NCEE 2009-4038

U.S. DEPARTMENT OF EDUCATION

Reading First Impact Study

Final Report

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

Reading First Impact Study
Final Report

NOVEMBER 2008

Beth C. Gamse, Project Director, Abt Associates
Robin Tepper Jacob, Abt Associates/University of Michigan
Megan Horst, Abt Associates
Beth Boulay, Abt Associates
Fatih Unlu, Abt Associates

Laurie Bozzi
Linda Caswell
Chris Rodger
W. Carter Smith
Abt Associates

Nancy Brigham
Sheila Rosenblum
Rosenblum Brigham Associates

With the assistance of
Howard Bloom
Yequin He
Corinne Herlihy
James Kemple
Don Laliberty
Ken Lam
Kenyon Maree
Rachel McCormick
Rebecca Unterman
Pei Zhu

NCEE 2009-4038
U.S. DEPARTMENT OF EDUCATION

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.

November 2008

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the reports.

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 thank 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, 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 (Sherry Sanborne and Alex Ratnofsky, for managing the student assessment, and the Student Assessment Coordinators and test administrators), and from Westover (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
Sean Reardon
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)
Jacob Klerman (A)
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

We also want to recognize the steady contributions of Abt-SRBI staff, including Brenda Rodriguez, Fran Coffey, Kay Ely, Joanne Melton, Judy Meyer, Lynn Reneau, Davyd Roskilly, Jon Schmalz, Estella Sena, and Judy Walker, who were instrumental in completing multiple data collections, and Eileen Fahey, Katheleen Linton, and Jan Nicholson for countless hours of production support. Finally, we want to acknowledge Diane Greene, whose wisdom helped us all.

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

Contents (continued)

Contents (continued)

Contents (continued)

List of Exhibits

List of Exhibits (continued)

List of Exhibits (continued)

List of Exhibits (continued)

List of Exhibits (continued)

List of Exhibits (continued)

Executive Summary

This report presents findings from the third and final year of the Reading First Impact Study (RFIS), 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 (PL 107-110, Title I, Part B, Subpart 1) established Reading First (RF) and mandated its evaluation. This evaluation is being conducted by Abt Associates and MDRC with collaboration from 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).

This report examines the impact of Reading First funding on 248 schools in 13 states and includes 17 school districts and one statewide program for a total of 18 sites. The study includes data from three school years: 2004-05, 2005-06 and 2006-07.

The Reading First Impact Study 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 primary measure of student reading achievement was the Reading Comprehension subtest from the Stanford Achievement Test—10 (SAT 10), given to students in grades one, two, and three. A secondary measure of student reading achievement in decoding was given to students in first grade. The measure of classroom reading instruction was derived from direct observations of reading instruction, and measures of program implementation were derived from surveys of educational personnel. Findings related to the first two questions are based on results pooled across the study's three years of data collection (2004-05, 2005-06, and 2006-07) for classroom instruction and reading comprehension, results from first grade students in one school year (spring 2007) for decoding, and aspects of program implementation from spring 2007 surveys. Key findings are as follows:

- Reading First produced a positive and statistically significant impact on amount of instructional time spent on the five essential components of reading instruction promoted by the program (phonemic awareness, phonics, vocabulary, fluency, and comprehension) in grades one and two. The impact was equivalent to an effect size of 0.33 standard deviations in grade one and 0.46 standard deviations in grade two.
- Reading First produced positive and statistically significant impacts on multiple practices that are promoted by the program, including professional development in scientifically based reading instruction (SBRI), support from full-time reading coaches, amount of reading instruction, and supports available for struggling readers.
- Reading First did not produce a statistically significant impact on student reading comprehension test scores in grades one, two or three.

- Reading First produced a positive and statistically significant impact on decoding among first grade students tested in one school year (spring 2007). The impact was equivalent to an effect size of 0.17 standard deviations.

Results are also presented from exploratory analyses that examine some hypotheses about factors that might account for the observed patterns of impacts. These analyses are considered exploratory because the study was not designed to provide a rigorous test of these hypotheses, and therefore the results must be considered as suggestive. Across different potential predictors of student outcomes, these exploratory analyses are based on different subgroups of students, schools, grade levels, and/or years of data collection. Key findings from these exploratory analyses are as follows:

- There was no consistent pattern of effects over time in the impact estimates for reading instruction in grade one or in reading comprehension in any grade. There appeared to be a systematic decline in reading instruction impacts in grade two over time.
- There was no relationship between reading comprehension and the number of years a student was exposed to RF.
- There is no statistically significant site-to-site variation in impacts, either by grade or overall, for classroom reading instruction or student reading comprehension.
- There is a positive association between time spent on the five essential components of reading instruction promoted by the program and reading comprehension measured by the SAT 10, but these findings are sensitive to both model specification and the sample used to estimate the relationship.

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 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, and 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.

Reading First grants were made available 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.[2] 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). States could reserve up to 20 percent of their Reading First funds to support staff development, technical assistance to districts and schools, and planning, administration and reporting. According to the program guidance, this funding provided "States with the resources and opportunity…to improve instruction beyond the specific districts and schools that receive Reading First subgrants." (U.S. Department of Education, 2002). Districts could reserve up to 3.5 percent of their Reading First funds for planning and administration (No Child Left Behind Act, 2001). For the purposes of this study, Reading First is defined as the receipt of Reading First funding at the school level.

The Reading First Impact Study

Research Design

The Reading First Impact Study uses a regression discontinuity design that capitalizes on the systematic processes some school districts used to allocate Reading First funds once their states had received RF grants.[3] A regression discontinuity design is the strongest quasi-experimental method available to produce unbiased estimates of program impacts. Under certain conditions, all of which are met by the present study, this method can produce unbiased estimates of program impacts. Within each district or site:

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;[4]
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; and
3) Funding decisions were based only on whether a school's rating was above or below its local cut-point; nothing superseded these decisions.

Also, assuming that the shape of the relationship between schools' ratings and outcomes is correctly modeled, once the above conditions have been met, 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 ratings used to determine funding decisions. Controlling for differences in schools' ratings allows one to control statistically for all systematic pre-existing differences between the two groups. One then can estimate the impact of Reading First by comparing the outcomes for Reading First schools and non-Reading First schools in the study

[2] Data were obtained from the SEDL website (www.sedl.org/readingfirst).

[3] Appendix A indicates when study sites first received their Reading First grants.

[4] Each study site could (and did) use different metrics to rate or rank schools; it is not necessary for all study sites to use the same metric.

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—it is their regression-adjusted outcomes that represent the best indications of what outcomes would have been for the treatment group (in this instance, Reading First schools) in the absence of the program being evaluated.

Study Sample

The study sample was selected purposively to meet the requirements of the regression discontinuity design by selecting a sample of sites that had used a systematic rating or ranking process to select their Reading First school grantees. Within these sites, the selection of schools focused on schools as close to the site-specific cut-points as possible in order to obtain schools that were as comparable as possible in the treatment and comparison groups.

The study sample includes 18 study sites: 17 school districts and one state-wide program. Sixteen districts and one state-wide program were selected from among 28 districts and one state-wide program that had demonstrably met the three criteria listed above. One other school district agreed to randomly assign some of its eligible schools to Reading First or a control group. The final selection reflected wide variation in district characteristics and provided enough schools to meet the study's sample size requirements. The regression discontinuity sites provide 238 schools for the analysis, and the randomized experimental site provides 10 schools. Half the schools at each site are Reading First schools and half are non-Reading First schools: in three sites, the study sample includes all the RF schools (in that site), in the remaining 15 sites, the study sample includes some, but not all, of the RF schools (in that site).

At the same time, the study deliberately endeavored to obtain a sample that was geographically diverse and as similar as possible to the population of all RF schools. The final study sample of 248 schools, 125 of which are Reading First schools, represents 44 percent of the Reading First schools in their respective sites (at the time the study selected its sample in 2004). The study's sample of RF schools is large, is quite similar to the population of all RF schools, is geographically diverse, and represents states (and districts) that received their RF grants across the range of RF state award dates. The average Year 1 grant for RF schools in the study sample ranged from about $81,790 to $708,240, with a mean of $188,782. This translates to an average of $601 per RF student. For more detailed information about the selection process and the study sample, see the study's Interim Report (Gamse, Bloom, Kemple & Jacob, 2008).

Data Collection Schedule and Measures

Exhibit ES.1 summarizes the study's three-year, multi-source data collection plan. The present report is based on data for school years 2004-05, 2005-06, and 2006-07. Data collection included student assessments in reading comprehension and decoding, and classroom observations of teachers' instructional practices in reading, teachers' instructional organization and order, and students' engagement with print. Data were also collected through surveys of teachers, reading coaches, and principals, and interviews of district personnel.

Exhibit ES.1: Data Collection Schedule for the Reading First Impact Study

Data Collection Elements	2004-2005		2005-2006		2006-2007	
	Fall	Spring	Fall	Spring	Fall	Spring
Student Testing	✓	✓		✓		✓
Stanford Achievement Test, 10th Edition (SAT 10)	✓	✓		✓		✓
Test of Silent Word Reading Fluency (TOSWRF)						✓
Classroom Observations		✓	✓	✓	✓	✓
Instructional Practice in Reading Inventory (IPRI)		✓	✓	✓	✓	✓
Student Time-on-Task and Engagement with Print (STEP)			✓	✓	✓	✓
Global Appraisal of Teaching Strategies (GATS)			✓	✓	✓	✓
Teacher, Principal, Reading Coach Surveys		✓				✓
District Staff Interviews		✓				✓

Exhibit ES.2 lists the principal domains for the study, the outcome measures within each domain, and the data sources for each measure. These include:

Student reading performance, assessed with the reading comprehension subtest of the Stanford Achievement Test, 10th Edition (SAT 10, Harcourt Assessment, Inc., 2004). The SAT 10 was administered to students in grades one, two and three during fall 2004, spring 2005, spring 2006, and spring 2007, with an average completion rate of 83 percent across all administrations. In the spring of 2007 only, first grade students were assessed with the Test of Silent Word Reading Fluency (TOSWRF, Mather et al., 2004), a measure designed to assess students' ability to decode words from among strings of letters. The average completion rate was 86 percent. Three outcome measures of student reading performance were created from SAT 10 and TOSWRF data.

Classroom reading instruction, 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 during scheduled reading blocks in each sampled classroom on two consecutive days during each wave of data collection: spring 2005, fall 2005 and spring 2006, and fall 2006 and spring 2007. The average completion rate was 98 percent across all years. The IPRI, which is designed to record instructional behaviors in a series of three-minute intervals, can be used for observations of varying lengths, reflecting the fact that schools' defined reading blocks can and do vary. Most reading blocks are 90 minutes or more. Eight outcome measures of classroom reading instruction were created from IPRI data to represent the components of reading instruction emphasized by the Reading First legislation.[5] Six of these measures are reported in terms of the amount of time spent on the

[5] 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.

Exhibit ES.2: Description of Domains, Outcome Measures, and Data Sources Utilized in the Reading First Impact Study

Domain	Outcome Measure and Description	Source
Student reading performance	**Mean scaled scores for 1st, 2nd, and 3rd grade students**, 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.	*Stanford Achievement Test, 10th Edition (SAT 10)*
	Percentage of 1st, 2nd, and 3rd grade students at or above grade level, 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.	*Stanford Achievement Test, 10th Edition (SAT 10)*
	Mean standard scores for 1st grade students, represented as a continuous measure of first grade students' decoding skill.	*Test of Silent Word Reading Fluency*
Classroom reading instruction	**Minutes of instruction in phonemic awareness**, or how much instructional time 1st and 2nd grade teachers spent on phonemic awareness.	*RFIS Instructional Practice in Reading Inventory*
	Minutes of instruction in phonics, or how much instructional time 1st and 2nd grade teachers spent on phonics.	*RFIS IPRI*
	Minutes of instruction in fluency building, or how much instructional time 1st and 2nd grade teachers spent on fluency building.	*RFIS IPRI*
	Minutes of instruction in vocabulary development, or how much instructional time 1st and 2nd grade teachers spent on vocabulary development.	*RFIS IPRI*
	Minutes of instruction in comprehension, or how much instructional time 1st and 2nd grade teachers spent on comprehension of connected text.	*RFIS IPRI*
	Minutes of instruction in all five dimensions combined, or how much instructional time 1st and 2nd grade teachers spent on all five dimensions combined.	*RFIS IPRI*
	Proportion of each observation with highly explicit instruction, or the proportion of time spent within the five dimensions when teachers used highly explicit instruction (e.g., instruction included teacher modeling, clear explanations, and the use of examples).	*RFIS IPRI*
	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 (e.g., teachers asked students to practice such word learning strategies as context, word structure, and meanings).	*RFIS IPRI*
Student engagement with print	**Percentage of 1st and 2nd grade students engaged with print**, represented as the per-classroom average of the percentage of students engaged with print across three sweeps in each classroom during observed reading instruction.	*RFIS Student Time-on-Task and Engagement with Print (STEP)*

Exhibit ES.2: Description of Domains, Outcome Measures, and Data Sources Utilized in the Reading First Impact Study (continued)

Domain	Outcome Measure and Description	Source
Professional development in scientifically based reading instruction	**Amount of PD in reading received by teachers**, or teachers' self-reported number of hours of professional development in reading during 2006-07.	*RFIS Teacher Survey*
	Teacher receipt of PD in the five essential components of reading instruction, or the number of essential components teachers reported were covered in professional development they received during 2006-07.	*RFIS Teacher Survey*
	Teacher receipt of coaching, or whether or not a teacher reported receiving coaching or mentoring from a reading coach in reading programs, materials, or strategies in 2006-07.	*RFIS Teacher Survey*
	Amount of time dedicated to serving as K-3 reading coach, or reading coaches' self-reported percentage of time spent as the K-3 reading coach for their school in 2006-07.	*RFIS Reading Coach Survey*
Amount of reading instruction	**Minutes of reading instruction per day**, or teachers' reported average amount of time devoted to reading instruction per day over the prior week.	*RFIS Teacher Survey*
Supports for struggling readers	**Availability of differentiated instructional materials for struggling readers**, or whether or not schools reported that specialized instructional materials beyond the core reading program were available for struggling readers.	*RFIS Reading Coach and Principal Surveys*
	Provision of extra classroom practice for struggling readers, or the number of dimensions in which teachers reported providing extra practice opportunities for struggling students in the past month.	*RFIS Teacher Survey*
Use of assessments	**Use of assessments to inform classroom practice**, or the number of instructional purposes for which teachers reported using assessment results.	*RFIS Teacher Survey*

various dimensions of instruction. Two of these measures are reported in terms of the proportion of the intervals within each observation .

Student engagement with print. Beginning in fall 2005, the study conducted 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 observation was completed by recording a time-sampled "snapshot" of student engagement three times in each observed classroom, for a total of three such "sweeps" during each STEP observation. The STEP was used to observe classrooms in fall 2005, spring 2006, fall 2006, and spring 2007, with an average completion rate of 98 percent across all years. One outcome measure was created using STEP data.

Professional development in scientifically based reading instruction, amount of reading instruction, supports for struggling readers, and use of assessments. Within these four domains, eight outcome measures were created based on data from surveys of principals, reading coaches, and teachers about school and classroom resources. The eight outcome measures represent aspects of scientifically based reading instruction promoted in the Reading First legislation and guidance. Surveys were fielded in spring 2005 and again in spring 2007 with an average completion rate across all respondents of 73 percent in spring 2005 and 86 percent in spring 2007. This final report includes findings from 2007 surveys only.

Additional data were collected by the study team in order to create measures used in correlational analyses. These data include:

The ***Global Appraisal of Teaching Strategies (GATS)***, a 12-item checklist designed to measure teachers' instructional strategies related to overall instructional organization and order, is adapted from The Checklist of Teacher Competencies (Foorman and Schatschneider, 2003). Unlike the IPRI, which focuses on discrete teacher behaviors, the GATS was designed to capture global classroom management and environmental factors. Items covered topics such as the teacher's organization of materials, lesson delivery, responsiveness to students, and behavior management. The GATS was completed by the classroom observer immediately after each IPRI observation, meaning that each sampled classroom was rated on the GATS twice in the fall and twice in the spring in both the 2005-2006 school year and the 2006-2007 school year. The GATS was fielded in fall 2005, spring 2006, fall 2006, and spring 2007, with an average completion rate of over 99 percent. A single measure from the GATS data was created for use in correlational analyses.

Average Impacts on Classroom Reading Instruction, Key Components of Scientifically Based Reading Instruction, and Student Reading Achievement

Exhibit ES.3 reports average impacts on classroom reading instruction and student reading comprehension pooled across school years 2004-05 and 2005-06 and 2006-07.[6] Exhibit ES.4 reports average impacts on key components of scientifically based reading instruction from spring 2007. Exhibit ES.5 reports the average impact on first graders' decoding skills from spring 2007. 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 typical study school. On average:

- Reading First had a statistically significant impact on the total time that teachers spent on the five essential components of reading instruction promoted by the program in grades one and two.
- Reading First had a statistically significant impact on the use of highly explicit instruction in grades one and two and on the amount of high quality student practice in grade two. Its estimated impact on high quality student practice for grade one was not statistically significant.
- Reading First had no statistically significant impacts on student engagement with print.
- Reading First had a statistically significant impact on the amount of professional development in reading teachers reported receiving; teachers in RF schools reported receiving 25.8 hours of professional development compared to what would have been expected without Reading First (13.7 hours). The program also had a statistically significant impact on teachers' self-reported receipt of professional development in the five essential components of reading instruction; teachers in RF schools reported receiving professional development on an average of 4.3 of 5 components, compared to what would have been expected without Reading First (3.7 components).

[6] Except for student engagement with print (STEP), which is pooled across the 2005-06 and 2006-07 school years only.

- A statistically significantly greater proportion (20 percent) of teachers in RF schools reported receiving coaching from a reading coach than would be expected without Reading First. The program also had a statistically significant impact on the amount of time reading coaches reported spending in their role as the school's reading coach; coaches in RF schools reported spending 91.1 percent of their time in this role, 33.5 percentage points more than would be expected without Reading First (57.6 percent).
- Reading First had a statistically significant impact on the amount of time teachers reported spending on reading instruction per day. Teachers in RF schools reported an average of 105.7 minutes per day, 18.5 minutes more than the 87.2 minutes that would be expected without Reading First.
- Reading First had a statistically significant impact on teachers' provision of extra classroom practice in the essential components of reading instruction in the past month; the impact was 0.2 components.
- There were no statistically significant impacts of Reading First on the availability of differentiated instructional materials for struggling readers or on teachers' reported use of assessments to inform classroom practice for grouping, diagnostic, and progress monitoring purposes.
- Reading First had no statistically significant impact on students' reading comprehension scaled scores or the percentages of students whose reading comprehension scores were at or above grade level in grades one, two or three. The average first, second, and third grade student in Reading First schools was reading at the 44^{th}, 39^{th}, and 39^{th} percentile respectively on the end-of-the-year assessment (on average over the three years of data collection).
- Reading First had a positive and statistically significant impact on average scores on the TOSWRF, a measure of decoding skill, equivalent to 2.5 standard score points, or an effect size of 0.17 standard deviations (See Exhibit ES.5). Because the test of students' decoding skills was only administered in a single grade and a single year, it is not possible to provide an estimate of Reading First's overall impact on decoding skills across multiple grades and across all three years of data collection, as was done for reading comprehension.

Exploratory Analyses of Variations in Impacts and Relationships among Outcomes

This report also presents results from exploratory analyses that examine some hypotheses about factors that might account for the pattern of observed impacts presented above. These exploratory analyses are based on analyses of subgroups of students, schools, grade levels, and/or years of data collection. The information is provided as possible avenues for further exploration or for improving Reading First or programs like Reading First. However, the study was not designed to provide a rigorous test of these hypotheses, and therefore the results are only suggestive. Findings from these exploratory analyses include the following:

- Data collected during three school years (2004-05, 2005-06 and 2006-07) were used to examine variation over time in program impacts. No consistent pattern of differential impacts over time was established.
- No relationship was found between the number of years a student was exposed to RF and student reading achievement.

- There was no statistically significant variation in impacts across sites in the study, either by grade or overall, for reading instruction or for reading comprehension.
- Correlational analyses, which are outside the causal framework of the main impact analyses presented in the report, indicate a positive and statistically significant association between time spent on the five essential components of reading instruction promoted by the program and students' reading comprehension. A one-minute increase in time devoted to instruction in the five dimensions per daily reading block was associated with a 0.07 point increase in scaled score points in first grade, and a 0.06 point increase in second grade. This relationship does not hold for models that include other potential mediators of student achievement. However, due to data limitations, these latter models could only be run on a subset of the data; thus, we do not know whether the differences in the findings across models are due to changes in the sample or changes in the model specification itself.

Exhibit ES.3: Estimated Impacts on Reading Comprehension, Instruction, and Percentage of Students Engaged with Print: 2005, 2006, and 2007 (pooled)[1]

	Actual Mean with Reading First	Estimated Mean without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Instruction					
Number of minutes of instruction in the five components combined					
Grade 1	59.23	52.31	6.92*	0.33*	(0.005)
Grade 2	59.08	49.30	9.79*	0.46*	(<0.001)
Percentage of intervals in five components with Highly Explicit Instruction					
Grade 1	29.39	26.10	3.29*	0.18*	(0.018)
Grade 2	30.95	27.95	3.00*	0.16*	(0.040)
Percentage of intervals in five components with High Quality Student Practice					
Grade 1	18.44	17.61	0.82	0.05	(0.513)
Grade 2	17.82	14.88	2.94*	0.16*	(0.019)
Reading Comprehension					
Scaled Score					
Grade 1	543.8	539.1	4.7	0.10	(0.083)
Grade 2	584.4	582.8	1.7	0.04	(0.462)
Grade 3	609.1	608.8	0.3	0.01	(0.887)
Percent Reading At or Above Grade Level					
Grade 1	46.0	41.8	4.2	--	(0.104)
Grade 2	38.9	37.3	1.6	--	(0.504)
Grade 3	38.7	38.8	-0.1	--	(0.973)
Percentage of Students Engaged with Print					
Grade 1	47.84	42.52	5.33	0.18	(0.070)
Grade 2	50.53	55.27	-4.75	-0.17	(0.104)

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 in 2006, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one 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; statistically significant findings at the $p \leq .05$ level are indicated by *.

[1]Except for STEP, which is pooled across 2006 and 2007 school years only.

EXHIBIT READS: The observed mean amount of time spent per daily reading block in instruction in the five components combined for first grade classrooms with Reading First was 59.23 minutes. The estimated mean amount of time without Reading First was 52.31 minutes. The impact of Reading First on the amount of time spent in instruction in the five components combined was 6.92 (or 0.33 standard deviations), which was statistically significant (p=.005).

SOURCES: RFIS SAT 10 administrations in the spring of 2005, 2006, and 2007, 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, spring 2006, fall 2006, and spring 2007; RFIS Student Time-on-Task and Engagement with Print, fall 2005, spring 2006, fall 2006, and spring 2007.

Exhibit ES.4: Estimated Impacts on Key Components of Scientifically Based Reading Instruction (SBRI): Spring 2007

Domain	Actual Mean With Reading First	Estimated Mean Without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Professional Development (PD) in SBRI					
Amount of PD in reading received by teachers (hours) [a]	25.84	13.71	12.13*	0.51*	(<0.001)
Teacher receipt of PD in the five essential components of reading instruction (0-5) [a]	4.30	3.75	0.55*	0.31*	(0.010)
Teacher receipt of coaching (proportion) [a]	0.83	0.63	0.20*	0.41*	(<0.001)
Amount of time dedicated to serving as K-3 reading coach (percent) [b,c]	91.06	57.57	33.49*	1.03*	(<0.001)
Amount of Reading Instruction					
Minutes of reading instruction per day [a]	105.71	87.24	18.47*	0.63*	(<0.001)
Supports for Struggling Readers					
Availability of differentiated instructional materials for struggling readers (proportion) [b]	0.98	0.97	0.01	0.15	(0.661)
Provision of extra classroom practice for struggling readers (0-4) [a]	3.79	3.59	0.19*	0.20*	(0.018)
Use of Assessments					
Use of assessments to inform classroom practice (0-3) [a]	2.63	2.45	0.18	0.19	(0.090)

NOTES:

[a] Classroom level outcome

[b] School level outcome

[c] The response rates for RF and nonRF reading coach surveys were statistically significantly different (p=0.037). Reading first schools were more likely to have had reading coaches and to have returned reading coach surveys.

[d] Missing data rates ranged from 0.1 to 3.3 percent for teacher survey outcomes (RF: 0.1 to 1.0 percent; non-RF: 0 to 4.9 percent) and 1.3 to 2.8 percent for reading coach and/or principal survey outcomes (RF: 0 to 1.6 percent; non-RF: 2.7 to 4.1 percent). Survey constructs (i.e., those outcomes comprised of more than one survey item) were computed only for observations with complete data, with one qualification: for the construct "minutes spent on reading instruction per day," the mean was calculated as the total number of minutes reported for last week (over a maximum of 5 days) divided by the number of days with non-missing values. Only those teacher surveys with missing data for all 5 days were missing 0.9 percent).

The complete Reading First Impact 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 effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools.

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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: The observed mean amount of professional development in reading received by teachers with Reading First was 25.84 hours. The estimated mean amount of professional development in reading received by teachers without Reading First was 13.71 hours. This impact of 12.13 hours was statistically significantly (p<.001).

SOURCES: RFIS, Teacher, Reading Coach, and Principal Surveys, spring 2007

Exhibit ES.5: Estimated Impacts of Reading First on Decoding Skill: Grade One, Spring 2007

	Actual Mean with Reading First	Estimated Mean without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Decoding Skill					
Standard Score	96.9	94.4	2.5 *	0.17 *	(0.025)
Corresponding Grade Equivalent	1.7	1.4			
Corresponding Percentile	42	35			

NOTES:

The Test of Silent Word Reading Fluency (TOSWRF) sample includes first-graders in 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 from spring 2007 TOSWRF test scores (1st grade).

The key metric for the TOSWRF analyses is the standard score, corresponding grade equivalents and percentiles are provided for reference. Although the publisher of the Test of Silent Word Reading Fluency states that straight comparisons between standard scores and grade equivalents will likely yield discrepancies due to the unreliability of the grade equivalents, they are provided because program criteria are sometimes based on grade equivalents (TOSWRF, Mather et al., 2004).

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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: The observed mean silent word reading fluency standard score for first-graders with Reading First was 96.9 standard score points. The estimated mean without Reading First was 94.4 standard score points. The impact of Reading First was 2.5 standard score points (or 0.17 standard deviations), which was statistically significant (p=.025).

SOURCES: RFIS TOSWRF administration in spring 2007

Summary

The findings presented in this report are generally consistent with findings presented in the study's Interim Report, which found statistically significant impacts on instructional time spent on the five essential components of reading instruction promoted by the program (phonemic awareness, phonics, vocabulary, fluency, and comprehension) in grades one and two, and which found no statistically significant impact on reading comprehension as measured by the SAT 10. In addition to data on the instructional and student achievement outcomes reported in the Interim Report, the final report also presents findings based upon information obtained during the study's third year of data collection: data from a measure of first grade students' decoding skill, and data from self-reported surveys of educational personnel in study schools.

Analyses of the impact of Reading First on aspects of program implementation, as reported by teachers and reading coaches, revealed that the program had statistically significant impacts on several domains. The information obtained from the Test of Silent Word Reading Fluency indicates that Reading First had a positive and statistically significant impact on first grade students' decoding skill.

The final report also explored a number of hypotheses to explain the pattern of observed impacts. Analyses that explored the association between the length of implementation of Reading First in the study schools and reading comprehension scores, as well as between the number of years students had been exposed to Reading First instruction and reading comprehension scores were inconclusive. No statistically significant variation across sites in the pattern of impacts was found. Correlational analyses suggest that there is a positive association between time spent on the five essential components of reading instruction promoted by the program and reading comprehension measured by the SAT 10, but these findings appear to be sensitive to model specification and the sample used to estimate the relationship.

The study finds, on average, that after several years of funding the Reading First program, it has a consistent positive effect on reading instruction yet no statistically significant impact on student reading comprehension. Findings based on exploratory analyses do not provide consistent or systematic insight into the pattern of observed impacts.

Chapter One: Overview of the Reading First Impact Study

The No Child Left Behind Act of 2001 (NCLB) established the Reading First (RF) 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). Abt partnered with other organizations, including MDRC, RMC Research, Rosenblum-Brigham Associates, and Westat.[7] 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 final report presents major findings based on data collected during the 2004-05, 2005-06, and 2006-07 school years. It reviews information about the study background, design, sample, and measures, and it updates information presented in the study's interim report with data from the final year of data collection.

Chapter One begins with an overview of the Reading First Program, describes the conceptual framework underlying the program and this evaluation as a whole, outlines the study's guiding evaluation questions, summarizes the study design, measures, and data collection activities, and presents a roadmap for the remainder of the report.

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.

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

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, and 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.

Reading First grants were made available 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.[8] 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). States could reserve up to 20 percent of their Reading First funds to support staff development, technical assistance to districts and schools, and planning, administration and reporting. According to the program guidance, this funding provided "states with the resources and opportunity…to improve instruction beyond the specific districts and schools that receive Reading First subgrants." (U.S. Department of Education, 2002). Districts could reserve up to 3.5 percent of their Reading First funds for planning and administration (No Child Left Behind Act, 2001). For the purposes of this study, Reading First is defined as the receipt of Reading First funding at the school level.

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 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.[9] Each dimension is described below.

Phonemic Awareness

Phonemic awareness instruction teaches students to distinguish and manipulate the sounds in words.[10] 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).

---

[8] Data were obtained from the SEDL website (www.sedl.org/readingfirst).

[9] 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.

[10] 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.

Phonics
Phonics instruction helps children learn and understand the relationships between the letters of written language and the sounds (phonemes) of spoken language. Instruction in phonics 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 NICHD, 2000).

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 NICHD, 2000).

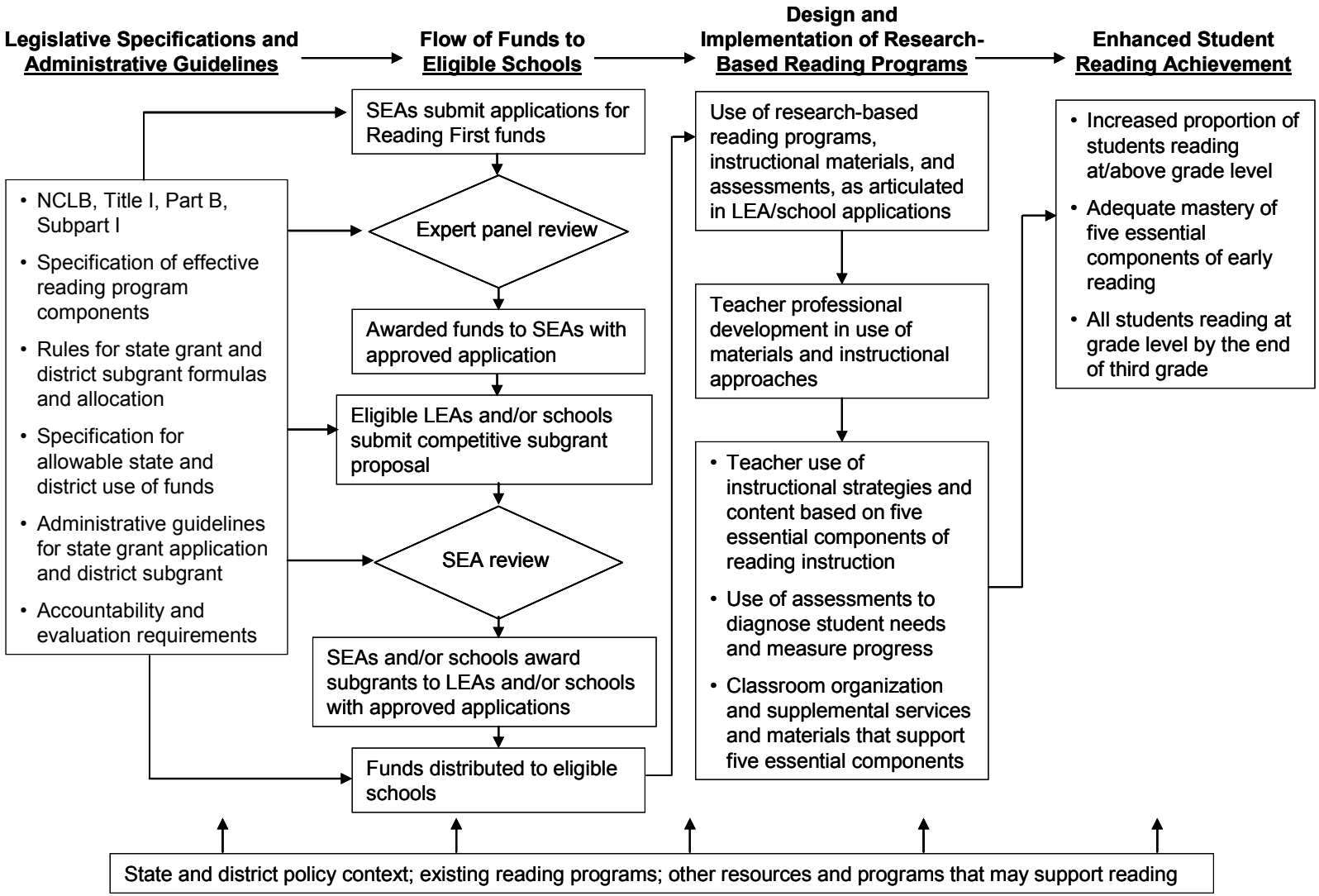
Vocabulary Development
Oral vocabulary refers to words used in speaking or recognized in listening. Reading vocabulary refers to words that are recognized or used in print. Instruction for beginning readers uses oral vocabulary to help them make sense of the words they see, and instruction that develops 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.

Conceptual Model

Exhibit 1.1 identifies the program's central goals and specifies the pathways through which the principles and components of the Reading First program are hypothesized to improve reading instruction, and subsequently student reading achievement. This conceptual framework provides a substantive backdrop for the Reading First Impact Study. The Reading First Impact Study has focused primarily on Column 3 (which specifies aspects of program implementation, including necessary components of scientifically based reading instruction hypothesized to achieve its longer term student achievement goals) and Column 4 (which details aspects of student reading achievement). The hypothesis underlying Reading First is that these outcomes will only be achieved 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.

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



Research Questions and Design

The Reading First Impact Study 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 Reading First Impact Study uses a regression discontinuity design (RDD) that capitalizes on the systematic processes some school districts used to allocate Reading First funds once their states had received RF grants.[11] A regression discontinuity design is the strongest quasi-experimental method available to produce unbiased estimates of program impacts. Under certain conditions, all of which are met by the present study, this method can produce unbiased estimates of program impacts. Within each district or site:

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;[12]
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; and
3) Funding decisions were based only on whether a school's rating was above or below its local cut-point; nothing superseded these decisions.

Also, assuming that the shape of the relationship between schools' ratings and outcomes is correctly modeled, once the above conditions have been met, 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 ratings used to determine funding decisions. Controlling for differences in schools' ratings allows one to control statistically for all systematic pre-existing differences between the two groups. One then can 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—it is their regression-adjusted outcomes that represent the best indications of what outcomes would have been for the treatment group (in this instance, Reading First schools) in the absence of the program being evaluated.[13]

[11] Appendix A indicates when study sites first received their Reading First grants.

[12] Each study site could (and did) use different metrics to rate or rank schools; it is not necessary for all study sites to use the same metric.

[13] See Appendix B of this report and Gamse, Bloom, Kemple & Jacob (2008) for a more extended discussion of the regression discontinuity design, the study sample, and the study's approach to estimating impacts.

Study Sample

The study sample was selected purposively to meet the requirements of the regression discontinuity design by selecting a sample of sites that had used a systematic rating or ranking process to select their Reading First school grantees. Within these sites, the selection of schools focused on schools as close to the site-specific cut-points as possible in order to obtain schools that were as comparable as possible in the treatment and comparison groups.

The study sample includes 18 study sites: 17 school districts and one state-wide program. Sixteen districts and one state-wide program were selected from among 28 districts and one state-wide program that had demonstrably met the three criteria listed above. One other school district agreed to randomly assign some of its eligible schools to Reading First or a control group. The final selection reflected wide variation in district characteristics and provided enough schools to meet the study's sample size requirements. The regression discontinuity sites provide 238 schools for the analysis, and the randomized experimental site provides 10 schools. Half the schools at each site are Reading First schools and half are non-Reading First schools: in three sites, the study sample includes all the RF schools (in that site), in the remaining 15 sites, the study sample includes some, but not all, of the RF schools (in that site).

At the same time, the study deliberately endeavored to obtain a sample that was geographically diverse and as similar as possible to the population of all RF schools. The final study sample of 248 schools, 125 of which are Reading First schools, represents 44 percent of the Reading First schools in their respective sites (at the time the study selected its sample in 2004). The study's sample of RF schools is large, is quite similar to the population of all RF schools, is geographically diverse, and represents states (and districts) that received their RF grants across the range of RF state award dates. The average Year 1 grant for RF schools in the study sample ranged from about $81,790 to $708,240, with a mean of $188,782. This translates to an average of $601 per RF student. Nationally, the median RF grant (based on data reported in the 2004-05 school year) is $138,000 (U.S. Department of Education, 2006). For more detailed information about the selection process and the study sample, see the study's Interim Report (Gamse, Bloom, Kemple & Jacob, 2008).

Data Collection and Outcome Measures

Exhibit 1.2 summarizes the study's three-year, multi-source data collection plan. The present report is based on data for school years 2004-05, 2005-06, and 2006-07. Data collection included student assessments in reading comprehension and decoding, and classroom observations of teachers' instructional practices in reading, teachers' instructional organization and order, and students' engagement with print. Data were also collected through surveys of teachers, reading coaches, and principals, and interviews of district personnel. Sample sizes and response rates for all data collection activities are presented in Exhibit 1.3; see Appendix C for detailed descriptions of the numbers of schools, classrooms, survey respondents, and students included in each separate data collection activity. See Appendix B, Part 5 for a discussion of how missing data were handled.

Exhibit 1.2: Data Collection Schedule for the Reading First Impact Study

Data Collection Elements	2004-2005		2005-2006		2006-2007	
	Fall	Spring	Fall	Spring	Fall	Spring
Student Testing	✓	✓		✓		✓
Stanford Achievement Test, 10th Edition (SAT 10)	✓	✓		✓		✓
Test of Silent Word Reading Fluency (TOSWRF)						✓
Classroom Observations		✓	✓	✓	✓	✓
Instructional Practice in Reading Inventory		✓	✓	✓	✓	✓
Student Time-on-Task and Engagement with Print (STEP)			✓	✓	✓	✓
Global Appraisal of Teaching Strategies (GATS)			✓	✓	✓	✓
Teacher, Principal, Reading Coach Surveys		✓				✓
District Staff Interviews		✓				✓

Exhibit 1.3: Summary of RFIS Data Collection Activities and Respective Response Rates, By Grade

	Fall 2004				Spring 2005				Fall 2005			
	RF		Non-RF		RF		Non-RF		RF		Non-RF	
	N	*(%)*	*N*	*(%)*	*N*	*(%)*	*N*	*(%)*	*N*	*(%)*	*N*	*(%)*
Student assessments (SAT 10)[a]												
Grade 1	5,417	72%	5,139	69%	7,791	84%	7,037	80%				
Grade 2	5,178	71%	4,978	70%	7,519	85%	7,046	82%				
Grade 3	5,281	73%	4,861	69%	7,362	84%	7,014	84%				
Student assessments (TOSWRF)[b]												
Grade 1												
Classroom observations (IPRI)												
Grade 1					809	97%	820	96%	720	98%	704	98%
Grade 2					766	96%	760	95%	664	97%	668	98%
Student engagement with print observations (STEP)[c]												
Grade 1									359	99%	349	99%
Grade 2									324	97%	329	98%
Global Appraisal of Teaching Strategies (GATS)[d]												
Grade 1									359	99%	351	99%
Grade 2									333	99%	335	99%
Surveys												
Grade 1 Teacher					396	73%	363	67%				
Grade 2 Teacher					362	73%	319	65%				
Grade 3 Teacher					318	71%	279	64%				
Reading Coach					118	95%	79	72%				
Principal					98	78%	89	72%				
Site/District Interviews					18	100%	18	100%				

Notes:

[a] In 12 sites, the SAT 10 classroom sample mirrors the observation (and TOSWRF) classroom samples; in the remaining 6 sites, state and district testing requirements meant that all classrooms were tested.

[b] The TOSWRF classroom sample mirrors the classrooms selected for classroom observations.

[c] In each round of two classroom observations, the STEP was administered once while the IPRI was administered twice.

[d] At the conclusion of each IPRI observation (two per classroom), the observer completed a GATS form for the classroom. Information presented here on the GATS was combined to produce a single record per classroom.

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. (See Appendix C for a complete discussion of the observation protocols).

The numbers reported here for SAT 10 student assessments differ from those in Exhibit 3.2 in the Interim Report because the Interim Report incorrectly presented the numbers of students eligible to be tested rather than the number of students tested. Note that the response rates (the number of students tested divided by the number of students eligible to be tested) were correct in Exhibit 3.2 in the Interim Report, and are reproduced here.

EXHIBIT READS: During fall 2004, there were 5,417 student assessments completed in Reading First grade 1 classrooms, corresponding to 72 percent of all eligible student assessments.

Exhibit 1.3: Summary of RFIS Data Collection Activities and Respective Response Rates, By Grade (continued)

	Spring 2006				Fall 2006				Spring 2007			
	RF		Non-RF		RF		Non-RF		RF		Non-RF	
	N	*(%)*	*N*	*(%)*	*N*	*(%)*	*N*	*(%)*	*N*	*(%)*	*N*	*(%)*
Student assessments (SAT 10)[a]												
Grade 1	6,522	86%	5,588	85%					6,954	88%	5,534	85%
Grade 2	6,497	86%	5,596	85%					6,777	90%	5,621	85%
Grade 3	6,254	87%	6,043	87%					6,172	86%	6,117	86%
Student assessments (TOSWRF)[b]												
Grade 1									5,520	87%	5,272	85%
Classroom observations (IPRI)												
Grade 1	718	99%	707	99%	738	100%	703	100%	734	99%	708	99%
Grade 2	666	100%	668	100%	684	99%	672	100%	684	99%	676	100%
Student engagement with print observations (STEP)[c]												
Grade 1	351	97%	347	98%	366	99%	343	97%	361	98%	349	97%
Grade 2	326	97%	330	99%	339	98%	332	99%	341	99%	333	98%
Global Appraisal of Teaching Strategies (GATS)[d]												
Grade 1	358	99%	354	99%	369	99%	352	100%	367	99%	354	99%
Grade 2	334	99%	334	100%	342	99%	336	99%	342	99%	338	99%
Surveys												
Grade 1 Teacher									328	87%	317	88%
Grade 2 Teacher									313	89%	304	87%
Grade 3 Teacher									286	84%	244	74%
Reading Coach									123	99%	105	89%
Principal									104	83%	99	80%
Site/District Interviews									18	100%	18	100%

Notes:

[a] In 12 sites, the SAT 10 classroom sample mirrors the observation (and TOSWRF) classroom samples; in the remaining 6 sites, state and district testing requirements meant that all classrooms were tested.

[b] The TOSWRF classroom sample mirrors the classrooms selected for classroom observations.

[c] In each round of two classroom observations, the STEP was administered once while the IPRI was administered twice.

[d] At the conclusion of each IPRI observation (two per classroom), the observer completed a GATS form for the classroom. Information presented here on the GATS was combined to produce a single record per classroom.

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. (See Appendix C for a complete discussion of the observation protocols).

The numbers reported here for SAT 10 student assessments differ from those in Exhibit 3.2 in the Interim Report because the Interim Report incorrectly presented the numbers of students eligible to be tested rather than the number of students tested. Note that the response rates (the number of students tested divided by the number of students eligible to be tested) were correct in Exhibit 3.2 in the Interim Report, and are reproduced here.

EXHIBIT READS: During spring 2006, there were 6,522 student assessments completed in Reading First grade 1 classrooms, corresponding to 86 percent of all eligible student assessments.

Exhibit 1.4 lists the principal domains for the study, the outcome measures within each domain, and the data sources for each measure.[14] These include:

Student reading performance, assessed with the reading comprehension subtest of the Stanford Achievement Test, 10th Edition (SAT 10, Harcourt Assessment, Inc., 2004). The SAT 10 was administered to students in grades one, two and three during fall 2004, spring 2005, spring 2006, and spring 2007, with an average completion rate of 83 percent across all administrations. In the spring of 2007 only, first grade students were assessed with the Test of Silent Word Reading Fluency (TOSWRF, Mather et al., 2004), a measure designed to assess students' ability to decode words from among strings of letters. The average completion rate was 86 percent. Three outcome measures of student reading performance were created from SAT 10 and TOSWRF data.

Individualized student testing on all five essential components of reading skill emphasized by Reading First was not conducted due to concerns about cost as well as about the burden of study data collection on schools and students. The study team selected reading comprehension as the central reading achievement construct for the study, recognizing that the other four essential components would not be assessed. The selection of reading comprehension reflected its importance as the "essence of reading" that sets the stage for children's later academic success (National Institute of Child Health and Human Development, 2000). The SAT 10 reading comprehension subtest chosen is feasible in group-administered settings and on a large scale, and this test was already being used by some study sites, which reduced the burden on schools and students.

Midway through the evaluation, the study team, in conjunction with IES, decided to add a test of skills that precede comprehension. The study added a decoding test to assess whether the Reading First program had an effect on this skill. Resources were insufficient to expand the data collection into all grades. Because the programmatic emphasis on decoding skill was hypothesized to be more intensive in first grade, the study added the Test of Silent Word Reading Fluency only in first grade.

Classroom reading instruction, 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 during scheduled reading blocks in each sampled classroom on two consecutive days during each wave of data collection: spring 2005, fall 2005 and spring 2006, and fall 2006 and spring 2007. The average completion rate was 98 percent across all years. The IPRI 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 used a booklet containing a series of individual IPRI forms, each of which corresponds to a three-minute interval of observation. The average reading block based on observational data was 108 minutes. Eight outcome measures of classroom instruction were created from IPRI data to represent the components of reading instruction emphasized by the Reading First legislation.[15]

[14] Appendix C presents more detailed information, including (where applicable) copies of measures developed specifically for the RFIS.

[15] 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.

Exhibit 1.4: Description of Domains, Outcome Measures, and Data Sources Utilized in the Reading First Impact Study

Domain	Outcome Measure and Description	Source
Student reading performance	**Mean scaled scores for 1st, 2nd, and 3rd grade students**, 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.	*Stanford Achievement Test, 10th Edition (SAT 10)*
	Percentage of 1st, 2nd, and 3rd grade students at or above grade level, 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.	*Stanford Achievement Test, 10th Edition (SAT 10)*
	Mean standard scores for 1st grade students, represented as a continuous measure of first grade students' decoding skill.	*Test of Silent Word Reading Fluency*
Classroom reading instruction	**Minutes of instruction in phonemic awareness**, or how much instructional time 1st and 2nd grade teachers spent on phonemic awareness.	*RFIS Instructional Practice in Reading Inventory*
	Minutes of instruction in phonics, or how much instructional time 1st and 2nd grade teachers spent on phonics.	*RFIS IPRI*
	Minutes of instruction in fluency building, or how much instructional time 1st and 2nd grade teachers spent on fluency building.	*RFIS IPRI*
	Minutes of instruction in vocabulary development, or how much instructional time 1st and 2nd grade teachers spent on vocabulary development.	*RFIS IPRI*
	Minutes of instruction in comprehension, or how much instructional time 1st and 2nd grade teachers spent on comprehension of connected text.	*RFIS IPRI*
	Minutes of instruction in all five dimensions combined, or how much instructional time 1st and 2nd grade teachers spent on all five dimensions combined.	*RFIS IPRI*
	Proportion of each observation with highly explicit instruction, or the proportion of time spent within the five dimensions when teachers used highly explicit instruction (e.g., instruction included teacher modeling, clear explanations, and the use of examples).	*RFIS IPRI*
	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 (e.g., teachers asked students to practice such word learning strategies as context, word structure, and meanings).	*RFIS IPRI*
Student engagement with print	**Percentage of 1st and 2nd grade students engaged with print**, represented as the per-classroom average of the percentage of students engaged with print across three sweeps in each classroom during observed reading instruction.	*RFIS Student Time-on-Task and Engagement with Print (STEP)*

Exhibit 1.4: Description of Domains, Outcome Measures, and Data Sources Utilized in the Reading First Impact Study (continued)

Domain	Outcome Measure and Description	Source
Professional development in scientifically based reading instruction	**Amount of PD in reading received by teachers**, or teachers' self-reported number of hours of professional development in reading during 2006-07.	*RFIS Teacher Survey*
	Teacher receipt of PD in the five essential components of reading instruction, or the number of essential components teachers reported were covered in professional development they received during 2006-07.	*RFIS Teacher Survey*
	Teacher receipt of coaching, or whether or not a teacher reported receiving coaching or mentoring from a reading coach in reading programs, materials, or strategies in 2006-07.	*RFIS Teacher Survey*
	Amount of time dedicated to serving as K-3 reading coach, or reading coaches' self-reported percentage of time spent as the K-3 reading coach for their school in 2006-07.	*RFIS Reading Coach Survey*
Amount of reading instruction	**Minutes of reading instruction per day**, or teachers' reported average amount of time devoted to reading instruction per day over the prior week.	*RFIS Teacher Survey*
Supports for struggling readers	**Availability of differentiated instructional materials for struggling readers**, or whether or not schools reported that specialized instructional materials beyond the core reading program were available for struggling readers.	*RFIS Reading Coach and Principal Surveys*
	Provision of extra classroom practice for struggling readers, or the number of dimensions in which teachers reported providing extra practice opportunities for struggling students in the past month.	*RFIS Teacher Survey*
Use of assessments	**Use of assessments to inform classroom practice**, or the number of instructional purposes for which teachers reported using assessment results.	*RFIS Teacher Survey*

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 only one instructional behavior/activity was observed, that interval was designated accordingly. In 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.

Student engagement with print. Beginning in fall 2005, the study conducted classroom observations using the Student Time-on-Task and Engagement with Print (STEP) instrument to measure the percentage of students engaged in academic work that are reading or writing print. The STEP was used to observe classrooms in fall 2005, spring 2006, fall 2006, and spring 2007, with an average completion rate of 98 percent across all years. 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 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.

Professional development in scientifically based reading instruction, amount of reading instruction, supports for struggling readers, and use of assessments. Within these four domains, eight outcome measures were created based on data from surveys of principals, reading coaches, and teachers about school and classroom resources. The eight outcome measures represent aspects of scientifically based reading instruction promoted by the Reading First legislation and guidance. Surveys were fielded in spring 2005 and again in spring 2007 with an average completion rate across all respondents of 73 percent in spring 2005 and 86 percent in spring 2007. This final report includes findings from 2007 surveys only.

Additional data were collected by the study team in order to create measures used in correlational analyses. These data include:

The ***Global Appraisal of Teaching Strategies (GATS)***, a 12-item checklist designed to measure teachers' instructional strategies related to overall instructional organization and order, is adapted from "The Checklist of Teacher Competencies" (Foorman and Schatschneider, 2003). Unlike the IPRI, which focuses on discrete teacher behaviors, the GATS was designed to capture global classroom management and environmental factors. Items covered topics such as the teacher's organization of materials, lesson delivery, responsiveness to students, and behavior management. The GATS was completed by the classroom observer immediately after each IPRI observation, meaning that each sampled classroom was rated on the GATS twice in the fall and twice in the spring in both the 2005-2006 school year and the 2006-2007 school year. The GATS was fielded in fall 2005, spring 2006, fall 2006, and spring 2007, with an average completion rate of over 99 percent. A single measure from the GATS data was created for use in correlational analyses.

Study's Methodological Approach

This section summarizes key features of the study's methodological approach, including use of multi-level models, determination of statistical significance, and multiple hypothesis testing. More detailed information about the study's approach is presented in Appendix B.

Approach to Estimating Impacts

As described in detail in Appendix B, and in the study's Interim Report, all impact estimates are regression-adjusted to control for (1) a linear specification of each site's specific rating variable for selecting Reading First schools, and (2) selected student background characteristics used in the analysis (Gamse, Bloom, Kemple, & Jacob, 2008).[16] The impacts have been estimated using multi-level models to account for the clustering of students within classrooms, classrooms within schools, and schools within

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

sites. Throughout this report, tables that display impact estimates present values in the "Actual Mean with Reading First" column that are actual, unadjusted values for Reading First schools. The 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 these are calculated by subtracting the impact estimates from the RF schools' actual mean values.

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. Conversely, 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.

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."[17] As a further safeguard, composite hypothesis tests were used to assess the overall statistical significance for groups of impact estimates within the core outcome domains described in Exhibit 1.4: student reading performance, classroom reading instruction, student engagement with print, professional development in SBRI, amount of reading instruction, supports for struggling readers, and use of assessments. 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 components might be due to chance. The composite tests therefore help to "qualify," or call into question, statements that are based on individual findings.[18]

[17] Researchers disagree about whether and how to account for multiple hypothesis testing (e.g., Gelman and Stern, 2006; Schochet, 2008; Shaffer, 1995).

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

Roadmap to this Report

Chapter Two addresses the study's first two evaluation questions about impacts on instruction and on reading achievement for the study sites. Chapter Three presents the results of several exploratory analyses, pertaining to variation in impacts and relationships among instructional practices and student reading comprehension (in response to the study's third research question).

Chapter Two: Impact Findings

This chapter addresses the study's first two evaluation questions pertaining to Reading First impacts on classroom reading instructional practices and reading comprehension test scores. The core impact results are averaged across the study's 18 sites and pooled across the 2004-05, 2005-06, and 2006-07 school years. The study pools estimates both to improve statistical power and to be more parsimonious with respect to findings. The differences in impacts among the three years are not statistically significant for data collected in all three years. (Appendix E presents impact estimates separately for each follow-up year.) In addition, the chapter presents Reading First impacts on measures administered in the spring of 2007: a measure of students' decoding skills administered to first graders and surveys administered to educational personnel.[19] As noted in Chapter One, all tables that display impact findings present values in the "Actual Mean with Reading First" column that are actual, unadjusted values for Reading First schools. The values in the "Estimated Mean without Reading First" column represent the best estimates of what would have happened in RF schools absent RF funding. Impact estimates are regression-adjusted to control for a linear specification of the rating variable used by sites to select Reading First schools. Estimates were obtained from multi-level statistical models that account for the clustering of students within classrooms, classrooms within schools, and schools within sites.[20] Impacts were estimated for each study site and then averaged across sites in proportion to their number of Reading First schools in the study sample.

Average Impacts on Reading Instruction

Exhibits 2.1, 2.2, and 2.3 present estimated impacts on classroom reading instruction and student engagement with print. These estimates are based on data from classroom observations conducted in the 18 study sites during the 2004-05, 2005-06, and 2006-07 school years.

- Reading First produced a statistically significant positive impact on the total time that teachers spent on the five essential components of reading instruction promoted by the program.

Exhibit 2.1 indicates that first- and second-grade teachers in Reading First schools spent 59 minutes, on average, during the approximately 112 minutes of the average daily reading block teaching phonemic awareness, phonics, vocabulary, fluency and/or comprehension.[21] This reflects a program impact of 6.9 additional minutes per daily reading block in grade one and 9.8 additional minutes per daily reading block in grade two. Over the course of a week, this represents an additional 35 minutes for grade one and 49 minutes for grade two.

[19] Appendix D presents 95 percent confidence intervals for main impacts in relevant metrics as well as effect sizes. Confidence intervals for estimated impacts are reported for reading comprehension, decoding, instructional outcomes, and student engagement with print.

[20] See Appendix B for a discussion of the study's approach to estimating impacts.

[21] The number of minutes of reading instruction used in impact analyses is based on observational data, which differs slightly from number of minutes reported on surveys.

Exhibit 2.1: Estimated Impacts on Instructional Outcomes: 2005, 2006, and 2007 (pooled)

Construct	Actual Mean With Reading First	Estimated Mean Without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Grade 1					
Minutes of instruction in the five dimensions combined	59.23	52.31	6.92*	0.33*	(0.005)
Percentage of intervals in five dimensions with highly explicit instruction	29.39	26.10	3.29*	0.18*	(0.018)
Percentage of intervals in five dimensions with High Quality Student Practice	18.44	17.61	0.82	0.05	(0.513)
Grade 2					
Number of minutes of instruction in the five dimensions combined	59.08	49.30	9.79*	0.46*	(<0.001)
Percentage of intervals in five dimensions with highly explicit instruction	30.95	27.95	3.00*	0.16*	(0.040)
Percentage of intervals in five dimensions with High Quality Student Practice	17.82	14.88	2.94*	0.16*	(0.019)

NOTES:

The complete Reading First Impact 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 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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: The observed mean amount of time spent per daily reading block in instruction in the five dimensions combined for first grade classrooms with Reading First was 59.23 minutes. The estimated mean amount of time without Reading First was 52.31 minutes. The impact of Reading First on the amount of time spent in instruction in the five dimensions combined was 6.92 (or 0.33 standard deviations), which was statistically significant (p=.005).

SOURCES: RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, spring 2006, fall 2006, and spring 2007

- Reading First produced a statistically significant positive impact on the use of highly explicit instruction in grades one and two, and a statistically significant increase in the amount of high quality student practice in grade two. Its estimated impact on high quality student practice for grade one was not statistically significant.

For first-grade classrooms in Reading First schools, 29 percent of the observation intervals with instruction in the five dimensions also involved highly explicit instruction (active teaching, modeling or explaining concepts, and helping children to use reading strategies). This average was 31 percent for second-grade classrooms. These findings represent a program impact of 3.29 percentage points for first grade and 3.00 percentage points for second grade.

For first-grade and second-grade classrooms in Reading First schools, approximately 18 percent of the observation intervals that included instruction in the five dimensions also involved high quality student practice (component-specific opportunities for students to practice their skills). These findings represent a

Exhibit 2.2: Estimated Impacts On the Number of Minutes in Instruction in Each of the Five Dimensions of Reading: 2005, 2006, and 2007 (pooled)

Number of minutes of instruction in:	Actual Mean With Reading First	Estimated Mean Without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Grade 1					
Phonemic Awareness	2.32	1.71	0.61*	0.23*	(0.030)
Phonics	21.32	18.45	2.86*	0.21*	(0.048)
Vocabulary	7.92	7.35	0.57	0.09	(0.386)
Fluency	4.67	3.43	1.24*	0.20*	(0.043)
Comprehension	23.01	21.23	1.78	0.12	(0.247)
Grade 2					
Phonemic Awareness	0.49	0.37	0.12	0.10	(0.319)
Phonics	13.92	10.65	3.27*	0.31*	(0.006)
Vocabulary	11.79	10.06	1.73*	0.20*	(0.036)
Fluency	4.14	3.56	0.58	0.11	(0.297)
Comprehension	28.74	24.73	4.01*	0.24*	(0.019)

NOTES:

The complete Reading First Impact 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 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; statistically significant findings at the $p \leq .05$ level are indicated by *.

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 2.32 minutes. The estimated mean amount of time without Reading First was 1.71 minutes. The impact of Reading First on the amount of time spent in instruction in phonemic awareness was 0.61 minutes (or 0.23 standard deviations), which was statistically significant (p=.030).

SOURCES: RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, spring 2006, fall 2006 and spring 2007

Exhibit 2.3: Estimated Impacts on the Percentage of Students Engaged with Print: 2006 and 2007

Construct	Actual Mean with Reading First	Estimated Mean without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Grade 1					
Percentage of students engaged with print					
Pooled (SY 2006, SY 2007)	47.84	42.52	5.33	0.18	(0.070)
Grade 2					
Percentage of students engaged with print					
Pooled (SY 2006, SY 2007)	50.53	55.27	-4.75	-0.17	(0.104)

NOTES:

The complete Reading First Impact 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 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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: For the 2006 and 2007 school years pooled, the actual average percentage of students engaged with print in first grade classrooms with Reading First was 47.84 percent. The estimated average percentage without Reading First was 42.52 percent. The impact of Reading First on the average percentage of student engagement with print was 5.33 percentage points (or 0.18 standard deviations), which was not statistically significant (p=.070).

SOURCE: RFIS Student Time-on-Task and Engagement with Print, fall 2005, spring 2006, fall 2006, and spring 2007

program impact of 2.94 percentage points for second grade and 0.82 percentage points for first grade (which was not statistically significant).

A composite test of the six impact estimates in Exhibit 2.1 was conducted by combining its three measures into one index and pooling the data for grades one and two. (See Appendix B, Exhibit B.7). This test indicates a statistically significant overall impact of Reading First on instructional practice.

Exhibit 2.2 presents separate impact estimates for each of the five Reading First instructional dimensions, illustrating the relative emphasis placed by Reading First schools on each dimension, how this emphasis differs by grade, and how Reading First impacts are distributed across the dimensions. The majority of Reading First instructional time focused on comprehension and phonics, and half of the program's statistically significant instructional impacts were on these two dimensions.

- First grade teachers in Reading First schools spent about 21.3 minutes on phonics and 23.0 minutes on comprehension per daily reading block. This reflects an estimated daily impact of 2.9 additional minutes for phonics (statistically significant) and 1.8 additional minutes for comprehension (not statistically significant). Although first grade teachers in Reading First schools spent relatively little time on phonemic awareness (an average of 2.3 minutes per

reading block) and fluency (4.7 minutes), program impacts on these dimensions were positive and statistically significant.

- Second grade teachers in Reading First schools spent 13.9 minutes on phonics and 28.7 minutes on comprehension per daily reading block. This reflects statistically significant impacts of 3.3 minutes for phonics and 4.0 minutes for comprehension. Reading First also produced a statistically significant impact on vocabulary instruction of 1.7 minutes per daily reading block.

Average Impacts on Student Engagement with Print

Exhibit 2.3 presents estimated impacts on the percentage of students engaged with print during observations of reading instruction within the reading block. The measure of student engagement with print used in impact analyses is the per-classroom average of the percentage of students engaged with print across three observation sweeps in each classroom.

Approximately 48 percent of first grade students and 51 percent of second grade students in Reading First schools were engaged with print during observations of reading instruction within the reading block. The estimated impact on student engagement with print was not statistically significant for grade one (5.33 percentage points) or grade two (-4.75 percentage points).

Exhibit 2.3 includes two statistical tests of program impacts on the percentage of students engaged with print, one for each grade. A composite test was conducted that pools findings across grades; it was not statistically significant. (See Appendix B, Exhibit B.7).

Average Impacts on Key Components of SBRI

The section below draws from self reported survey data collected at both the school level (surveys of principals and reading coaches) and the classroom level (teacher surveys)[22] to assess the extent to which components of scientifically based reading instruction (SBRI) have been implemented in study schools. Data on such school and classroom level practices can provide information about the levels of these practices and whether Reading First has had an impact on them.

Exhibit 2.4 lists eight outcome measures that represent four domains—professional development in SBRI, amount of reading instruction, supports for struggling readers, and use of assessments. Two outcome measures are at the school-level and six outcome measures are at the classroom-level.[23] For each measure, RDD estimation methods were used to determine if statistically significant differences exist between the treatment and comparison groups.

[22] This section reports on 2007 survey findings only.

[23] See Appendix C for a detailed description of the eight survey outcome variables, including the survey items, the item metrics, the outcome specifications, and the internal consistency reliability (as applicable).

Exhibit 2.4: Estimated Impacts on Key Components of Scientifically Based Reading Instruction (SBRI): Spring 2007

Domain	Actual Mean With Reading First	Estimated Mean Without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Professional Development (PD) in SBRI					
Amount of PD in reading received by teachers (hours) [a]	25.84	13.71	12.13*	0.51*	(<0.001)
Teacher receipt of PD in the five essential components of reading instruction (0-5) [a]	4.30	3.75	0.55*	0.31*	(0.010)
Teacher receipt of coaching (proportion) [a]	0.83	0.63	0.20*	0.41*	(<0.001)
Amount of time dedicated to serving as K-3 reading coach (percent) [b,c]	91.06	57.57	33.49*	1.03*	(<0.001)
Amount of Reading Instruction					
Minutes of reading instruction per day [a]	105.71	87.24	18.47*	0.63*	(<0.001)
Supports for Struggling Readers					
Availability of differentiated instructional materials for struggling readers (proportion) [b]	0.98	0.97	0.01	0.15	(0.661)
Provision of extra classroom practice for struggling readers (0-4) [a]	3.79	3.59	0.19*	0.20*	(0.018)
Use of Assessments					
Use of assessments to inform classroom practice (0-3) [a]	2.63	2.45	0.18	0.19	(0.090)

NOTES:

[a] Classroom level outcome

[b] School level outcome

[c] The response rates for RF and nonRF reading coach surveys were statistically significantly different (p=0.037). Reading first schools were more likely to have had reading coaches and to have returned reading coach surveys.

[d] Missing data rates ranged from 0.1 to 3.3 percent for teacher survey outcomes (RF: 0.1 to 1.0 percent; non-RF: 0 to 4.9 percent) and 1.3 to 2.8 percent for reading coach and/or principal survey outcomes (RF: 0 to 1.6 percent; non-RF: 2.7 to 4.1 percent). Survey constructs (i.e., those outcomes comprised of more than one survey item) were computed only for observations with complete data, with one qualification: for the construct "minutes spent on reading instruction per day," the mean was calculated as the total number of minutes reported for last week (over a maximum of 5 days) divided by the number of days with non-missing values. Less than one percent of teachers (0.9 percent) were missing data for all 5 days.

The complete Reading First Impact 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 effect size of the impact is the impact divided by the actual standard deviation of the outcome for the non-Reading First Schools.

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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: The observed mean amount of professional development in reading received by teachers with Reading First was 25.84 hours. The estimated mean amount of professional development in reading received by teachers without Reading First was 13.71 hours. This impact of 12.13 hours was statistically significantly (p<.001).

SOURCES: RFIS, Teacher, Reading Coach, and Principal Surveys, spring 2007

Exhibit 2.4 indicates that Reading First had a significant impact on the amount, content, and type of professional development received by teachers in grades one through three, according to teacher and reading coach self-reports. More specifically, there were statistically significant impacts on all four outcome measures in the domain of professional development in SBRI:

- Reading First had a statistically significant impact on the amount of professional development in reading teachers reported receiving; this impact was 12.1 hours.
- Reading First had a statistically significant impact on teachers' self-reported receipt of professional development in the five essential components of reading instruction. Teachers in RF schools reported receiving professional development in an average of 4.3 components, 0.6 components more than would be expected without Reading First (3.7 components).
- A statistically significantly greater proportion (20 percent) of teachers in RF schools reported receiving coaching from a reading coach than would be expected without Reading First.
- Reading First had a statistically significant impact on the amount of time reading coaches reported spending in their role as the school's reading coach. Reading coaches in RF schools reported spending 91.1 percent of their time in this role, 33.5 percentage points more than would be expected without Reading First (57.6 percent).

Reading First had a statistically significant impact on the amount of time teachers reported spending on reading instruction per day. Teachers in RF schools reported an average of 105.7 minutes per day, 18.5 minutes more than would be expected without Reading First (87.2 minutes).

Reading First had mixed impacts on the availability of supports for struggling readers.

- Reading First had a statistically significant impact on teachers' provision of extra classroom practice in the essential components of reading instruction in the past month; the estimated impact was 0.2 components.
- There was no statistically significant impact of Reading First on the availability of differentiated instructional materials for struggling readers.

There was no statistically significant impact of Reading First on the teachers' reported use of assessments to inform classroom practice for grouping, diagnostic, and progress monitoring purposes.

To assess the overall impact of Reading First on these survey items, two composite tests were conducted. The first composite test combined the two outcome measures from the reading coach and/or principal survey data into a single school-level index; the second composite test combined the six outcome measures from the teacher survey data into a single classroom-level index (See Appendix B, Exhibit B.7). These tests indicate a statistically significant overall impact of Reading First on the implementation of scientifically based reading instruction both at the school-level and the classroom-level.

In conclusion, estimated impacts based on survey data from RF and non-RF schools in the study sample indicate that statistically significant impacts of Reading First are evident in six of the eight outcome measures, including the four outcome measures in the *professional development in SBRI* domain, the single outcome measure in the *amount of reading instruction* domain, and one of two outcome measures in the *supports available for struggling readers* domain. There was no statistically significant impact of

RF in the *use of assessments* domain. These data indicate that RF schools are consistently reporting higher levels of implementation of SBRI practices than would have occurred absent RF.

Average Impacts on Reading Achievement

Average Impacts on Reading Comprehension

Exhibit 2.5 presents estimated Reading First impacts on student reading comprehension scores on the SAT 10. These findings reflect impact estimates that are averaged across the 18 study sites and pooled across the three study follow-up years (2004-2005, 2005-2006, and 2006-2007). Impact estimates are regression-adjusted to control for a linear specification of the rating variable used by sites to select Reading First schools and for selected school and student background characteristics. Estimates were obtained from multi-level statistical models that account for the clustering of students within classrooms, classrooms within schools, and schools within sites. Impacts were estimated for each study site and then averaged across sites in proportion to their number of Reading First schools in the study sample.

- Impacts on student reading comprehension test scores were not statistically significant.

Estimated impacts were not statistically significant for grade one (4.7 scaled score points or an effect size of 0.10 standard deviations), grade two (1.7 scaled score points or an effect size of 0.04 standard deviations), or grade three (0.3 scaled score points or an effect size of 0.01 standard deviations).[24] The average first, second, and third grade student in Reading First schools was reading at the 44th, 39th, and 39th percentile, respectively, on the end-of-the-year assessment (on average over the three years of data collection).

Exhibit 2.5 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. (See Appendix B, Exhibit B.7).[25]

Average Impacts on Decoding Skills for Students in Grade One in Spring 2007

For the final year of data collection, first grade students were also assessed with the Test of Silent Word Reading Fluency (TOSWRF, Mather et al., 2004). The TOSWRF is a short three-minute assessment that measures students' ability to identify words quickly and correctly. This assessment was added to explore whether Reading First has an impact on decoding skills, another of the five components of reading skill targeted by Reading First (along with comprehension, vocabulary, phonemic awareness, and fluency). The assessment was added in the last year of the study's data collection, which means that the TOSWRF

[24] The study also examined third grade reading achievement scores on state-required assessments for the core sample for 2006 scores only (excluding one site that had no third grade assessment and another site that did not use a percent proficient metric). These results are shown in Appendix E, Part 3. The results are consistent with the Grade Three results for the SAT 10.

[25] For technical reasons, the index used in the composite test for student reading performance includes only the two SAT 10 measures for which data are available across grades. The TOSWRF could not be included in the index because data were only available for one grade.

Exhibit 2.5: Estimated Impacts on Reading Comprehension: Spring 2005, 2006, and 2007 (Pooled)

Construct	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 Scaled Score					
Grade 1					
Scaled Score	543.8	539.1	4.7	0.10	(0.083)
Corresponding Grade Equivalent[a]	1.7	1.7			
Corresponding Percentile	*44*	*41*			
Grade 2					
Scaled Score	584.4	582.8	1.7	0.04	(0.462)
Corresponding Grade Equivalent[a]	*2.5*	2.4			
Corresponding Percentile	*39*	*38*			
Grade 3					
Scaled Score	609.1	608.8	0.3	0.01	(0.887)
Corresponding Grade Equivalent[a]	*3.3*	*3.3*			
Corresponding Percentile	*39*	*39*			
Panel 2					
All Sites					
Percent Reading At or Above Grade Level[b]					
Grade 1	46.0	41.8	4.2		(0.104)
Grade 2	38.9	37.3	1.6		(0.504)
Grade 3	38.7	38.8	-0.1		(0.973)

NOTES:

The complete Reading First Impact 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one 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).

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; statistically significant findings at the $p \leq .05$ level are indicated by *.

[a] Grade equivalent scores are based on a nine-month school year, are reported in decimal format (year.month), and provide an estimate of the performance that an average student at a grade level is assumed to demonstrate on the test at a particular month in the school year. For example, a score of 1.7 represents a performance level typical of a first grade student in the seventh month of the school year.

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

EXHIBIT READS: The observed mean reading comprehension score for first-graders with Reading First was 543.8 scaled score points. The estimated mean without Reading First was 539.1 scaled score points. The impact of Reading First was 4.7 scaled score points (or 0.10 standard deviations), which was not statistically significant (p=.083). The observed average percent of first-graders reading at or above grade level with Reading First was 46.0 percentage points. The estimated average percent without Reading First was 41.8 percentage points. The impact of Reading First on the percent of first grade students reading at or above grade level was 4.2 percentage points, which was not statistically significant (p=.104).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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).

was administered to first grade students only once in the spring of 2007. Thus, unlike the reading comprehension impact estimates, which are available for grades one, two, and three, and pooled across three school years, the decoding results reflect only one of the three follow up years of data collection and are available for only grade one.

Exhibit 2.6 summarizes findings from an analysis of Reading First's impact on TOSWRF scores for first grade students in spring 2007.

- Reading First produced a statistically significant positive impact on TOSWRF scores of 2.5 standard score points, equal to an effect size of 0.17 standard deviations.

Exhibit 2.6: Estimated Impacts of Reading First on Decoding Skill: Grade One, Spring 2007

	Actual Mean with Reading First	Estimated Mean without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Decoding Skill					
Standard Score	96.9	94.4	2.5 *	0.17 *	(0.025)
Corresponding Grade Equivalent[a]	1.7	1.4			
Corresponding Percentile	42	35			

NOTES:

The Test of Silent Word Reading Fluency (TOSWRF) sample includes first-graders in 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 from spring 2007 TOSWRF test scores (1st grade).

The key metric for the TOSWRF analyses is the standard score, corresponding grade equivalents and percentiles are provided for reference. Although the publisher of the Test of Silent Word Reading Fluency states that straight comparisons between standard scores and grade equivalents will likely yield discrepancies due to the unreliability of the grade equivalents, they are provided because program criteria are sometimes based on grade equivalents.

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; statistically significant findings at the $p \leq .05$ level are indicated by *.

[a] Grade equivalent scores are based on a nine-month school year, are reported in decimal format (year.month), and provide an estimate of the performance that an average student at a grade level is assumed to demonstrate on the test at a particular month in the school year. For example, a score of 1.7 represents a performance level typical of a first grade student in the seventh month of the school year.

EXHIBIT READS: The observed mean silent word reading fluency standard score for first-graders with Reading First was 96.9 standard score points. The estimated mean without Reading First was 94.4 standard score points. The impact of Reading First was 2.5 standard score points (or 0.17 standard deviations), which was statistically significant (p=.025).

SOURCES: RFIS TOSWRF administration in spring 2007

Summary

The findings presented in this chapter are generally consistent with findings presented in the study's Interim Report, which found statistically significant impacts on instructional time spent on the five essential components of reading instruction promoted by the program (phonemic awareness, phonics, vocabulary, fluency, and comprehension) in grades one and two, and which found no statistically significant impact on reading comprehension as measured by the SAT 10. In addition to data on the instructional and student achievement outcomes reported in the Interim Report, the final report also presents findings based upon information obtained during the study's third year of data collection: data from a measure of first grade students' decoding skill and data from self-reported surveys of educational personnel in study schools.

The additional data sources provide more information about the contexts within which the Reading First program has operated. The information obtained from the Test of Silent Word Reading Fluency indicates that Reading First had a positive and statistically significant impact on first grade students' decoding skill. Through surveys, Reading First school personnel reported implementing the key programmatic components outlined in the enabling legislation.

A frequent criticism of the interim report was that the scientifically based reading practices promoted by Reading First have been diffused to non-Reading First schools, thus diluting the impact of Reading First (see, for example, the response to the Interim Report by the Reading First Federal Advisory Committee, 2008). States could reserve up to 20 percent of their Reading First funds to support staff development, technical assistance to districts and schools, and planning, administration, and reporting. According to the program guidance, this funding provided "states with the resources and opportunity…to improve instruction beyond the specific districts and schools that receive Reading First subgrants" (U.S. Department of Education, 2002).

The results from both observational and survey data indicate that Reading First produced statistically significant impacts on instruction and reading program implementation. These differences are inconsistent with the view that the treatment had diffused to the extent that diffusion means that practices were the same in RF and non-RF schools. However, there are no data available on reading practices in study schools prior to Reading First implementation. Thus, the study cannot provide a definitive statement as to the presence or absence of diffusion.

Chapter Three: Exploratory Analyses of Variations in Impacts and Relationships among Outcomes

The Reading First Impact Study was designed to test the impact of the receipt of Reading First funds at the school level. The study was conducted in 248 schools located in 18 sites in 13 states. The study focused on student reading achievement, as well as teachers' classroom reading practices. Analyses of impact were conducted for data collected during three school years (2004-05, 2005-06, and 2006-07), representing between one and four years of program implementation, depending on the site.

The results reported in Chapter Two indicate that the receipt of Reading First funding at the school level produced an impact on the amount of time teachers spent on the five components of reading instruction promoted by the program and on first graders' decoding skills, but not on student reading comprehension. The sections below describe exploratory analyses that examine some hypotheses about factors that might account for the observed pattern of impacts. The results are based on analyses of subgroups of students, schools, grade levels, and/or years of data collection. The information provides possible avenues for further exploration or for improving Reading First or programs like Reading First. Because the study was not designed to provide a rigorous test of the hypotheses explored in this chapter, the results are only suggestive. The methodological literature about subgroup analyses highlights the importance of specifying hypotheses in advance, limiting the number of additional tests, and interpreting results with considerable caution. (See, for example, Hernandez, Boersma, Murray, Steyerberg, 2006; Rothwell, 2005; Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J., & Drazen, J.M., 2007).

The first section of this chapter examines variation in impacts. The second section examines the relationship between classroom reading instruction and student achievement.

Variation in Impacts

The core impact analyses reported in Chapter Two are average impacts, meant to represent the impact for the average Reading First school in the sample. It is reasonable to wonder whether these overall averages might be masking differences in impacts that could be attributed to variation in: 1) time of RF implementation; 2) student exposure to RF; or 3) sites. The following section explores these hypotheses.

Variation in Impacts Over Time

This section explores the question of whether the impact estimates presented in Chapter Two—which are pooled across three school years—may be masking changes in impacts over time.[26]

Three approaches were used to address the question of possible changes in impacts over time. First, we examined estimated impacts on instructional and reading comprehension outcomes for each year of the study (and pooled) at a given grade level. Next, we conducted two types of statistical tests. The first test, which is a more restrictive test, assessed whether there was a linear trend (year-to-year change) of impacts

[26] Additional analyses of student achievement trends for the RFIS study sample, including patterns of mean SAT 10 scores in grades one through three and state-mandated reading assessments in grade three, are presented in Appendix E.

over time for successive cohorts of first, second, and third graders (if applicable). The second test, a global F-test, assessed whether there was any overall variation in the impacts over the study years for a given grade level. If inconsistencies in statistical significance were found between these two tests, then the results of either test were interpreted with caution.

For instructional time in the five dimensions combined, Exhibit 3.1 indicates that when impacts are estimated separately for each grade and year, those impacts decrease over time for each grade. [27] For example, for minutes of instruction in the five dimensions combined in Grade One, the impact was 8.89 minutes per reading block in Spring 2005, 8.71 minutes in School Year 2006, 5.92 minutes in School year 2007, and 6.92 minutes for all years pooled. The first statistical test of a linear time trend for these impacts suggests a statistically significant annual decline in impacts on time in the five dimensions of 2.6 minutes per daily reading block for grade one and 2.9 minutes per daily reading block for grade two (Exhibit 3.2). However, the second global F-test for each grade of the null hypothesis of no variation across three years suggests that the variation for grade one was not statistically significant while the variation for grade two was statistically significant (Exhibit 3.2). Thus, readers should be particularly cautious when inferring a systematic pattern of decline in impacts on time in the five dimensions for first grade. At the same time, it does appear that the decline in impacts on time in the five dimensions for second grade was more systematic.

Findings for reading comprehension scores, estimated separately for each grade and year, suggest that impacts increased over time for each grade (Exhibit 3.3). For example, in Grade One, the impact was 2.2 scaled score points in Spring 2005, 5.3 scaled score points in Spring 2006, 7.5 scaled score points in Spring 2007, and 4.7 scaled score points for all years pooled. The first statistical test of a linear trend for impacts suggests that only for grade three was there a statistically significant increase. Estimates of a linear impact trend for all three grades pooled indicate a statistically significant increase of 2.5 scaled score points per year (Exhibit 3.2). However, the global F-test of the null hypothesis of no variation across three years was not statistically significant for any grade (Exhibit 3.2). Thus, readers should be cautious about inferring a systematic pattern of increasing impacts over time on reading comprehension.

In sum, these analyses do not provide conclusive support for the hypothesis that the core impact estimates presented in Chapter Two are masking variation in impacts over time in either reading instruction in grade one or in student reading comprehension in grades one, two or three. For reading instruction, there appears to be a systematic decline in impacts in grade two.

Variation in Impacts on Reading Comprehension Associated with Student Exposure to Reading First Schools

Reading First is intended to provide students with a complete instructional program from kindergarten through third grade. However, because of student mobility and the coincident timing of both the start of the program and of the study, many students in the study sample may not have experienced the fullest exposure possible (four full school years, K through 3) to Reading First instructional practices and support services. For example, in the group of study sites that began implementing RF in 2004-2005, third

[27] These same analyses were also conducted for each dimension separately (phonemic awareness, phonics, vocabulary, fluency, and comprehension) and results are presented in Appendix E, Exhibits E.1 and E.2. Results of these analyses for the STEP are also presented in Appendix E, Exhibit E.3.

Exhibit 3.1: Estimated Impacts on Instructional Outcomes: 2005, 2006, and 2007, and Pooled

Construct	Actual Mean With Reading First	Estimated Mean Without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Grade 1					
Minutes of instruction in the five dimensions combined					
Spring 2005	59.23	50.34	8.89*	0.43*	(0.007)
School year 2006	59.49	50.78	8.71*	0.42*	(0.010)
School year 2007	58.93	53.00	5.92	0.28	(0.050)
Pooled 3 years (Sp05, Sy06, Sy07)	59.23	52.31	6.92*	0.33*	(0.005)
Percentage of intervals in five dimensions with highly explicit instruction					
Spring 2005	29.71	22.38	7.33*	0.41*	(0.003)
School year 2006	29.76	27.90	1.86	0.10	(0.326)
School year 2007	28.73	25.90	2.83	0.16	(0.169)
Pooled 3 years (Sp05, Sy06, Sy07)	29.39	26.10	3.29*	0.18*	(0.018)
Percentage of intervals in five dimensions with High Quality Student Practice					
Spring 2005	21.31	22.05	-0.74	-0.04	(0.749)
School year 2006	17.99	16.25	1.75	0.10	(0.295)
School year 2007	17.24	15.55	1.69	0.10	(0.300)
Pooled 3 years (Sp05, Sy06, Sy07)	18.44	17.61	0.82	0.05	(0.513)
Grade 2					
Minutes of instruction in the five dimensions combined					
Spring 2005	58.33	45.25	13.07*	0.62*	(<0.001)
School year 2006	60.14	49.30	10.84*	0.51*	(0.001)
School year 2007	58.57	52.06	6.51*	0.31*	(0.029)
Pooled 3 years (Sp05, Sy06, Sy07)	59.08	49.30	9.79*	0.46*	(<0.001)
Percentage of intervals in five dimensions with highly explicit instruction					
Spring 2005	32.02	25.15	6.86*	0.36*	(0.008)
School year 2006	31.33	24.38	6.95*	0.36*	(0.001)
School year 2007	30.02	31.97	-1.95	-0.10	(0.309)
Pooled 3 years (Sp05, Sy06, Sy07)	30.95	27.95	3.00*	0.16*	(0.040)
Percentage of intervals in five dimensions with High Quality Student Practice					
Spring 2005	22.86	18.96	3.90	0.22	(0.083)
School year 2006	16.40	13.04	3.35*	0.19*	(0.043)
School year 2007	16.40	14.24	2.16	0.12	(0.212)
Pooled 3 years (Sp05, Sy06, Sy07)	17.82	14.88	2.94*	0.16*	(0.019)

NOTES:

The complete Reading First Impact Study (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).

Impact estimates are statistically adjusted 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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: The observed mean amount of time spent per daily reading block in instruction in the five dimensions combined for first grade classrooms with Reading First was 59.23 minutes in spring 2005. The estimated mean amount of time without Reading First was 50.34 minutes. The impact of Reading First on the amount of time spent in instruction in the five dimensions combined was 8.89 minutes, which was statistically significant (p=.007).

SOURCES: RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, spring 2006, fall 2006, and spring 2007

Exhibit 3.2: Change Over Time in Program Impact on Reading Comprehension and Instruction

		Reading Comprehension (SAT 10 Scaled Score)	Reading Instruction (min. in 5 Dimensions)
Grade 1	Linear Year-to-Year Change	2.82	-2.59*
	SE	2.07	1.22
	p-value	0.174	0.034
	F-test for overall variation across years	0.808	1.48
	p-value	0.446	0.22
Grade 2	Linear Year-to-Year Change	0.53	-2.88*
	SE	1.77	1.25
	p-value	0.766	0.021
	F-test for overall variation across years	0.072	5.03*
	p-value	0.931	0.025
Grade 3	Linear Year-to-Year Change	3.81*	n.a.
	SE	1.74	n.a.
	p-value	0.029	n.a.
	F-test for overall variation across years	2.630	n.a.
	p-value	0.072	n.a.
All Available Grades [a]	Linear Year-to-Year Change	2.477*	-2.36*
	SE	1.08	0.87
	p-value	0.022	0.007
	F-test for overall variation across years	2.712	4.46*
	p-value	0.066	0.035

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 in 2006, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one RF school could not be included in the analysis because test score data were not available.

[a] For Reading Comprehension, grades 1-3 were included in the analysis. For Reading Instruction, only grades 1 and 2 were included in the analysis because instructional data were only available for these two grades.

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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: For grade 1, the program impact on reading comprehension increases by 2.82 scaled score points per year between 2005 and 2007. This change was not statistically significant (p=.174). The program impact on instruction in the five dimensions of reading instruction decreases by -2.59 minutes per daily reading block per year. This change was statistically significant (p=.034).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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, spring 2006, fall 2006, and spring 2007

Exhibit 3.3: Estimated Impacts on Reading Comprehension: Spring 2005, 2006, and 2007, and Pooled

Construct	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 Scaled Score					
Grade 1: Spring 2005	541.2	538.9	2.2	0.05	(0.524)
Spring 2006	545.7	540.4	5.3	0.11	(0.152)
Spring 2007	545.3	537.8	7.5	0.15	(0.052)
Pooled 3 years (2005, 2006, and 2007)	543.8	539.1	4.7	0.10	(0.083)
Grade 2: Spring 2005	583.5	582.4	1.2	0.03	(0.654)
Spring 2006	585.3	583.7	1.6	0.04	(0.620)
Spring 2007	584.8	582.3	2.5	0.06	(0.415)
Pooled 3 years (2005, 2006, and 2007)	584.4	582.8	1.7	0.04	(0.462)
Grade 3: Spring 2005	607.4	609.9	-2.5	-0.06	(0.306)
Spring 2006	609.5	610.0	-0.5	-0.01	(0.860)
Spring 2007	610.6	605.1	5.5	0.14	(0.082)
Pooled 3 years (2005, 2006, and 2007)	609.1	608.8	0.3	0.01	(0.887)
Panel 2					
All Sites					
Percent Reading At or Above Grade Level[1]					
Grade 1: Spring 2005	43.8	41.6	2.2		(0.529)
Spring 2006	47.3	43.0	4.3		(0.217)
Spring 2007	47.5	40.3	7.3*		(0.047)
Pooled 3 years (2005, 2006, and 2007)	46.0	41.8	4.2		(0.104)
Grade 2: Spring 2005	38.0	38.0	0.0		(0.996)
Spring 2006	39.9	39.6	0.3		(0.926)
Spring 2007	39.0	34.1	4.9		(0.121)
Pooled 3 years (2005, 2006, and 2007)	38.9	37.3	1.6		(0.504)
Grade 3: Spring 2005	36.0	39.3	-3.3		(0.255)
Spring 2006	39.9	40.8	-0.9		(0.801)
Spring 2007	40.5	34.8	5.6		(0.101)
Pooled 3 years (2005, 2006, and 2007)	38.7	38.8	-0.1		(0.973)

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 in 2006, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one 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.
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; statistically significant findings at the $p \leq .05$ level are indicated by *.
[1] The "at or above grade level" variable is dichotomous, therefore effect sizes are not appropriate.
EXHIBIT READS: The observed mean reading comprehension score for first-graders with Reading First was 541.2 scaled score points in spring 2005. The estimated mean without Reading First was 538.9 scaled score points. The impact of Reading First was 2.2 scaled score points (or 0.05 standard deviations), which was not statistically significant (p=.524). The observed average percent of first-graders reading at or above grade level with Reading First was 43.8 percentage points in spring 2005. The estimated average percent without Reading First was 41.6 percentage points. The impact of Reading First on the percent of first grade students reading at or above grade level was 2.2 percentage points, which was not statistically significant (p=.529).
SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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)

grade students in 2004-2005 were exposed to RF for only one year, while third graders in those same sites in 2006-2007 were exposed to RF for up to three years. The cross-sectional design of the study, in which all third grade students' scores are pooled across years—regardless of number of years of exposure—does not account for differing amounts of exposure. As a result, the program's observed effects may have been diluted, if in fact more years of exposure were related to greater impacts.

To address this issue, a separate analysis was conducted (see Appendix F) to assess the effect of three years of observed program exposure. The sample for this analysis comprised all third-graders in spring 2007 who were in a Reading First school during spring 2007 and 2005 (the program group) or in a non-Reading First school at both times (the comparison group). Given existing data, this is the best possible approximation to students with three years of program exposure.[28]

Program impacts for this subsample were then estimated for spring 2007 test scores.

- These findings suggest an average impact of 4.3 scaled score points (not statistically significant), which represents an effect size of 0.11 standard deviations (Exhibit 3.4). This estimate is smaller than that of 5.5 scaled score points (not statistically significant), which represents an effect size of 0.14, for all third-graders in spring 2007.

These impact estimates may be biased if Reading First caused a difference in the types of students who move from or stay at the same school. Because the study does not include pre-Reading First characteristics for students in the study sample, this question cannot be examined directly. As a result, the findings presented in this section should be interpreted with caution. Also, students who remain in schools with the same treatment status for three years likely differ along a number of important dimensions from students who do not, so the results of this analysis may have limited external validity.

Variation in Impacts Across Sites

This section explores whether the impact estimates presented in Chapter Two—which reflect averages across the 18 study sites—may be masking systematic differences in impacts among the sites. Study sites differ in both local conditions and in the timing that they received their Reading First grants, thus the exploratory analyses presented here explore a) site-by-site variation, and b) variation across early and late award sites.

[28] In the spring of 2005, the study tested students in all eligible classrooms in grades one through three in study schools. In subsequent waves of testing, the study tested students in a randomly selected subsample of classrooms in those study schools with four or more eligible classrooms per grade, on average, and continued to test all eligible students in eligible classrooms in those schools with three or fewer classrooms per grade level, on average. Because not all classrooms (and those classrooms' students) were tested in 2006, it is not possible to determine how many third graders tested in 2007 had also been in study schools in **both** 2005 and 2006. Also, because not all third grade students were tested in all study schools in 2007, this sample does not encompass all students who remained in the same type of school (within the study sample) for three years.

Exhibit 3.4: Estimated Impacts of Reading First on the Reading Comprehension of Students With Three Years of Exposure: Spring 2005-Spring 2007

	Actual Mean with Reading First	Estimated Mean without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (P-value)
Students With Three Years of Exposure					
Grade 3, Spring 2007					
Reading Comprehension					
Scaled Score	613.6	609.3	4.3	0.11	(0.223)
Corresponding Grade Equivalent	*3.5*	*3.3*			
Corresponding Percentile	*43*	*39*			

NOTES:

The Three-Year Exposure sample includes 243 schools from 18 sites (17 school districts and 1 state) located in 13 states. 123 schools are Reading First schools and 120 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 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; statistically significant findings at the $p \leq .05$ level are indicated by *.

EXHIBIT READS: The observed mean reading comprehension score for third-graders with three years of exposure to Reading First was 613.6 scaled score points. The estimated mean without Reading First was 609.3 scaled score points. The impact of Reading First was 4.3 scaled score points (or 0.11 standard deviations), which was not statistically significant (p=.223).

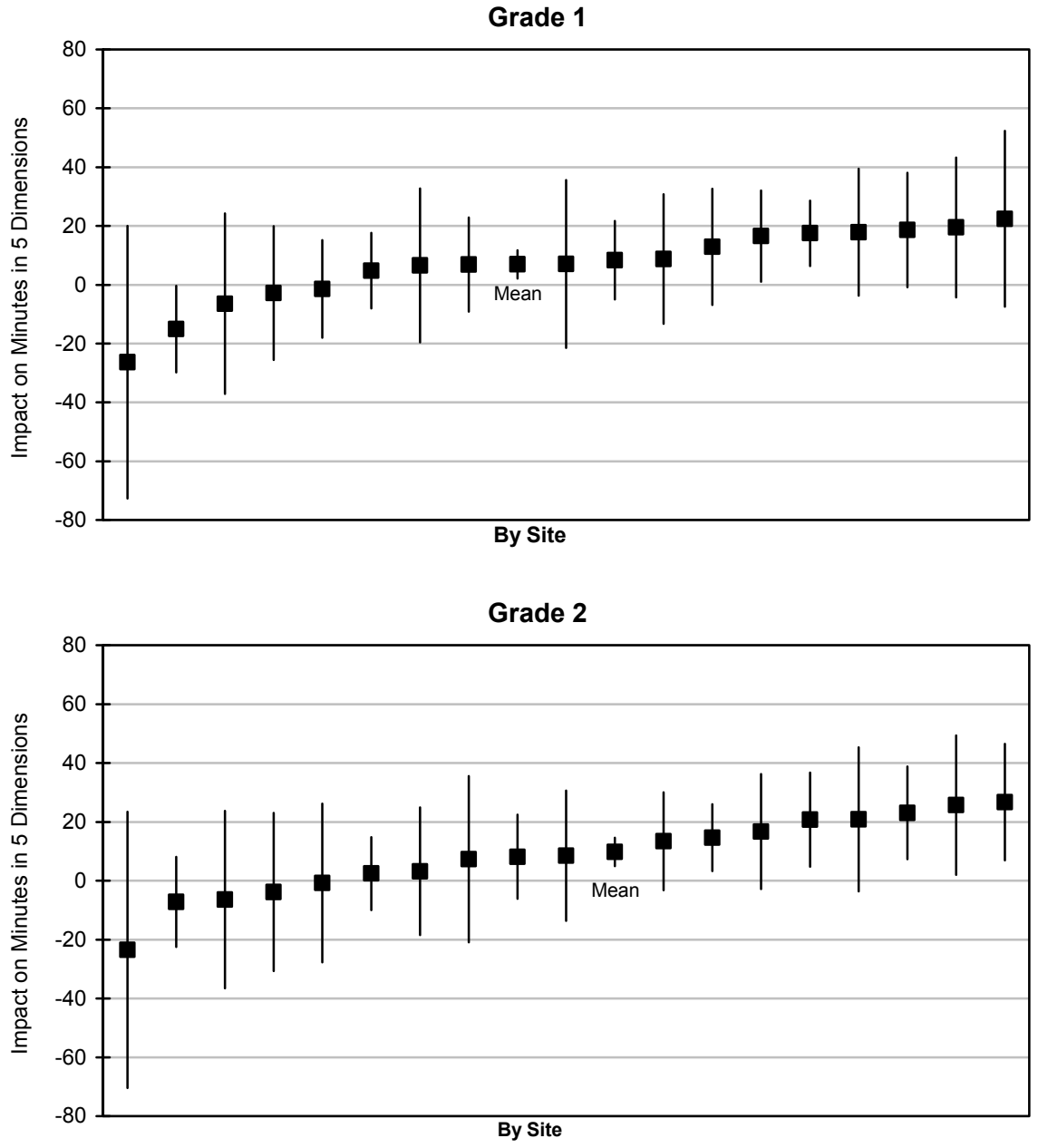
SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, as well as from state/district education agencies in those sites that already used the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR)

Site-by-Site Variation

If variation in Reading First impacts across study sites exists, it could represent important differences in program effectiveness by site, which are masked by average impacts. This variation might help to identify conditions under which the program is more (or less) effective. Because the present study was designed primarily to estimate average program impacts, there are limits to its statistical power and methodological ability to support causal inferences about impact variation. Nevertheless, information from the study about impact variation can help to provide a broader context for assessing its findings about average impacts.

Exhibits 3.5 and 3.6 graphically illustrate the impact estimates and 95 percent confidence intervals for instructional time in the five dimensions of reading and student test scores by site. This provides a visual representation of the variability in impacts as well as the uncertainty that exists about this variability.

Exhibit 3.5: Fixed Effect Impact Estimates for Instruction, by Site, by Grade



NOTES:

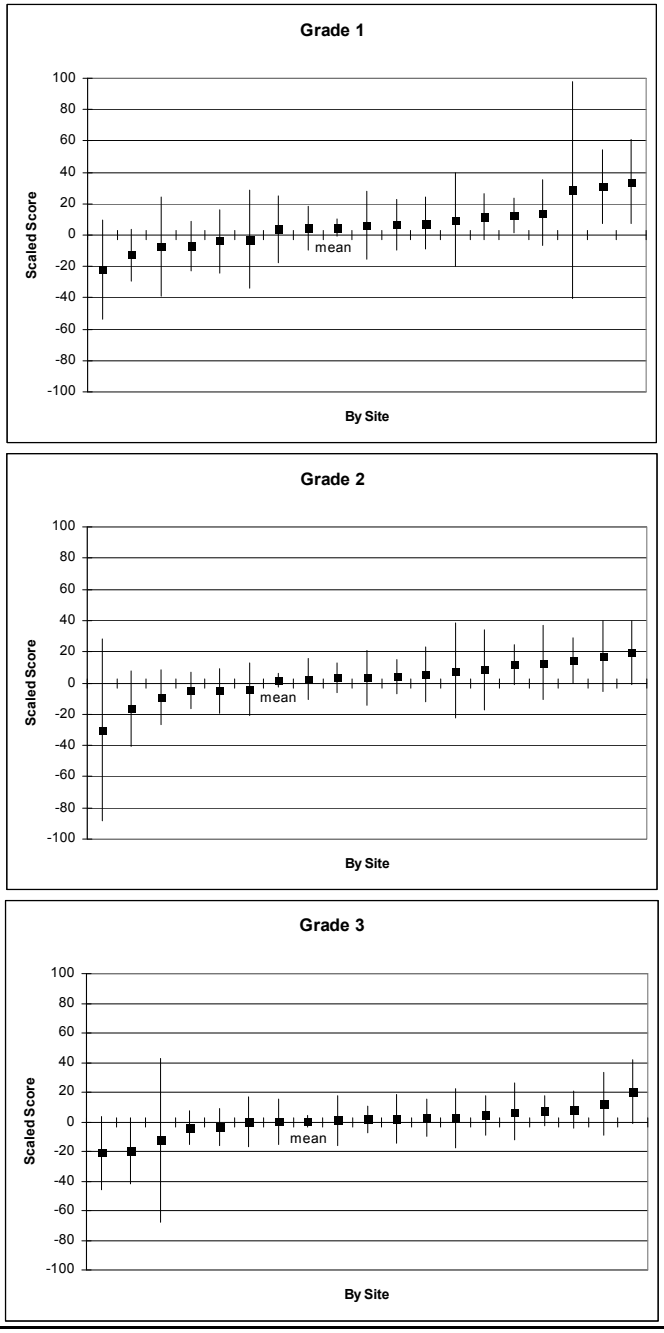
The complete Reading First Impact Study (RFIS) sample includes 248 schools from 18 sites (17 school districts and 1 state) in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools.

Impact estimates are statistically adjusted to reflect the regression discontinuity design of the study.

Boxes in exhibit represent mean impact estimates and lines represent 95 percent confidence intervals for each site.

SOURCE: RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, spring 2006, fall 2006 and spring 2007

Exhibit 3.6: Fixed Effect Impact Estimates for Reading Comprehension, by Site, by Grade



NOTES:

The complete Reading First Impact Study (RFIS) 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one 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.

Boxes in exhibit represent mean impact estimates and lines represent 95 percent confidence intervals for each site.

SOURCES: RFIS SAT 10 administration in the spring of 2007, 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)

A formal test of whether this variation is real (and whether it is statistically significant at the conventional $p<.05$ level or whether it reflects random error) was conducted for each outcome by grade and then pooled across grades (Exhibit 3.7).

- Estimated impacts on instructional time in the five dimensions per daily reading block ranged across site and grade from reductions of more than 20 minutes to increases of more than 20 minutes. Estimated impacts on reading comprehension scores ranged across sites and grade from reductions of nearly 30 scaled score points to increases of more than 35 scaled score points. However, formal tests indicated that this site-to-site variation was not statistically significant for either outcome, either by grade or overall, for classroom reading instruction or student reading comprehension, and therefore do not support the hypothesis that there is systematic variation site-to-site.

Exhibit 3.7: F-Test of Variation in Impacts Across Sites

		Reading Instruction (min. in 5 Dimensions)	Reading Comprehension (SAT 10 Scaled Score)
Grade 1	F-stat	1.34	1.424
	p-value	0.172	0.114
Grade 2	F-stat	1.31	1.076
	p-value	0.190	0.371
Grade 3	F-stat	n/a	0.903
	p-value	n/a	0.570
All Available Grades [a]	F-stat	1.47	1.142
	p-value	0.108	0.305

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 in 2006, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one RF school could not be included in the analysis because test score data were not available.

[a] For Reading Comprehension, grades 1-3 were included in the analysis. For Reading Instruction, only grades 1 and 2 were included in the analysis because instructional data were only available for these two grades.

Impact estimates are statistically adjusted (e.g., take into account 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; statistically significant findings at the $p\leq.05$ level are indicated by *.

EXHIBIT READS: The F-statistic for the joint F-test of whether the program impact is the same across all sites for first grade reading instruction is 1.34, which was not statistically significant (p=.172).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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, spring 2006, fall 2006 and spring 2007

Variation in Impacts Between Early and Late Award Sites

The RFIS Interim Report presented analyses that examined differences among two groups of sites that were identified at the outset of the study based on the timing of their grant awards. Early award sites (10 sites with 111 Reading First schools in the sample) received their initial Reading First grants between April and December 2003. Late award sites (8 sites with 137 Reading First schools in the sample)

received their initial Reading First grants between January and August 2004. When the data collection period for the study ended (in June 2007), early award sites had been funded for an average of 46 months, and late award sites had been funded for an average of 37 months.

The analyses conducted for this report update those from the Interim Report for the two main outcomes (reading instruction and reading comprehension) by incorporating data from the 2006-07 school year (see Appendix G).[29] For minutes of instruction in the five dimensions, Exhibit 3.8 indicates statistically significant impacts for late award sites, but not early award sites. For reading comprehension, as measured by scaled scores on the SAT 10, Exhibit 3.9 indicates no statistically significant impacts for early award sites and only one statistically significant impact (in Grade Two) for late award sites.

There is no statistically significant difference between estimated impacts in late award versus early award sites in minutes of instruction in the five dimensions for either Grade One or Grade Two (Exhibit 3.10). The composite test (on an index that combines the three instructional outcomes and pools data from first and second grades) of differences between the two groups of sites was, however, statistically significant. The difference between estimated impacts in late award versus early award sites for average scaled scores in student reading comprehension was statistically significant for only Grade Two (Exhibit 3.10). The composite test (on an index that combines scaled scores and indicators of students' at or above grade level performance and pools data across three grades) was not statistically significant. The inconsistent findings do not support the hypothesis that there is systematic variation across early and late award sites.

Exploring the Relationship between Classroom Reading Instruction and Student Achievement

The study provides a rigorous test of the extent to which the receipt of RF funding at the school level had an impact on instruction and reading achievement. However, another question of interest is whether the scientifically based reading instruction promoted by RF is related to student achievement, regardless of where it is implemented. Although the study design does not support a causal analysis of this question, the relationship between the study's instructional data and the study's achievement data (for grades one and two only) can be estimated using correlational techniques.

This section, therefore, explores the following research question: *What is the relationship between the degree of implementation of scientifically based reading instruction and student achievement?* by using hierarchical linear modeling to explore the observed correlations between instructional practices and student achievement in the RFIS sample of schools. These analyses are outside the causal research design (i.e., regression discontinuity design) described in Chapter Two, and can therefore provide evidence only about observed statistical associations between classroom instruction and student achievement in the study sample.

[29] This specific set of analyses was not conducted for the Student Engagement with Print measure.

Exhibit 3.8: Estimated Impacts on Classroom Instruction: 2005, 2006, and 2007 (pooled), 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					
Number of minutes of instruction in the five dimensions combined					
Grade 1	62.02	60.00	2.02	0.10	0.640
Grade 2	63.04	57.49	5.55	0.26	0.223
Percentage of intervals in five dimensions with highly explicit instruction					
Grade 1	29.90	26.12	3.78	0.21	0.067
Grade 2	31.34	31.38	-0.04	0.00	0.987
Percentage of intervals in five dimensions with High Quality Student Practice					
Grade 1	18.18	20.06	-1.88	-0.11	0.336
Grade 2	17.66	14.14	3.53	0.20	0.073
Late Award Sites					
Number of minutes of instruction in the five dimensions combined					
Grade 1	57.04	46.30	10.74*	0.52*	<0.001
Grade 2	55.98	42.90	13.08*	0.62*	<0.001
Percentage of intervals in five dimensions with highly explicit instruction					
Grade 1	28.98	25.98	3.01	0.17	0.109
Grade 2	30.65	25.25	5.40*	0.28*	0.004
Percentage of intervals in five dimensions with High Quality Student Practice					
Grade 1	18.63	15.70	2.93	0.17	0.073
Grade 2	17.95	15.41	2.54	0.14	0.113

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. 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).

Impact estimates are statistically adjusted 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; 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 62.02 minutes. The estimated mean amount of time without Reading First was 60.00 minutes. The impact of Reading First on the amount of time spent in instruction in the five dimensions was 2.02 minutes (or 0.10 standard deviations), which was not statistically significant (p=.640).

SOURCES: RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, spring 2006, fall 2006 and spring 2007

Exhibit 3.9: Estimated Impacts on Reading Comprehension: Spring 2005, 2006, and 2007 (pooled), 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					
Reading Comprehension					
Grade 1: Scaled Score	546.6	543.8	2.9	0.06	(0.569)
Corresponding Grade Equivalent	*1.8*	1.7			
Corresponding Percentile	*47*	44			
Grade 2: Scaled Score	587.4	591.8	-4.4	-0.10	(0.287)
Corresponding Grade Equivalent	*2.6*	2.7			
Corresponding Percentile	*41*	45			
Grade 3: Scaled Score	613.1	617.0	-3.9	-0.10	(0.343)
Corresponding Grade Equivalent	*3.5*	3.6			
Corresponding Percentile	*43*	46			
Late Award Sites					
Reading Comprehension					
Grade 1: Scaled Score	541.6	536.0	5.6	0.11	(0.061)
Corresponding Grade Equivalent	*1.7*	*1.6*			
Corresponding Percentile	*43*	*39*			
Grade 2: Scaled Score	582.1	576.1	6.0 *	0.14 *	(0.021)
Corresponding Grade Equivalent	*2.4*	2.3			
Corresponding Percentile	*38*	33			
Grade 3: Scaled Score	606.0	602.4	3.5	0.09	(0.108)
Corresponding Grade Equivalent	*3.1*	3.0			
Corresponding Percentile	*36*	34			

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. Among them, there are 8 late award sites, with 137 schools, and 10 early award sites, with 111 schools. For grade 2 in 2006, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one 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.

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; 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 late award sites was 541.6 scaled score points. The estimated mean without Reading First was 536.0 scaled score points. The impact of Reading First was 5.6 scaled score points (or 0.11 standard deviations), which was not statistically significant (p=.061).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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)

Exhibit 3.10: Award Group Differences in Estimated Impacts on Reading Comprehension and Classroom Instruction: 2005, 2006, and 2007 (pooled)

	Difference in Impact (Early - Late)	Effect Size of Difference	Statistical Significance of Differences (p-value)
Average Scaled Score			
Grade 1	-2.8	-0.06	(0.636)
Grade 2	-10.4*	-0.25*	(0.032)
Grade 3	-7.4	-0.19	(0.110)
Number of minutes spent in instruction in five dimensions combined			
Grade 1	-8.72	-0.42	(0.092)
Grade 2	-7.53	-0.35	(0.155)
Percentage of observation intervals in five dimensions with			
Highly Explicit Instruction			
Grade 1	0.78	0.04	(0.779)
Grade 2	-5.44	-0.28	(0.068)
High Quality Student Practice			
Grade 1	-4.81	-0.29	(0.059)
Grade 2	0.98	0.05	(0.696)

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. 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 data (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.

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

A composite test on an index that combines scaled scores and indicators of students' at or above grade level performance and pools data across three grades of differences between early and late sites was not statistically significant (p=.082).

A composite test on an index that combines the three instructional outcomes and pools data from first and second grades of differences between early and late sites was statistically significant (p=.037).

EXHIBIT READS: The estimated difference in impact between early and late award sites in grade 1 was -2.8 scaled score points. The effect size of the difference was -0.06 standard deviations. The estimated difference was not statistically significant (p=.636).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006, and 2007 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, spring 2006, fall 2006, and spring 2007

Specifically, this section examines statistical associations between several aspects of reading instruction, each of which is developed from observational data collected using the study's Instructional Practice in Reading Inventory, and student reading achievement, based on students' test scores on the reading comprehension subtest of the Stanford Achievement Test, 10th Edition (SAT 10). The measures of reading instruction used in this analysis are the same as those selected to represent the degree of implementation of scientifically based reading instruction in Chapter Two of this report. They include:

- average time spent per daily reading block in the five core dimensions of scientifically based reading instruction combined (referred to as "*time in the five dimensions*"),[30]
- average time spent per daily reading block in each of the five dimensions of scientifically based reading instruction (phonemic awareness, phonics, vocabulary, fluency and comprehension) separately,
- the proportion of three-minute time intervals during reading instruction in the five dimensions of reading instruction that involve highly explicit instruction (referred to as "*highly explicit instruction*"), and
- the proportion of three-minute time intervals during reading instruction in the five dimensions of reading instruction that involve high quality student practice (referred to as "*high quality student practice*").

This section also presents supplementary analyses that test whether there are other factors that might account for any observed relationship between the predictors outlined above and student reading comprehension. The study cannot possibly account for the complete set of alternative predictors in these models because it did not measure all the variables that are possibly related to both instruction and comprehension; nonetheless, three variables thought to be the most compelling are explored.

All analyses are conducted using data from all schools included in the study: those with and without Reading First funding, without accounting for treatment group. Results are also presented separately by treatment status. Instructional variables from classroom observations and the SAT 10 test scores from all three years of data collection (2005, 2006, and 2007) are included in these analyses. The unit of observation is the classroom within a given school year.[31] In Year One, the classroom instruction measures are derived from classroom observations conducted in the spring of 2005; in Years Two and Three, they represent the average of the fall and spring observations.

Caveats

The results described below should be interpreted with considerable caution. These analyses are outside the causal research design (i.e., regression discontinuity design) described in Chapter Two, and so do not provide evidence of a causal link between instructional practices and student reading comprehension.

Estimation Model

The analyses use a two-level hierarchical linear model to account for the repeated measures within classrooms, as well as indicator variables for schools to account for the nesting of classrooms within schools. More specifically, covariates in the models include:

- site indicators,

[30] These five dimensions of reading instruction (phonemic awareness, phonics, vocabulary, fluency and comprehension) are outlined in the Reading First legislation and in the guidance provided to states about Reading First.

[31] A 'classroom' is defined as having the same teacher at the same grade level in the same school. Since some teachers moved to other schools, and some to other grades within the same school over the study's three years of data collection, all classrooms are not necessarily represented in multiple years.

- school indicators,
- percentage of male students in the classroom,
- classroom level average of student age at start of school year,
- date of the post-test at the classroom level,
- school-level pre-program reading performance measure.[32]

In order to account for possible modeling differences associated with the year of data collection, all of the covariates (except school indicators) are interacted with indicators for each data collection period.[33] Site indicators are interacted with the predictors and covariates to allow the estimation of separate regression coefficients in each site. Each regression coefficient is then weighted according to the number of RF schools in the site prior to averaging across sites.

The multi-level model presented in (1) below estimates the degree to which variation in a particular predictor (PRE_{tj}) is associated with variation in the mean classroom-level reading comprehension test scores, controlling for the covariates listed above. For each grade, the model takes the following form:

$$Y_{tjkm} = \sum_{mt} \beta_{0m} ST_{mk} YR_t + \sum_{m} \beta_{1m} ST_{mk} PRE_{tj} + \sum_{k} \beta_{2k} SC_{jk} + \sum_{mt} \beta_{3m} ST_{mk} \overline{Y_{-1km}}\ YR_t \quad (1)$$
$$+ \sum_{t} \gamma_t Z_{jk} YR_t + \sum_{nt} \theta_n X_{nijkm} YR_t + \upsilon_{jk} + \varepsilon_{ijk}$$

where:

Y_{tjkm} = the average post-test score in year t, for classroom j, in school k, in site m,

ST_{mk} = one if school k is in site m and zero otherwise, m = 1 to 18,

PRE_{tj} = value of the predictor of interest in classroom j in year t,

SC_{jk} = the indicator variable for school k. In other words, it equals one if classroom j is in school k and zero otherwise, k = 1 to 248,

$\overline{Y_{-1km}}$ = the mean baseline pretest for school k (standardized and centered by site),

YR_t = indicator for follow-up years; 2005, 2006 or 2007,

Z_{tjk} = a variable indicating when the post-test in year t was given for classroom j in school k (site-centered),

X_{nijkm} = classroom average of the n^{th} demographic student characteristic in classroom j in school k, in site m

υ_{jk} and ε_{ijk} = classroom- level random error term and the residual, respectively, assumed to be independently and identically distributed.

[32] Different pre-program performance measures were constructed for early and late award sites. For the ten early award sites and one late award site (which had no fall 2004 test data due to a hurricane), performance on a state reading test (when available, an average of test scores from up to three pre-RF years) was used as a school level pretest measure. For late award sites except for the one without available fall 2004 data, the mean fall 2004 SAT 10 test scores for each school/grade were used as the pretest measure.

[33] This accounts for year-to-year variation in the levels of the outcome measure as well as the relationship between covariates and outcome measures.

The average estimated value of β_{1m} (m = 1, 2, …, 18), weighted by the number of RF schools in each site, captures the overall relationship between student test scores and the predictor of interest.[34] An important distinction between the model described here and those employed for the main impact analyses is the use of school level indicators in place of the rating variable. These school level indicators were introduced to control for *unobservable and time-invariant* school characteristics that affected the outcome and the predictors.

Findings

Descriptive statistics and bivariate correlations between all of the predictors as well as the outcome are presented in Exhibits 3.11 and 3.12. Correlation coefficients between the outcome and predictors range from -0.06 to 0.27, and from -0.00 to 0.30 for grades one and two, respectively.

The remainder of this section presents estimates of the relationship between student reading comprehension and the key measures of instruction listed above. First, the association between student reading comprehension and time spent on each of the five dimensions of reading instruction (phonemic awareness, phonics, comprehension, vocabulary, and fluency) was examined (Exhibit 3.13, Models I-V). A sixth model estimated the relationship between all five dimensions and comprehension; this model explores the relationship between comprehension and the time spent on a specific dimension controlling for the time spent on the other four dimensions. These analyses were conducted separately for grades one and two. Findings indicate that:

- In grade one, when tested individually, time spent on comprehension and vocabulary were both significantly and positively related to student achievement. Specifically, a one-minute difference per daily reading block in the time spent on comprehension is associated with a 0.15 scaled score point difference in student achievement, and a one-minute difference per daily reading block in the time spent on vocabulary is associated with a 0.22 point difference in student reading comprehension.
- Time spent on phonics in grade one, however, was significantly and negatively related to student reading comprehension. In particular, a one-minute difference per daily reading block in the time spent on phonics per daily reading block was associated with a –0.10 point difference in student test scores.
- In the model that tested the joint association between reading achievement and time spent on each dimension in grade one, only time spent on comprehension remained a significant predictor.
- In grade two, time spent on phonics was significantly and negatively related to student reading comprehension. Similar to the finding in grade one, a one-minute difference per daily reading block in the time spent on phonics was associated with a –0.15 point difference in student test scores.
- Time spent on comprehension was also significantly related to student reading comprehension in grade two, such that a one-minute difference per daily reading block in the time spent on comprehension was associated with a 0.12 point difference in student reading comprehension.

[34] Note that models that jointly tested multiple predictors were also estimated. In such cases, the overall relational coefficient for each predictor was calculated in a similar manner.

Exhibit 3.11: Descriptive Statistics

	Mean	Std Dev	N	Min	Max
Panel A: GRADE 1					
SAT10 Test Score	544.7	23.2	2199	423.0	629.7
Minutes spent on…					
Phonemic Awareness	1.64	2.35	2199	0.00	22.59
Phonics	19.21	11.25	2199	0.00	63.99
Comprehension	21.95	11.73	2199	0.00	72.26
Vocabulary	7.17	5.18	2199	0.00	31.82
Fluency	4.22	5.18	2199	0.00	44.74
Five dimensions combined	54.19	18.36	2199	0.00	132.15
Percentage of Intervals in the five dimensions with highly explicit instruction	28.48	13.88	2199	0.00	78.46
Percentage of Intervals in the five dimensions with high quality student practice	17.89	12.18	2199	0.00	81.53
Observation length	108.57	26.71	2199	30.00	237.75
Gats score	4.40	0.58	1403	1.98	5.00
Percentage of students engaged with print	46.26	22.49	1399	0.00	100.00
Pretest (Z-scored)	0.01	1.02	2199	-4.47	2.71
Panel B: GRADE 2					
SAT10 Test Score	586.1	19.0	2133	515.7	664.3
Minutes spent on…					
Phonemic Awareness	0.39	0.99	2133	0.00	15.27
Phonics	11.41	9.04	2133	0.00	59.69
Comprehension	26.70	13.24	2133	0.00	91.20
Vocabulary	10.32	6.67	2133	0.00	57.83
Fluency	3.57	4.65	2133	0.00	43.77
Five dimensions combined	52.37	18.28	2133	5.01	123.84
Percentage of Intervals in the five dimensions with highly explicit instruction	29.99	14.30	2133	0.00	92.15
Percentage of Intervals in the five dimensions with high quality student practice	17.38	11.99	2133	0.00	72.31
Observation length	106.15	26.43	2133	36.75	210.00
Gats score	4.41	0.59	1371	1.40	5.00
Percentage of students engaged with print	50.88	22.40	1363	0.00	100.00
Pretest (Z-scored)	0.01	1.02	2133	-3.92	2.89

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 in 2006, one non-RF school could not be included in the analysis because test score data were not available.

EXHIBIT READS: The mean grade one SAT 10 score was 544.7, with a standard deviation of 23.2 across 2,199 observations. The minimum score was 423.0, and the maximum score was 629.7.

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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, spring 2006, fall 2006, and spring 2007; RFIS Global Appraisal of Teaching Strategies, fall 2005, spring 2006, fall 2006, and spring 2007; RFIS Student Time-on-Task and Engagement with Print, fall 2005, spring 2006, fall 2006, and spring 2007

Exhibit 3.12: Bivariate Correlation Coefficients between Test Scores and Predictors

Panel A:GRADE 1

	SAT10 Test Score	Minutes spent on Phonemic Awareness	Minutes spent on Phonics	Minutes spent on Comprehen-sion	Minutes spent on Vocabulary	Minutes spent on Fluency Building	Minutes spent on the Five Dimensions Combined	Percentage of Intervals in the five dimensions with highly explicit instruction	Percentage of Intervals in the five dimensions with high quality student practice	Observation length	GATS score
SAT10 Test Score											
Minutes spent on Phonemic Awareness	-0.053										
Minutes spent on Phonics	-0.063	0.177									
Minutes spent on Comprehension	0.150	-0.066	-0.132								
Minutes spent on Vocabulary	0.091	0.065	0.079	0.165							
Minutes spent on Fluency Building	0.068	-0.049	0.062	0.052	0.001						
Minutes spent on the Five Dimensions Combined	0.095	0.199	0.590	0.610	0.444	0.347					
Percentage of Intervals in the five dimensions with highly explicit instruction	0.074	0.144	0.164	0.009	0.370	-0.092	0.202				
Percentage of Intervals in the five dimensions with high quality student practice	0.055	0.197	0.136	-0.015	0.045	0.207	0.170	0.183			
Observation length	0.017	0.088	0.314	0.335	0.262	0.222	0.554	-0.030	-0.004		
GATS score	0.269	0.060	0.160	0.092	0.165	0.073	0.228	0.183	0.195	-0.005	
Percentage of students engaged with print	0.174	-0.077	0.065	-0.022	-0.005	0.139	0.047	0.066	0.109	-0.095	0.165

Exhibit 3.12: Bivariate Correlation Coefficients between Test Scores and Predictors (continued)

Panel B:GRADE 2

	SAT10 Test Score	Minutes spent on Phonemic Awareness	Minutes spent on Phonics	Minutes spent on Comprehen-sion	Minutes spent on Vocabulary	Minutes spent on Fluency Building	Minutes spent on the Five Dimensions Combined	Percentage of Intervals in the five dimensions with highly explicit instruction	Percentage of Intervals in the five dimensions with high quality student practice	Observation length	GATS score
SAT10 Test Score											
Minutes spent on Phonemic Awareness	-0.003										
Minutes spent on Phonics	-0.129	0.210									
Minutes spent on Comprehension	0.093	-0.078	-0.136								
Minutes spent on Vocabulary	0.027	0.008	0.073	0.138							
Minutes spent on Fluency Building	-0.030	0.015	0.100	0.010	-0.033						
Minutes spent on the Five Dimensions Combined	0.006	0.108	0.459	0.705	0.493	0.300					
Percentage of Intervals in the five dimensions with highly explicit instruction	0.123	0.102	0.079	0.072	0.369	-0.085	0.210				
Percentage of Intervals in the five dimensions with high quality student practice	0.059	0.123	0.155	0.072	0.075	0.152	0.201	0.232			
Observation length	-0.091	0.033	0.288	0.370	0.246	0.177	0.547	-0.014	-0.041		
GATS score	0.303	0.015	0.072	0.199	0.136	0.096	0.247	0.220	0.220	-0.038	
Percentage of students engaged with print	0.173	-0.030	0.027	0.005	-0.057	0.069	0.010	0.069	0.011	-0.091	0.190

Exhibit 3.13: Regression Coefficients for the Relationship between Classroom Reading Instruction and Reading Comprehension

	I	II	III	IV	V	VI
Panel A: GRADE 1						
Minutes in…						
Phonemic Awareness	-0.220 (0.316)	-	-	-	-	-0.102 (0.656)
Phonics	-	-0.103* (0.024)	-	-	-	-0.072 (0.135)
Comprehension	-	-	0.148* (<0.001)	-	-	0.131* (0.005)
Vocabulary	-	-	-	0.219* (0.017)	-	0.175 (0.062)
Fluency	-	-	-	-	0.146 (0.206)	0.148 (0.212)
Panel B: GRADE 2						
Minutes in…						
Phonemic Awareness	-0.128 (0.769)	-	-	-	-	0.158 (0.729)
Phonics	-	-0.150* (<0.001)	-	-	-	-0.138* (0.003)
Comprehension	-	-	0.115* (<0.001)	-	-	0.099* (0.002)
Vocabulary	-	-	-	0.086 (0.139)	-	0.084 (0.159)
Fluency	-	-	-	-	0.004 (0.966)	0.074 (0.443)

NOTES:

Sample sizes for grade 1 and 2 analyses are 2,199 and 2,133 classrooms, respectively. 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available.

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

EXHIBIT READS: For grade 1, the regression coefficient between minutes spent teaching phonemic awareness and student achievement is -.22, which means that a one-minute difference in the amount of time spent teaching phonemic awareness per daily reading block is associated with a –0.22 point difference in student test scores. This association is not statistically significant (p=0.316).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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, spring 2006, fall 2006, and spring 2007

- These two predictors remained significant in the specification that tested all five predictors jointly in grade two.

These analyses were also run separately by treatment status to see whether the relationship between instruction and comprehension differed between the two groups of schools. As shown in Exhibits 3.14 and 3.15, except in phonics in grade one (p=.035), there are no statistically significant differences in the estimates for the treatment and comparison groups in either grade. However, note that in Exhibit 3.14, Model II, in which phonics is included on its own, the difference between the estimated coefficients for the treatment and the comparison groups is not statistically significant. Overall, therefore, the results

suggest that the estimated relationship between student reading comprehension and key measures of reading instruction do not differ across the treatment and comparison groups.

Exhibit 3.14: Regression Coefficients Between Classroom Reading Instruction and Reading Comprehension by Treatment Status—Grade 1

	I	II	III	IV	V	VI
Panel A: Treatment Group						
Minutes in…						
Phonemic Awareness	-0.401	-	-	-	-	-0.185
	(0.176)					(0.555)
Phonics	-	-0.182*	-	-	-	-0.160*
		(0.006)				(0.027)
Comprehension	-	-	0.143*	-	-	0.076
			(0.039)			(0.308)
Vocabulary	-	-	-	0.226	-	0.186
				(0.076)		(0.168)
Fluency	-	-	-	-	0.100	0.171
					(0.546)	(0.331)
Panel B: Comparison Group						
Minutes in…						
Phonemic Awareness	0.121	-	-	-	-	0.237
	(0.771)					(0.590)
Phonics	-	0.003	-	-	-	0.064
		(0.965)				(0.409)
Comprehension	-	-	0.143*	-	-	0.169*
			(0.028)			(0.018)
Vocabulary	-	-	-	0.051	-	-0.012
				(0.732)		(0.940)
Fluency	-	-	-	-	0.279	0.332
					(0.152)	(0.128)
Panel C: P-values from t-tests comparing treatment and comparison estimates						
Minutes in…						
Phonemic Awareness	0.307	-	-	-	-	0.434
Phonics	-	0.057	-	-	-	0.035*
Comprehension	-	-	1.000	-	-	0.367
Vocabulary	-	-	-	0.372	-	0.339
Fluency	-	-	-	-	0.484	0.565

NOTES:

Sample size for grade 1 analysis is 2,199 classrooms. The complete Reading First Impact 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.

A two-tailed test of significance was used; statistically significant findings at the $p \leq .05$ level are indicated by *. In panels A and B, p-values are in parentheses.

EXHIBIT READS: For the treatment group in grade 1, the regression coefficient between minutes in phonemic awareness and student achievement is -.401, which means that a one-minute difference in the time spent teaching phonemic awareness per daily reading block is associated with a –0.40 point difference in student test scores. This association is not statistically significant (p=.176).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006, and 2007, 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, spring 2006, fall 2006, and spring 2007

Exhibit 3.15: Regression Coefficients Between Classroom Reading Instruction and Reading Comprehension by Treatment Status—Grade 2

	I	II	III	IV	V	VI
Panel A: Treatment Group						
Minutes in…						
Phonemic Awareness	0.025	-	-	-	-	0.541
	(0.970)					(0.451)
Phonics	-	-0.073	-	-	-	-0.027
		(0.270)				(0.709)
Comprehension	-	-	0.102*	-	-	0.097
			(0.031)			(0.066)
Vocabulary	-	-	-	0.078	-	0.056
				(0.347)		(0.528)
Fluency	-	-	-	-	-0.067	-0.004
					(0.626)	(0.978)
Panel B: Comparison Group						
Minutes in…						
Phonemic Awareness	-0.748	-	-	-	-	-0.633
	(0.523)					(0.626)
Phonics	-	-0.063	-	-	-	-0.062
		(0.423)				(0.466)
Comprehension	-	-	0.147*	-	-	0.123*
			(0.001)			(0.013)
Vocabulary	-	-	-	0.126	-	0.112
				(0.161)		(0.228)
Fluency	-	-	-	-	0.229	0.329
					(0.160)	(0.056)
Panel C: P-values from t-tests comparing treatment and comparison estimates						
Minutes in…						
Phonemic Awareness	0.568	-	-	-	-	0.418
Phonics	-	0.922	-	-	-	0.754
Comprehension	-	-	0.496	-	-	0.718
Vocabulary	-	-	-	0.695	-	0.663
Fluency	-	-	-	-	0.166	0.145

NOTES:

Sample size for grade 2 analysis is 2,133 classrooms. The complete Reading First Impact 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available.

A two-tailed test of significance was used; statistically significant findings at the $p \leq .05$ level are indicated by *. In panels A and B p-values are in parentheses.

EXHIBIT READS: For the treatment group in grade 2, the regression coefficient between minutes in phonemic awareness and student achievement is .025, which means that a one-minute difference in the time spent teaching phonemic awareness per daily reading block is associated with a 0.03 point difference in student test scores. This association is not statistically significant (p=.970).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006, and 2007, 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, spring 2006, fall 2006, and spring 2007

Next, the associations between student reading comprehension and three more broadly defined measures of reading instruction were examined (Exhibit 3.16). These measures are total time spent on the five dimensions, percentage of classroom observation intervals in which teachers used highly explicit instructional strategies associated with the five dimensions, and percentage of intervals in which students were provided with high quality reading practice. First, three models were fit using each measure as a predictor of student reading comprehension separately. Then, all three measures were included together in a fourth model.

Exhibit 3.16: Regression Coefficients Between Broadly Defined Measures of Classroom Instruction and Reading Comprehension

	I	II	III	IV
Panel A: GRADE 1				
Minutes in the five dimensions	0.073*	-	-	0.073*
	(0.014)			(0.019)
Percentage of Intervals in the five dimensions with highly explicit instruction	-	-0.023	-	-0.039
		(0.479)		(0.247)
Percentage of Intervals in the five dimensions with high quality student practice	-	-	0.040	0.038
			(0.270)	(0.311)
Panel B: GRADE 2				
Minutes in the five dimensions	0.051*	-	-	0.058*
	(0.034)			(0.023)
Percentage of Intervals in the five dimensions with highly explicit instruction	-	0.007	-	-0.004
		(0.778)		(0.886)
Percentage of Intervals in the five dimensions with high quality student practice	-	-	0.022	0.008
			(0.450)	(0.790)

NOTES:

These analyses use available data from all years (Grade 1 and Grade 2 analysis sample sizes are 2,199 and 2,133 classrooms, respectively). The complete Reading First Impact 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available.

A two-tailed test of significance was used; statistically significant findings at the $p \leq .05$ level are indicated by *. P-values are in parentheses.

EXHIBIT READS: For grade 1, the regression coefficient between minutes spent teaching the five dimensions of reading and student achievement is .073, which means that a one-minute difference in the amount of time spent teaching the five dimensions of reading per daily reading block is associated with a 0.07 point difference in student test scores. This association is statistically significant (p=.014).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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, spring 2006, fall 2006, and spring 2007

When tested individually, total time spent on the five dimensions of reading was significantly and positively related to reading achievement in both grades. As Model I in Exhibit 3.16, Panel A shows, a one-minute difference in the total time spent on five dimensions per daily reading block was associated with a 0.07 point difference in student test scores in grade one. In grade two, a one-minute difference in time spent teaching the five dimensions per daily reading block was associated with a 0.05 point difference in student test scores. When tested jointly with the other two main predictors of interest, the same relationship was observed between total time spent on the five dimensions of reading and student reading comprehension in both grades (Model IV). Results of these analyses run separately by treatment

status indicate that there are no statistically significant differences across the two groups of schools (see Exhibit 3.17).

The previous analysis suggests that time spent in the five dimensions of reading is positively related to levels of student reading comprehension. However, it is quite possible that some other variable(s), not included in these models, may actually account for the observed relationship. For example, teachers who spend more time on the five dimensions of reading may simply devote more time to reading, have more organized classrooms, or have students who spend more classroom time engaged with print material. Therefore, in addition to the primary predictors, three other measures—length of the reading block, a global measure of instructional quality (instructional organization and order), and percentage of students engaged with print—were also tested as alternative predictors of student reading comprehension.

Because two of the alternative predictors (instructional organization and order and percentage of students engaged with print) were not collected in the first study year, the model that jointly tested the three main predictors was re-estimated on two subsamples of 1,399 Grade One and 1,363 Grade Two classrooms for which all six predictors (three main and three alternative) were available (Exhibit 3.18, Model I). All further analyses were conducted using this subsample.

Since the subsamples used to estimate Model I in Exhibit 3.18 are substantially different (and only about two-thirds as large) as the full samples used to estimate Model IV in Exhibit 3.16, the results of analyses using the subsamples should be interpreted with caution. We cannot know whether we would have observed the same pattern of results if we had been able to use the full sample for these analyses. For example, Exhibits 3.16 and 3.18 indicate that even before adding the alternative predictors to the model, the relationships are substantively different when estimating with the subsample rather than the full sample, such that the relationship between minutes spent in the five dimensions of reading is no longer statistically significant in either first or second grade in the subsample. In addition, in first grade, the relationship between highly explicit instruction is negative and statistically significant and the relationship between high quality student practice is positive and statistically significant in the subsample, when neither was statistically significant in the full sample.

The alternative hypotheses were tested by estimating a single model that included all six primary and secondary predictors (Exhibit 3.18, Model II). The exhibit presents separate estimates from these analyses for grades one and two.

- In grade one, when jointly tested using the classrooms for which all six predictors were available, one of the primary predictors (the measure accounting for the presence of highly explicit instruction in the five dimensions) was significantly linked to achievement. More specifically, a one-percentage point difference in number of the intervals that included highly explicit instruction in the five dimensions was related to a –0.14 points difference in student test scores.
- None of the three primary predictors were statistically significantly related to student test scores in grade two, when the model was estimated with all six predictors.

Exhibit 3.17: Regression Coefficients Between Broadly Defined Measures of Classroom Instruction and Reading Comprehension by Grade and Treatment Status

	I	II	III	IV	V	VI	VII	VIII
			GRADE 1				*GRADE 2*	
Panel A: Treatment Group								
Minutes in the five dimensions	0.042 (0.318)	-	-	0.032 (0.488)	0.075* (0.043)	-	-	0.093* (0.018)
Percentage of Intervals in the five dimensions with highly explicit instruction	-	-0.015 (0.757)	-	-0.016 (0.755)	-	-0.005 (0.903)	-	-0.030 (0.485)
Percentage of Intervals in the five dimensions with high quality student practice	-	-	0.053 (0.321)	0.071 (0.209)	-	-	0.040 (0.366)	0.011 (0.821)
Panel B: Comparison Group								
Minutes in the five dimensions	0.0124* (0.009)	-	-	0.136* (0.006)	0.062 (0.081)	-	-	0.064 (0.098)
Percentage of Intervals in the five dimensions with highly explicit instruction	-	-0.036 (0.456)	-	-0.052 (0.296)	-	0.033 (0.367)	-	0.034 (0.406)
Percentage of Intervals in the five dimensions with high quality student practice	-	-	0.001 (0.993)	-0.011 (0.984)	-	-	-0.011 (0.796)	-0.040 (0.405)
Panel C: P-values from t-tests comparing treatment and comparison estimates								
Minutes in the five dimensions	0.194	-	-	0.121	0.799	-	-	0.598
Percentage of Intervals in the five dimensions with highly explicit instruction	-	0.760	-	0.619	-	0.480	-	0.273
Percentage of Intervals in the five dimensions with high quality student practice	-	-	0.494	0.365	-	-	0.415	0.448

NOTES:

These analyses use available data from all years (Grade 1 and Grade 2 analysis sample sizes are 2,199 and 2,133 classrooms, respectively). 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available.

A two-tailed test of significance was used; statistically significant findings at the $p \leq .05$ level are indicated by *. P-values are in parentheses.

EXHIBIT READS: For the treatment group in grade 1, the regression coefficient between minutes spent teaching the five dimensions of reading and student achievement is .042, which means that a one-minute difference in the amount of time spent teaching the five dimensions of reading per daily reading block is associated with a 0.04 point difference in student test scores. This association is not statistically significant (p=.318).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006, and 2007, 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, spring 2006, fall 2006, and spring 2007

Exhibit 3.18: Regression Coefficients Between All Predictors and Reading Comprehension

	I	II
Panel A: GRADE 1		
Minutes in the five dimensions	0.089 (0.056)	0.078 (0.171)
Percentage of Intervals in the five dimensions with highly explicit instruction	-0.126* (0.019)	-0.136* (0.013)
Percentage of Intervals in the five dimensions with high quality student practice	0.128* (0.034)	0.118 (0.059)
Observation length		-0.022 (0.645)
GATS score		3.702* (0.002)
Percentage of students engaged with print		0.036 (0.194)
Panel B: GRADE 2		
Minutes in the five dimensions	0.042 (0.273)	-0.010 (0.825)
Percentage of Intervals in the five dimensions with highly explicit instruction	-0.015 (0.715)	-0.039 (0.376)
Percentage of Intervals in the five dimensions with high quality student practice	0.006 (0.909)	-0.016 (0.761)
Observation length		0.018 (0.638)
GATS score		5.407* (<0.001)
Percentage of students engaged with print		0.002 (0.939)

NOTES:

These analyses use the sample of classrooms for which all predictors are available (Grade 1 and Grade 2 analysis sample sizes are 1,399 and 1,363 classrooms, respectively). The complete Reading First Impact 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available.

A two-tailed test of significance was used; statistically significant findings at the $p \leq .05$ level are indicated by *. P-values are in parentheses.

EXHIBIT READS: Controlling for the other variables in the model, the regression coefficient between minutes spent teaching the five dimensions of reading and student achievement is .089, which means that a one-minute difference in the time spent teaching the five dimensions per daily reading block of reading is associated with a 0.09 difference in student test scores. This association is not statistically significant (p=.056).

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006 and 2007, 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, spring 2006, fall 2006, and spring 2007

- Among the secondary predictors, the relationship between instructional organization and order was positively and statistically significantly related to student test scores in both first and second grade. A one point difference in the measure of instructional organization and order (measured on a five point scale) was associated with a 3.7 point difference in test scores in first grade and a 5.4 point increase in second grade.

These analyses were conducted separately by treatment status to determine whether the relationship between instruction and comprehension differed between the two groups of schools. Results are shown in Exhibit 3.19. Again, there is no pattern of statistically significant differences across the two groups of schools.

Summary

In sum, the correlational analyses described above indicate a positive association between time spent on the five essential components of reading instruction promoted by the program and reading comprehension as measured by the SAT 10, but these findings are sensitive to both model specification and the sample used to estimate the relationship. In addition, these analyses do not support causal inferences.

Exhibit 3.19: Regression Coefficients Between All Predictors and Reading Comprehension by Treatment Status

	Grade 1		Grade 2	
	I	II	III	IV
Panel A: Treatment Group				
Minutes in the five dimensions	0.081	0.030	0.071	0.018
	(0.330)	(0.772)	(0.274)	(0.822)
Percentage of Intervals in the five dimensions with highly explicit instruction	-0.056	-0.035	-0.058	-0.125
	(0.546)	(0.743)	(0.414)	(0.147)
Percentage of Intervals in the five dimensions with high quality student practice	0.052	0.023	0.120	0.076
	(0.595)	(0.827)	(0.151)	(0.395)
Observation length	-	-0.041	-	0.014
		(0.690)		(0.867)
GATS score	-	1.846	-	6.854*
		(0.343)		(<0.001)
Percentage of students engaged with print	-	0.078	-	0.005
		(0.100)		(0.903)
Panel B: Comparison Group				
Minutes in the five dimensions	0.151*	0.141	0.045	-0.062
	(0.039)	(0.211)	(0.404)	(0.376)
Percentage of Intervals in the five dimensions with highly explicit instruction	-0.203*	-0.215*	0.015	0.013
	(0.012)	(0.019)	(0.781)	(0.830)
Percentage of Intervals in the five dimensions with high quality student practice	0.144	0.164	-0.148	-0.138
	(0.135)	(0.108)	(0.053)	(0.406)
Observation length	-	-0.113	-	-0.011
		(0.157)		(0.839)
GATS score	-	6.301*	-	4.813*
		(0.006)		(0.012)
Percentage of students engaged with print	-	-0.016	-	-0.046
		(0.745)		(0.211)
Panel C: P-values from t-tests comparing treatment and comparison estimates				
Minutes in the five dimensions	0.523	0.462	0.764	0.450
Percentage of Intervals in the five dimensions with highly explicit instruction	0.228	0.203	0.414	0.186
Percentage of Intervals in the five dimensions with high quality student practice	0.503	0.334	0.018*	0.254
Observation length	-	0.573	-	0.800
GATS score	-	0.135	-	0.417
Percentage of students engaged with print	-	0.165		0.336

NOTES:
These analyses use the sample of classrooms for which all predictors are available (Grade 1 and Grade 2 analysis sample sizes are 1,399 and 1,363, respectively). The complete Reading First Impact 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 in 2006, one non-RF school could not be included in the analysis because test score data were not available.
A two-tailed test of significance was used; statistically significant findings at the $p \leq .05$ level are indicated by *. In panels A and B, p-values are in parentheses.
EXHIBIT READS: Controlling for the other variables in the model, the regression coefficient between minutes spent teaching the five dimensions of reading and student achievement is .081, which means that a one-minute difference in the time spent teaching the five dimensions of reading per daily reading block is associated with a 0.08 difference in student test scores. This association is not statistically significant (p=.330).
SOURCES: RFIS SAT 10 administration in the spring of 2006 and 2007, 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, spring 2006, fall 2006, and spring 2007; RFIS Global Appraisal of Teaching Strategies, fall 2005, spring 2006, fall 2006, and spring 2007; and RFIS Student Time-on-Task and Engagement with Print, fall 2005, spring 2006, fall 2006, and spring 2007

Summary

This chapter explored a number of hypotheses to explain the pattern of observed impacts. Analyses that explored the association between the length of implementation of Reading First in the study schools and reading comprehension scores, as well as between the number of years students had been exposed to Reading First instruction and reading comprehension scores were inconclusive. No statistically significant variation across sites in the pattern of impacts was found. Correlational analyses indicate a positive association between time spent on the five essential components of reading instruction promoted by the program and reading comprehension as measured by the SAT 10, but these findings appear to be sensitive to model specification and the sample used to estimate the relationship.

The study finds, on average, that after several years of funding the Reading First program, it has a consistent positive effect on reading instruction yet no statistically significant impact on student reading comprehension. Findings based on exploratory analyses do not provide consistent or systematic insight into the pattern of observed impacts.