacts from the second to the third year of funding only. The study will be able to report on changes from the second to third year of funding for late award sites in the final report, which will include data from 2006-07. Exhibit 4.8 summarizes the findings for these analyses by displaying the impacts for Implementation Years 1 and 2, which correspond to calendar years 2005 and 2006, for late award sites (Panel 1) and for Implementation Years 2 and 3 (or 2005 and 2006) for early award sites (Panel 2).[51]

None of the year-to-year differences in impacts was statistically significant for either the late award sites (Panel 1) or the early award sites (Panel 2). Thus, Reading First's impacts on student reading comprehension and teachers' instructional behaviors do not appear to have increased (or decreased) systematically over time as the sites gained more experience with the program.

Findings for late award sites indicate statistically significant and positive impacts on the percentage of Grade 1 students reading at or above grade level in Year 2 and the percentage of Grade 2 students reading at or above grade level in Year 1. Also, for late award sites, Reading First produced positive and statistically significant impacts on minutes of instruction in the five dimensions for Grades 1 and 2 in both Year 1 and Year 2. None of the estimated impacts for early award sites was statistically significant, although the direction of (not significant) estimated impacts on the percentage of students reading at or above grade level was negative for all three grades. Also, the (nonsignificant) estimated impacts on instruction in the five dimensions for early award sites were positive. On balance, the findings in Exhibit 4.8 do not support the hypothesis that program impacts increased with program maturity.

[51] This table does not include impacts on the percentage of students engaged with print because these data are available for one year only. Impacts for all outcomes by subgroup by year can be found in Appendix G.

**Exhibit 4.8: Estimated Impacts on Reading Comprehension and Minutes in the Five Dimensions, by Implementation Year, Calendar Year, and Award Status**

| | Implementation Year | | | | | |
|---|---|---|---|---|---|---|
| | Year 1 | | Year 2 | | Year 3 | |
| | Impact | (p-value) | Impact | (p-value) | Impact | (p-value) |
| ***Panel 1*** | | | | | | |
| **Late Award Sites** | **2005** | | **2006** | | **2007** | |
| Grade 1 | | | | | | |
| Percent reading at or above grade level (%) | 6.3 | (0.077) | 9.4* | (0.024) | N/A | N/A |
| Instruction in five dimensions (minutes) | 11.51* | (0.001) | 12.03* | (0.004) | N/A | N/A |
| Grade 2 | | | | | | |
| Percent reading at or above grade level (%) | 6.3* | (0.028) | 5.7 | (0.155) | N/A | N/A |
| Instruction in five dimensions (minutes) | 14.84* | (<0.001) | 16.11* | (<0.001) | N/A | N/A |
| Grade 3 | | | | | | |
| Percent reading at or above grade level (%) | 1.7 | (0.537) | 4.2 | (0.269) | N/A | N/A |
| Instruction in five dimensions (minutes) | N/A | N/A | N/A | N/A | N/A | N/A |
| ***Panel 2*** | | | | | | |
| **Early Award Sites** | **2004** | | **2005** | | **2006** | |
| Grade 1 | | | | | | |
| Percent reading at or above grade level (%) | N/A | N/A | -2.6 | (0.708) | -1.9 | (0.751) |
| Instruction in five dimensions (minutes) | N/A | N/A | 5.49 | (0.376) | 4.16 | (0.457) |
| Grade 2 | | | | | | |
| Percent reading at or above grade level (%) | N/A | N/A | -8.2 | (0.163) | -6.8 | (0.303) |
| Instruction in five dimensions (minutes) | N/A | N/A | 10.93 | (0.083) | 4.56 | (0.410) |
| Grade 3 | | | | | | |
| Percent reading at or above grade level (%) | N/A | N/A | -9.9 | (0.110) | -7.7 | (0.225) |
| Instruction in five dimensions (minutes) | N/A | N/A | N/A | N/A | N/A | N/A |

**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

Implementation year represents the number of years since sites received notice of their Reading First grants. For early award sites, this occurred in 2003, and Year 1, 2, and 3 refer to the 2003-2004, 2004-2005, and 2005-2006 school years, respectively. For late award sites, notification of funding occurred in 2004, and Years 1 and 2 refer to the 2004-2005 and 2005-2006 school years, respectively (data are available for the 2004-2005 and 2005-2006 school years only).

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

**EXHIBIT READS: The impact of Reading First on the percent of students reading at or above grade level in grade one, for late award sites, in implementation Year 1 and Calendar Year 2005, was 6.3 percentage points, which was not statistically significant at the $p \leq .05$ level (p=.077).**

*Sources: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); and RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006.*

The following sections examine the unexpected pattern of differences across groups of sites more systematically. First, impacts were estimated with data pooled across follow-up periods to increase precision. Next, the discussion describes other differences between the two groups of sites and explores whether impacts varied systematically with these differences.

### Differences in Impacts for Early and Late Award Sites

Exhibit 4.9 presents estimates of Reading First impacts on reading comprehension scores. None of the estimated impacts for early award sites were statistically significant; estimated impacts for the early award sites were negative— - 0.2, - 4.8, and - 7.0 scaled score points—equivalent to effect sizes of 0.00, - 0.11, and - 0.17 standard deviations, respectively. In contrast, estimates for late award sites were positive for all three grades (6.6, 6.1, and 2.4 scaled score points) and were statistically significant for grades one and two. These findings were equivalent to effect sizes of 0.13, 0.14, and 0.06 standard deviations, respectively. Exhibit 4.9 illustrates a similar pattern of findings for program impacts on the percentage of students reading at or above grade level.

Differences in impacts on reading comprehension test scores between early and late award sites were statistically significant for grades two and three, and not statistically significant for grade one (see bottom panel, Exhibit H.1). As with the full sample impact analysis, a composite test was conducted to assess the overall difference in program impacts on student reading comprehension by creating an index which combines scaled scores with indicators of student performance at or above grade level and which pools the data for all three grades. The test demonstrates that overall, Reading First produced a positive and statistically significant impact on reading test scores for the late award sites, and that the estimated impact for the early award sites was negative but not statistically significant. The test also indicates that the overall difference in impacts on test scores between the two groups of sites was statistically significant.

**Exhibit 4.9: Estimated Impacts on Reading Comprehension: Spring 2005 and 2006, by Award Status**

| | Actual Mean with Reading First | Estimated Mean without Reading First | Impact | Effect Size of Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|---|
| **Early Award Sites** | | | | | |
| Scaled Score | | | | | |
| Grade 1 | 546.7 | 547.0 | -0.22 | 0.00 | (0.966) |
| Grade 2 | 587.3 | 592.0 | -4.78 | -0.11 | (0.290) |
| Grade 3 | 612.2 | 619.1 | -6.98 | -0.17 | (0.101) |
| Percent Reading At or Above Grade Level | | | | | |
| Grade 1 | 48.0 | 50.4 | -2.34 | N/A | (0.665) |
| Grade 2 | 41.0 | 48.7 | -7.69 | N/A | (0.140) |
| Grade 3 | 41.8 | 50.9 | -9.04 | N/A | (0.081) |
| **Late Award Sites** | | | | | |
| Scaled Score | | | | | |
| Grade 1 | 540.3 | 533.7 | 6.58* | 0.13* | (0.039) |
| Grade 2 | 582.0 | 575.9 | 6.09* | 0.14* | (0.021) |
| Grade 3 | 605.5 | 603.0 | 2.43 | 0.06 | (0.283) |
| Percent Reading At or Above Grade Level | | | | | |
| Grade 1 | 43.3 | 35.8 | 7.55* | N/A | (0.011) |
| Grade 2 | 37.2 | 31.1 | 6.10* | N/A | (0.023) |
| Grade 3 | 34.8 | 31.8 | 2.97 | N/A | (0.245) |

**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available. There are 8 late award sites, with 137 schools, and 10 early award sites, with 111 schools.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005 and 2006 SAT-10 test scores (by grade).

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

**EXHIBIT READS: The observed mean reading comprehension score for first-graders with Reading First in the early award sites was 546.7 scaled score points. The estimated mean without Reading First was 547.0 scaled score points. The impact of Reading First was -0.2 scaled score points (or 0.00 standard deviations), which was not statistically significant at the $p \leq .05$ level ($p = .966$). The observed average percent reading at or above grade level for first-graders with Reading First in the early award sites was 48.0 percentage points. The estimated average percent without Reading First was 50.4 percentage points. The impact of Reading First on the percent of first grade students reading at grade level was -2.3 percentage points, which was not statistically significant at the $p \leq .05$ level ($p = .665$).**

*Sources: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR).*

Exhibit 4.10 presents estimates of Reading First impacts on classroom instruction. For early award sites, estimated impacts on the number of minutes of instruction in the five dimensions of reading instruction were not statistically significant. In Grade 2, Reading First increased the incidence of high quality student practice. For late award sites, the findings indicate that Reading First produced positive and statistically significant impacts on teachers' instructional behavior, increasing time in the five dimensions by 11.6 minutes per daily reading block for Grade 1 and 15.6 minutes for Grade 2. These results were equivalent to effect sizes of 0.56 and 0.74 standard deviations, respectively.

There was no clear pattern in Exhibit 4.10 for differences in program impacts in grades 1 and 2 across the two award groups on highly explicit instruction or high quality student practice. With one exception, the differences in impacts between early and late award sites were not statistically significant. In Grade 2, however, the difference between the estimated impact in early award sites and the estimated impact in late award sites on the highly explicit instruction measure was statistically significant. The impact was greater in the late award sites.

The differences in estimated impacts on student engagement with print across the award subgroups were not statistically significant. These differences were consistent with observed differences in impacts on test scores. Exhibit 4.11 indicates that estimated impacts on student engagement with print for the early award sites were negative, while those for the late award sites were either positive or less negative.

The overall difference in impacts on classroom instruction was evaluated using a composite test using an index that combines the three instructional outcomes and pools data from first and second grades. The composite test suggests that overall, Reading First produced a positive and statistically significant impact on reading instruction in the late award sites, and that the estimated impact in the early award sites was positive but not statistically significant. The overall difference in impacts on instruction between the two groups of sites was not statistically significant.

## Exhibit 4.10: Estimated Impacts on Reading Instruction, by Award Status

| Instructional Outcomes | Actual Mean with Reading First | Estimated Mean without Reading First | Impact | Effect Size of Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|---|
| **Early Award Sites** | | | | | |
| Number of minutes of instruction in five dimensions combined | | | | | |
| Grade 1 | 62.6 | 57.8 | 4.73 | 0.23 | (0.336) |
| Grade 2 | 64.0 | 56.5 | 7.49 | 0.35 | (0.149) |
| Percent of intervals in five dimensions with highly explicit instruction | | | | | |
| Grade 1 | 30.8 | 26.4 | 4.32 | 0.24 | (0.080) |
| Grade 2 | 31.7 | 29.3 | 2.39 | 0.12 | (0.391) |
| Percent of intervals in five dimensions with high quality student practice | | | | | |
| Grade 1 | 19.3 | 20.1 | -0.85 | -0.05 | (0.720) |
| Grade 2 | 18.6 | 13.3 | 5.26* | 0.29* | (0.022) |
| **Late Award Sites** | | | | | |
| Number of minutes of instruction in five dimensions combined | | | | | |
| Grade 1 | 56.9 | 45.4 | 11.57* | 0.56* | (0.001) |
| Grade 2 | 56.2 | 40.5 | 15.63* | 0.74* | (<0.001) |
| Percent of intervals in five dimensions with highly explicit instruction | | | | | |
| Grade 1 | 29.0 | 25.9 | 3.14 | 0.18 | (0.135) |
| Grade 2 | 31.4 | 21.0 | 10.46* | 0.54* | (<0.001) |
| Percent of intervals in five dimensions with high quality student practice | | | | | |
| Grade 1 | 19.2 | 16.9 | 2.27 | 0.14 | (0.223) |
| Grade 2 | 18.9 | 16.3 | 2.61 | 0.15 | (0.162) |

**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. There are 8 late award sites, with 137 schools, and 10 early award sites, with 111 schools.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005, fall 2005, and spring 2006 IPRI data (by grade).

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

**EXHIBIT READS: The observed mean amount of time spent in instruction in the five dimensions (phonemic awareness, phonics, vocabulary, fluency, and comprehension) in first grade classrooms with Reading First in early award sites was 62.6 minutes. The estimated mean amount of time without Reading First was 57.8 minutes. The impact of Reading First on the amount of time spent in instruction in the five dimensions was 4.73 minutes (or 0.23 standard deviations), which was not statistically significant at the $p \leq .05$ level (p=.336).**

*Sources: RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006.*

**Exhibit 4.11: Estimated Impacts on the Percentage of Students Engaged with Print: Fall 2005 and Spring 2006, by Award Status**

| Construct | Actual Mean with Reading First | Estimated Mean without Reading First | Impact (percentage points) | Effect Size of Impact | Statistical Significance of Impact (p-value) |
|---|---|---|---|---|---|
| ***Grade 1*** | | | | | |
| Early award schools | 48.24 | 51.34 | -3.10 | -0.11 | (0.622) |
| Late award schools | 45.88 | 35.10 | 10.78* | 0.37* | (0.019) |
| ***Grade 2*** | | | | | |
| Early award schools | 50.76 | 66.53 | -15.77* | -0.55* | (0.008) |
| Late award schools | 48.93 | 52.18 | -3.24 | -0.11 | (0.523) |

**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the fall 2005 and spring 2006 STEP data (by grade).

Values in the "Actual Mean with Reading First" column are actual, unadjusted values for Reading First schools; values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding and are calculated by subtracting the impact estimates from the RF schools' actual mean values.

A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

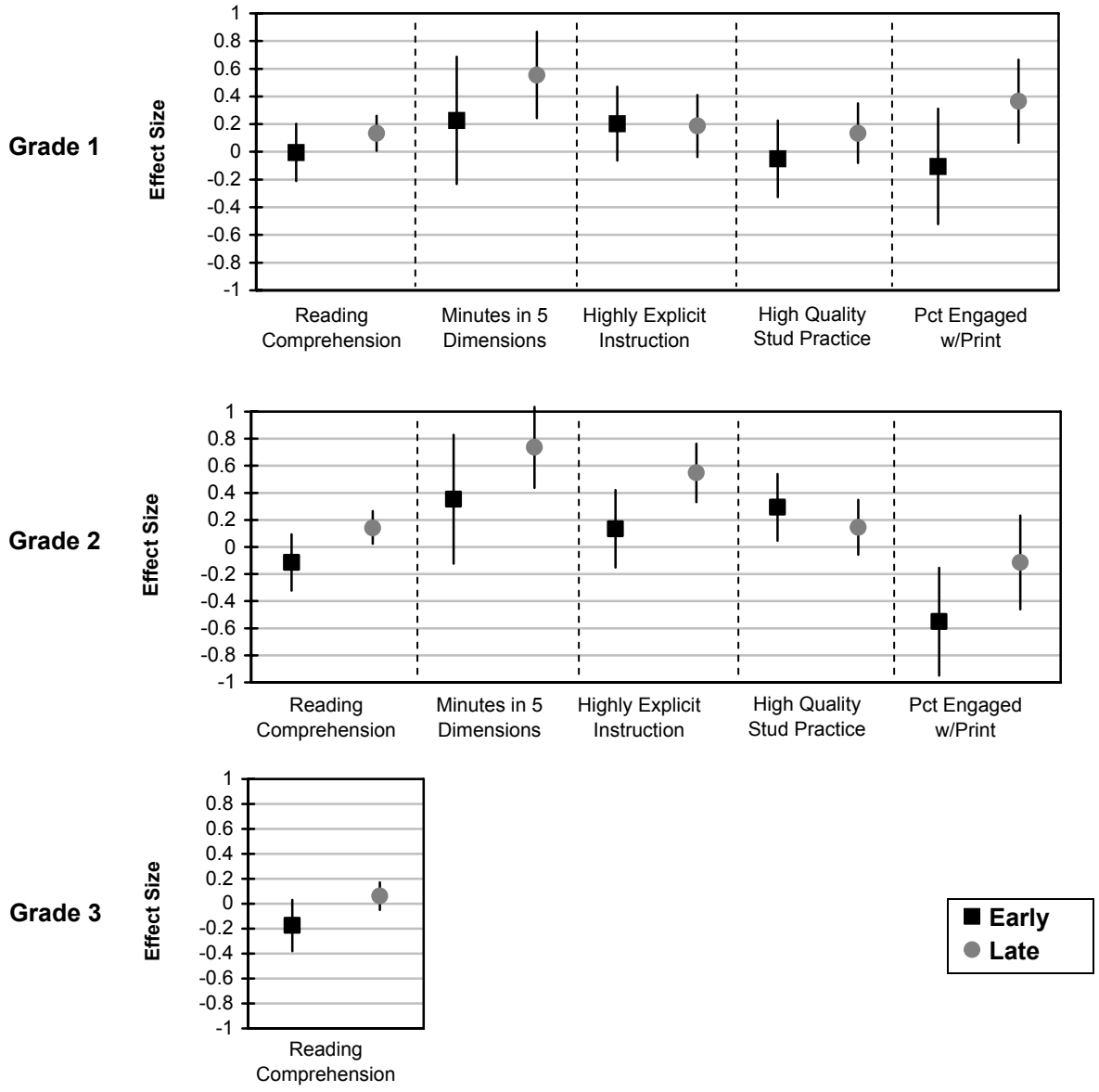
**EXHIBIT READS: The observed average percentage of students engaged with print in first grade classrooms with Reading First in early award sites was 48.24 percent. The estimated average percentage without Reading First was 51.34 percent. The impact of Reading First on the percentage of first grade students engaged with print in early award sites was –3.10 percentage points (or -0.11 standard deviations), which was not statistically significant at the $p \leq .05$ level (p=.622).**

*Sources: RFIS Student Time-on-Task and Engagement with Print, fall 2005 and spring 2006.*

Exhibit 4.12 provides a visual representation of the preceding impact analyses for early and late award sites. There are three panels in the exhibit, one for each grade. Impact estimates (in effect size) for early award sites are represented by small squares, and their 95 percent confidence intervals are represented by vertical lines above and below each square; late award sites are represented by small circles. The wider the confidence interval, the less reliable the impact estimate. If a 95 percent confidence interval does not include zero, the estimated impact is statistically significant (p-value less than or equal to 0.05).

These findings illustrate that for Grades 1 and 2, impacts for late award sites were consistently statistically significant and positive for both classroom instruction in the five dimensions of reading and student reading comprehension. The impacts on these outcomes were not statistically significant for the early award sites, and the pattern of impacts reflects a mix of positive and negative estimates.

## Exhibit 4.12: Estimated Impacts and Confidence Intervals for Key Outcomes, in Effect Size, by Grade, by Award Status




**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. There are 10 early award sites, with 111 schools, and 8 late award sites, with 137 schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available.

Impact estimates are statistically adjusted (e.g., take each school's rating, site-specific funding cut-point, and other covariates into account) to reflect the regression discontinuity design of the study.

The effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools pooled across the spring 2005 and 2006 SAT 10 test scores; spring 2005, fall 2005, and spring 2006 IPRI data; and fall 2005 and spring 2006 STEP data (by grade).

**EXHIBIT READS: For each outcome and grade level, impact estimates and 95 percent confidence intervals are presented in effect size terms. For grade 1, none of the impact estimates across the two award groups are statistically significantly different from each other, although for four of the five outcomes the estimates are (nonsignificantly) lower for the early award sites than for the late award sites. (See Exhibits 4.9, 4.10, and 4.11 for actual impact estimates by award status.)**

*Source: RFIS SAT 10 administration in the spring of 2005 and 2006, as well as from state/district education agencies in those sites that already use the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR); RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, and spring 2006; and RFIS Student Time-on-Task and Engagement with Print, fall 2005 and spring 2006.*

The pattern of findings discussed above raises two important questions for the RFIS. First, what characteristics of these two groups of sites might help to explain the observed variation? The next section below attempts to shed some light on this question by describing some of the differences in characteristics, and by examining the relationship between differences in Reading First impacts and related differences in selected characteristics. Note that the analyses presented here are exploratory, and cannot provide definitive evidence about what caused the observed differences in impacts across the two groups.

A second question arises when one considers the juxtaposition of impacts for late and early award sites. Specifically, in late award sites, where impacts on teachers' instruction in the five dimensions were positive and statistically significant, impacts on reading comprehension test scores were consistently positive and statistically significant for Grades 1 and 2. In early award sites, estimated impacts on teachers' instruction in the five dimensions were positive but not statistically significant, and estimated impacts on student reading comprehension were negative and not statistically significant. The study's final report will explore the relationship between the magnitude of observed impacts on teachers' instructional behavior and observed impacts on student reading comprehension.

## A Preliminary Exploration of Factors That Could Be Related to Program Impacts

This section presents a preliminary exploration of factors that could be related to the differences observed between the impacts of Reading First for early and late award sites. First, the two subgroups of sites are compared on a broad range of characteristics, some of which indicate statistically significant differences. Next, the discussion examines the relationship between program impacts and two selected characteristics of sites in more detail: (1) the amount of Reading First funding allocated per K-3 student in Reading First schools, and (2) the levels of reading comprehension exhibited by students in non-Reading First schools in fall 2004.

It should be noted, however, that it is not possible to provide conclusive evidence about what caused the observed differences between Reading First impacts for early award sites and late award sites, for at least three reasons. First, given the small number of sites in the study sample and the high level of impact estimation error for each site, there is little statistical power to distinguish impact differences, whether across sites or across subgroups of sites. Thus, only large differences can be statistically significant. This is a consequence of the fact that the study was designed to provide valid and reliable estimates of overall average program impacts. A comprehensive examination of variations in impacts across sites or subgroups of sites would require a much larger sample than is represented in the RFIS.

Second, there are more potential factors that could differentiate between the two subgroups of sites than there are sites in total; consequently, there are too few degrees of freedom to estimate a precise statistical model of the determinants of impacts by site. Third, holding aside the degrees of freedom issue, any such model can produce biased estimates because it cannot control for potentially important factors that have not been measured. For this set of reasons, the findings presented below can only be considered as suggestive. Nevertheless, the present analysis would be incomplete without having considered empirically which (observable) factors might be related to the differences in observed program impacts for the two subgroups of sites.

### Related Differences Between Site Award Subgroups

Other potentially relevant ways in which the two subgroups of sites differ provide a context for interpreting the observed impact differences for early and late award sites. Toward this end, Exhibits 4.13 and 4.14 provide as comprehensive a comparison of the two subgroups as is possible given available data. The information in Exhibit 4.13 indicates that:

- On average, late award sites allocated more Reading First funding per school and per student than did early award sites. Hence, there may have been a greater concentration of resources to produce change in the late award sites.
- On average, third grade students from schools without Reading First in the late award sites were less likely to be reading at grade level than those from the early award sites. There may have been a greater margin for improvement in the late award sites (since the study does not have data from early award sites from before they began their implementation of RF, it is not possible to know definitively that early award sites had more or less room for improvement).

**Exhibit 4.13: Characteristics of Early and Late Award Sites**

| Characteristic | Early Award Sites | Late Award Sites |
|---|---|---|
| Average number of months of Reading First funding (current as of May 2006) | 34 months | 25 months |
| Percent of schools in LEA receiving a Reading First grant | 35 percent | 16 percent |
| Average Reading First grant amount (per school) | $97,776 | $143,850 |
| Average Reading First grant amount (per student) | $432 | $574 |
| Fall 2004 reading performance of comparison schools (percent of students at or above grade level–grades 1, 2, and 3) [a] | 54 percent | 43 percent |

**Notes:**

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. There are 10 early award sites, with 111 schools, and 8 late award sites, with 137 schools.

[a] The RFIS SAT 10 administration in fall 2004 occurred an average of 15 months after Reading First funds were made available in early award sites and an average of 5 months after Reading First funds were made available in the late award sites.

**EXHIBIT READS: Schools in early award sites had received Reading First funding for an average of 34 months (as of May 2006).**

*Sources: RFIS SAT 10 administration in fall 2004, http://www.sedl.org/readingfirst/welcome.html, http://www.ed.gov/programs/readingfirst/awards.html*

Exhibit 4.14 compares baseline characteristics of the two subgroups of sites. The top panel compares their student characteristics, the middle panel compares their school characteristics, and the bottom line compares their prior test performance. Findings indicate that the two groups differ on several characteristics, including percent eligible for free and reduced price lunch, locale, and prior student reading performance.

**Exhibit 4.14: Baseline Characteristics of RFIS Reading First Schools, by Award Status**

| Characteristic | Reading First Schools (Early award sites) | Reading First Schools (Late award sites) | Difference[b] | Statistical Significance of Difference (p-value) |
|---|---|---|---|---|
| **Students** | | | | |
| Demographic information | | | | |
| Male (%) | 52.0 | 52.5 | -0.58 | (0.245) |
| Race (%) | | | | |
| Asian | 2.0 | 4.0 | -1.99* | (0.033) |
| Black | 33.0 | 37.9 | -4.87 | (0.436) |
| Hispanic | 33.3 | 20.9 | 12.37* | (0.025) |
| White | 31.2 | 36.7 | -5.47 | (0.327) |
| American Indian/Alaskan | 0.5 | 0.5 | -0.04 | (0.772) |
| Free and Reduced Price Lunch | 68.3 | 79.8 | -11.54* | (0.001) |
| **Schools** | | | | |
| Eligible for Title 1(%) | 94.5 | 100.0 | -5.45* | (0.048) |
| Locale (%) | | | | |
| Large City | 18.2 | 55.7 | -37.53* | (<0.001) |
| Mid-size City | 74.5 | 7.1 | 67.40* | (<0.001) |
| Other | 7.3 | 37.1 | -29.87* | (<0.001) |
| Size | | | | |
| Total Number of Students | 466.3 | 481.5 | -15.20 | (0.655) |
| Number of Students in Grade 3 | 67.1 | 75.2 | -8.11 | (0.152) |
| Student/Teacher Ratio | 15.0 | 15.2 | -0.23 | (0.646) |
| **Third Grade Reading Performance** | | | | |
| Deviation from State RF Mean Proficiency Rate (%)[a] | 1.8 | -4.0 | 5.77* | (0.009) |
| **Number of Schools** | 55 | 70 | | |

**Notes:**

The early sites include 111 schools from 10 sites located in 7 states; 55 schools are Reading First and 56 are non-Reading First schools.

The late sites include 137 schools from 8 sites located in 8 states; 70 schools are Reading First and 67 are non-Reading First.

[a] A school's proficiency score is defined as the percentage of third grade students (or fourth or fifth grade when third grade is unavailable) in the school who scored at or above the state-defined proficiency threshold on the state's reading assessment. The values in this row represent the average percentage point deviation from the mean proficiency score for the Reading First schools in the state.

[b] A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq .05$ level are indicated by *.

*Sources: Data on baseline characteristics are from the Common Core of Data.*

### Associations Between Program Impacts and Two Site Characteristics

Several exploratory analyses were conducted to examine potential relationships between the impacts of Reading First and the amount of Reading First funding per K-3 student and the fall 2004 reading achievement of students in non-Reading First schools. These analyses illustrate ways that relationships between site characteristics and program impacts can be studied.

The analysis for each site characteristic has two parts. First, sites were separated into two subgroups based on the characteristic of interest. These subgroups were as balanced as possible with respect to the number of Reading First schools. Thus, the 18 study sites were split into two roughly equivalent subgroups based on their Reading First funding per K-3 student (after being ordered from lowest to highest per-student allocations). Estimated program impacts for the two subgroups were then compared. A similar analysis was conducted based on two subgroups of sites that were defined in terms of the test scores of students in non-Reading First schools (after being ordered from lowest to highest based on fall 2004 reading performance).[52]

Results of these analyses (see Exhibits H-4 to H-9, panel 3, in Appendix H) suggest some observed differences in impacts, although none of these differences was statistically significant. Hence, the tests do not provide reliable evidence of an existing relationship between program impacts and either Reading First funding per K-3 student or the fall 2004 student reading achievement in non-Reading First schools.

A second part of the analysis was conducted for each of the two site characteristics by estimating an interaction between a continuous measure of the characteristic (at the site-level) and the treatment indicator (for Reading First status) in the statistical model used to estimate program impacts. The sign and size of the coefficient for this interaction reflects the linear relationship that exists between the impact of Reading First and the characteristic (or moderator). Results of these tests (see Appendix H) indicate that sites with higher allocations of Reading First funds per K-3 student had larger program impacts on student achievement than did sites with lower allocations. This relationship was statistically significant for grades one and two.

## Summary

At its core, Reading First is a federal funding process designed to influence local education policy and teacher behavior with the ultimate goal of improving student reading proficiency. Reading First funding deliberately targets classroom reading instruction as a necessary precursor to improved student reading performance. Yet improving students' reading performance is a priority for all schools, and particularly for those whose students are reading below grade level.

However, after up to three years of funding, the study finds, on average, that Reading First's impact on student reading achievement was not statistically detectable. Furthermore, the Reading First Impact Study indicates that schools receiving Reading First grants are still well short of the program's

[52] For both analyses, a robustness test was conducted by repeating the analysis after dropping the site from each subgroup that is closest to the cut-point between them. This was repeated again after dropping the two sites from each subgroup that are closest to the cut-point. The conclusion of each analysis was not highly sensitive to this deletion of sites, although levels of statistical significance declined with the corresponding decline in sample size.

ultimate goal of ensuring that all students are reading at grade level by the end of third grade. Half or more of the third grade students in the study sample's Reading First schools were performing below grade level three years into the initiative, according to SAT 10 grade level norms (which may differ from states' definitions of on or above grade level). Yet the findings indicate that Reading First did produce some positive and statistically significant improvements in first and second grade students' reading comprehension test scores in a group of sites that had received their RF funds between January and August 2004, and those same sites also experienced positive and statistically significant effects from Reading First on the instructional time that first and second grade teachers spent on the five dimensions of reading. The final report will address the question of whether changes in teacher instructional practices are associated with student reading performance.

The RFIS has completed its third year of data collection, which will provide considerably more data for the final report, including an additional year of data on students' reading comprehension, teachers' classroom instruction (three years in total), and student engagement with print (two years in total). The final report will draw upon additional data collected in the 2006-07 school year, including assessments of first grade students' decoding skills and surveys of educational personnel. The availability of these additional data will allow the study team to answer questions about the impact of Reading First more definitively, to explore relationships between observed impacts of Reading First on instructional outcomes and reading achievement, and to assess whether there are statistically significant and educationally meaningful variations in impacts. The additional data and analysis of factors that may influence the implementation and impact of the program may shed further light on the ability of Reading First to achieve its ultimate goal.