Reading First Impact Study

Final Report

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

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Evaluation and Regional Assistance
Phoebe Cottingham
Commissioner

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

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)

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 44th, 39th, and 39th 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≤.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≤.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

Legislative Specifications and Administrative Guidelines → **Flow of Funds to Eligible Schools** → **Design and Implementation of Research-Based Reading Programs** → **Enhanced Student Reading Achievement**

- NCLB, Title I, Part B, Subpart I
- Specification of effective reading program components
- Rules for state grant and district subgrant formulas and allocation
- Specification for allowable state and district use of funds
- Administrative guidelines for state grant application and district subgrant
- Accountability and evaluation requirements

SEAs submit applications for Reading First funds

↓

Expert panel review

↓

Awarded funds to SEAs with approved application

↓

Eligible LEAs and/or schools submit competitive subgrant proposal

↓

SEA review

↓

SEAs and/or schools award subgrants to LEAs and/or schools with approved applications

↓

Funds distributed to eligible schools

Use of research-based reading programs, instructional materials, and assessments, as articulated in LEA/school applications

↓

Teacher professional development in use of materials and instructional approaches

↓

- Teacher use of instructional strategies and content based on five essential components of reading instruction
- Use of assessments to diagnose student needs and measure progress
- Classroom organization and supplemental services and materials that support five essential components

- Increased proportion of students reading at/above grade level
- Adequate mastery of five essential components of early reading
- All students reading at grade level by the end of third grade

State and district policy context; existing reading programs; other resources and programs that may support reading

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≤.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≤.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≤.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≤.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≤.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≤.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≤.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≤.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≤.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≤.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
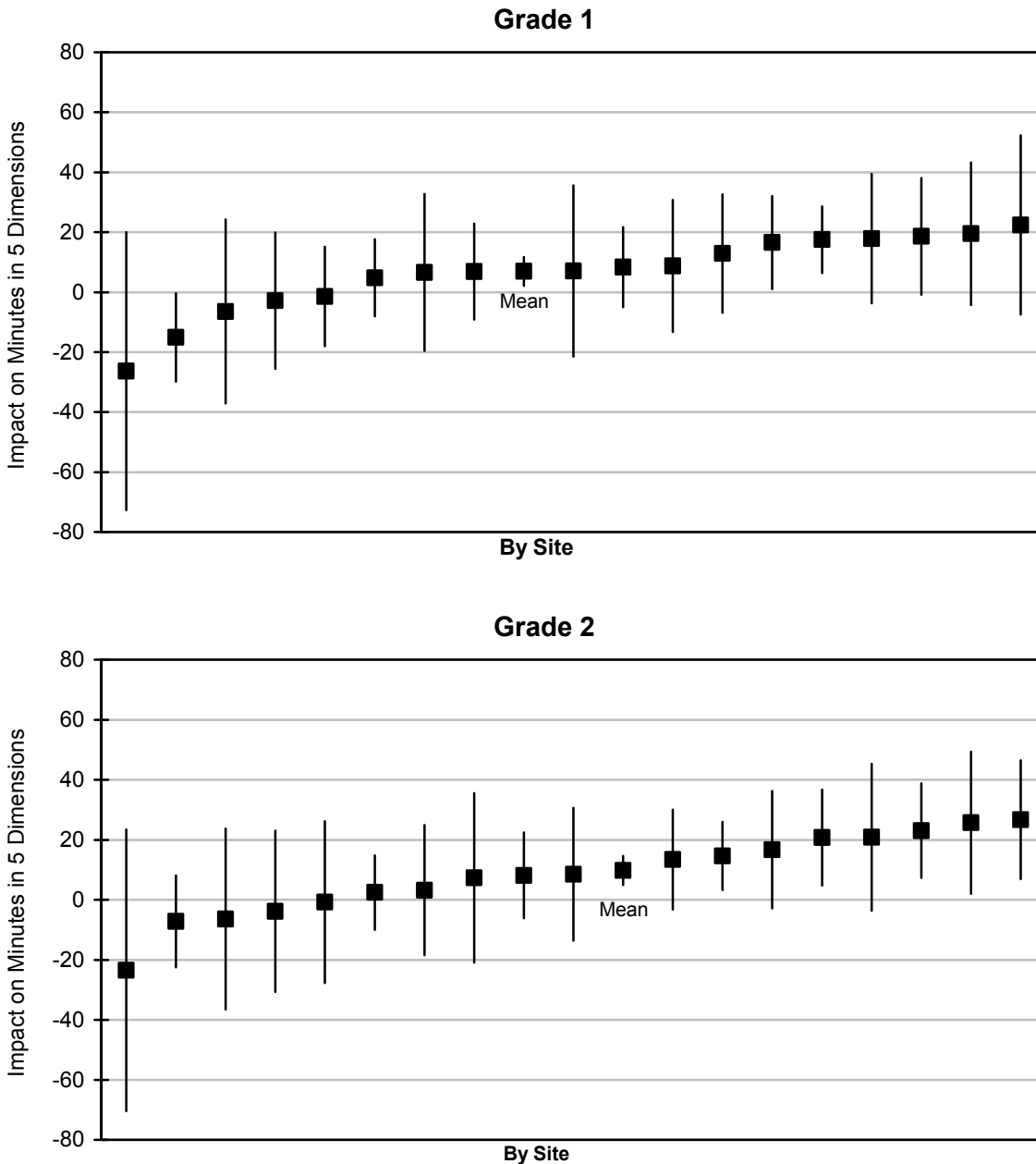
NOTES:

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≤.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≤.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≤.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 nth 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 Comprehension	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≤.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≤.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≤.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≤.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≤.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 (0.330)	0.030 (0.772)	0.071 (0.274)	0.018 (0.822)
Percentage of Intervals in the five dimensions with highly explicit instruction	-0.056 (0.546)	-0.035 (0.743)	-0.058 (0.414)	-0.125 (0.147)
Percentage of Intervals in the five dimensions with high quality student practice	0.052 (0.595)	0.023 (0.827)	0.120 (0.151)	0.076 (0.395)
Observation length	-	-0.041 (0.690)	-	0.014 (0.867)
GATS score	-	1.846 (0.343)	-	6.854* (<0.001)
Percentage of students engaged with print	-	0.078 (0.100)	-	0.005 (0.903)
Panel B: Comparison Group				
Minutes in the five dimensions	0.151* (0.039)	0.141 (0.211)	0.045 (0.404)	-0.062 (0.376)
Percentage of Intervals in the five dimensions with highly explicit instruction	-0.203* (0.012)	-0.215* (0.019)	0.015 (0.781)	0.013 (0.830)
Percentage of Intervals in the five dimensions with high quality student practice	0.144 (0.135)	0.164 (0.108)	-0.148 (0.053)	-0.138 (0.406)
Observation length	-	-0.113 (0.157)	-	-0.011 (0.839)
GATS score	-	6.301* (0.006)	-	4.813* (0.012)
Percentage of students engaged with print	-	-0.016 (0.745)	-	-0.046 (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≤.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.

Appendix A: State and Site Award Data

Appendix A presents additional information on when Reading First Impact Study sample sites first received Reading First awards (Exhibit A.1).

Exhibit A.1: Award Date by Site in Order of Date when Reading First Funds Were First Made Available for Implementation

	Date Initial Reading First Award Was Announced	Date when Reading First Funds Were First Made Available for Implementation
Site 9	03/2003	04/2003
Site 12	04/2003	05/2003
Site 2	06/2003	06/2003
Site 6	05/2003	06/2003
Site 5	02/2003	07/2003
Site 4	05/2003	07/2003
Site 18	06/2003	08/2003
Site 10*	10/2003	08/2003
Site 11*	10/2003	10/2003
Site 17*	08/2003	12/2003
Site 14	01/2004	02/2004
Site 8	01/2004	03/2004
Site 3	03/2004	04/2004
Site 13	01/2004	04/2004
Site 15	10/2003	05/2004
Site 1	05/2004	06/2004
Site 7	05/2004	06/2004
Site 16	03/2004	08/2004

NOTE:

Sites 10, 11 and 17 "backdated" the point at which schools could begin spending their grant money. It is not an error that the schools appear to have been given their money before their grants were announced.

SOURCE: Reading First District Coordinators

Appendix B: Methods

This appendix describes the general regression discontinuity approach used to estimate the impacts of Reading First and presents the specific models used to estimate impacts. In addition, it describes how the issue of multiple hypothesis testing was addressed and provides information about statistical precision.

Part 1: Regression Discontinuity Design

Approach

The Reading First Impact Study is based on a regression discontinuity design that capitalizes on the systematic process used by a number of school districts to allocate their Reading First funds.[1] A regression discontinuity design is the strongest quasi-experimental method that exists for estimating program impacts. Under certain conditions (which are met by the present study) this method can approach the rigor of a randomized experiment.[2] The conditions include:

1) Schools eligible for Reading First grants were rank-ordered for funding based on a quantitative rating, such as an indicator of past student reading performance or poverty.

2) A cut-point in the rank-ordered priority list separated schools that did or did not receive Reading First grants, and this cut-point was set without knowing which schools would then receive funding.

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

4) The shape of the relationship between schools' ratings and outcomes is correctly modeled.

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

Exhibit B.1 illustrates how the dividing line, or "cut-point," between the last funded school and the first school *not funded* on the district's priority list (or between the 10th and 11th schools on this hypothetical district's list) creates a "discontinuity" that makes it possible to estimate program impacts on future outcomes. The vertical axis of the exhibit represents a future outcome measure for each school, such as its

---

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

[2] Regression discontinuity analysis was introduced by Thistlethwaite and Campbell (1960) and has more recently experienced a resurgence of interest (e.g., Cappelleri et al., 1991; Cook, 2008; Goldberger, 1972; Hahn, Todd and Van Der Klaauw, 2001; Mohr, 1995; Reichardt, Trochim, and Cappelleri, 1995; and Trochim, 1990).

average student reading score in a subsequent year. The horizontal axis represents the rating used to determine each school's priority for Reading First (in this example, the percentage of past students eligible for free or reduced price meals). Schools to the left of the cut-point do not receive Reading First funding and serve as a "comparison group" for the impact analysis; these schools are referred to as non-Reading First schools. Schools to the right of the cut-point receive Reading First funding; these schools represent the "treatment group" for the impact analysis, and are referred to as Reading First schools.

Exhibit B.1: Regression Discontinuity Analysis for a Hypothetical School District
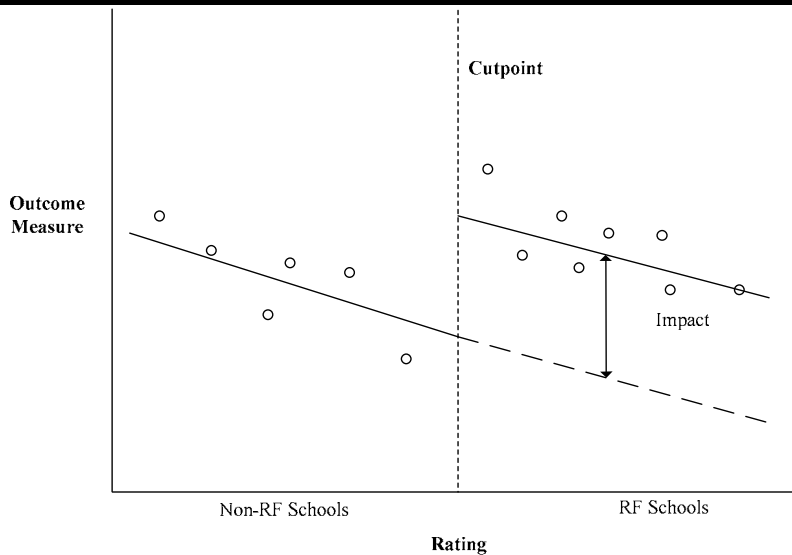
The exhibit illustrates a downward-sloping relationship between schools' ratings and their future outcomes. This implies that schools with a higher proportion of past (and thus future) students who live in poverty will tend to have lower levels of future student achievement. In the absence of Reading First, average student achievement at non-Reading First schools would therefore tend to be higher than at Reading First schools. Consequently, the average outcome for non-Reading First schools most likely over-states what this average would have been for Reading First schools without the program (their "counterfactual"). Because of this, a simple comparison of average outcomes for Reading First schools and non-Reading First schools would understate the impact of Reading First.

Given the way that schools were selected for Reading First, however, it is possible to obtain unbiased estimates of the program's impacts on future outcomes by controlling statistically for the relationships that exist between school outcomes and ratings. (These relationships comprise the "regression" part of regression discontinuity analysis.) Intuitively, this analysis would proceed as follows. The first step is to fit a regression line through the data points for non-Reading First schools, as indicated by the solid line to the left of the cut-point in Exhibit B.1. The second step is to extrapolate the fitted line across the cut-point to predict what student achievement would have been for Reading First schools—in the absence of the program. This is indicated by the dashed line in the exhibit. The third step is to fit a regression line through the data points for Reading First schools, as indicated by the solid line to the right of the cut-point. (For the purpose of this hypothetical example, the two fitted lines are assumed to have the same slope and are thus parallel, which simplifies the analysis but is not necessary.) The impact of Reading First thus can be measured by the vertical distance between the solid fitted line for Reading First schools

(what actually happened in Reading First schools after the program was launched) and the dashed extrapolated line for Reading First schools (the counterfactual prediction of what would have happened in Reading First schools without the program). This distance is indicated by a two-sided arrow.

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

The approach works properly, if schools' ratings are the only thing that determines their selection for Reading First. Consequently, only background characteristics that are correlated with ratings can be correlated with selection for the program. In other words, the only characteristics that can differ systematically between Reading First schools and non-Reading First schools are those correlated with their ratings. Controlling statistically for the ratings thereby controls for any systematic pre-existing differences between the two groups of schools.[3] It is this control that makes unbiased impact estimates possible.

Seventeen of the 18 sites in the Reading First Impact Study (16 school districts and one state program) allocated their Reading First grants in ways that meet the requirements of a regression discontinuity design. Each site prioritized its eligible schools according to a specified quantitative indicator, in most cases, an indicator based on a measure of student poverty, student performance, or both.[4] (See Exhibit B.2 for the criteria used by each site to rate its schools for Reading First.) Each site then allocated its Reading First funds according to the prioritized list, funding the top priority school first, the second priority school next, and so on through the list, until all available resources were allocated. In the context of this study, these sites are referred to as regression discontinuity design (RDD) sites.

The study sample was drawn from Reading First schools and non-Reading First schools whose ratings were as close as possible to their sites' local cut-point.[5] Half of the schools in the study sample are Reading First schools and half are non-Reading First schools.[6] Only 9 of the 248 sample schools from study sites had their rating-based Reading First funding status changed. Consequently, the study's sites support what is called a "sharp" regression discontinuity analysis, which is the strongest form of the design.[7]

---

[3] It is because regression discontinuity analysis utilizes "selection on observables" (i.e., values of the rating) that it can produce unbiased impact estimates (Cain, 1975). This feature is what distinguishes the approach from other quasi-experimental designs.
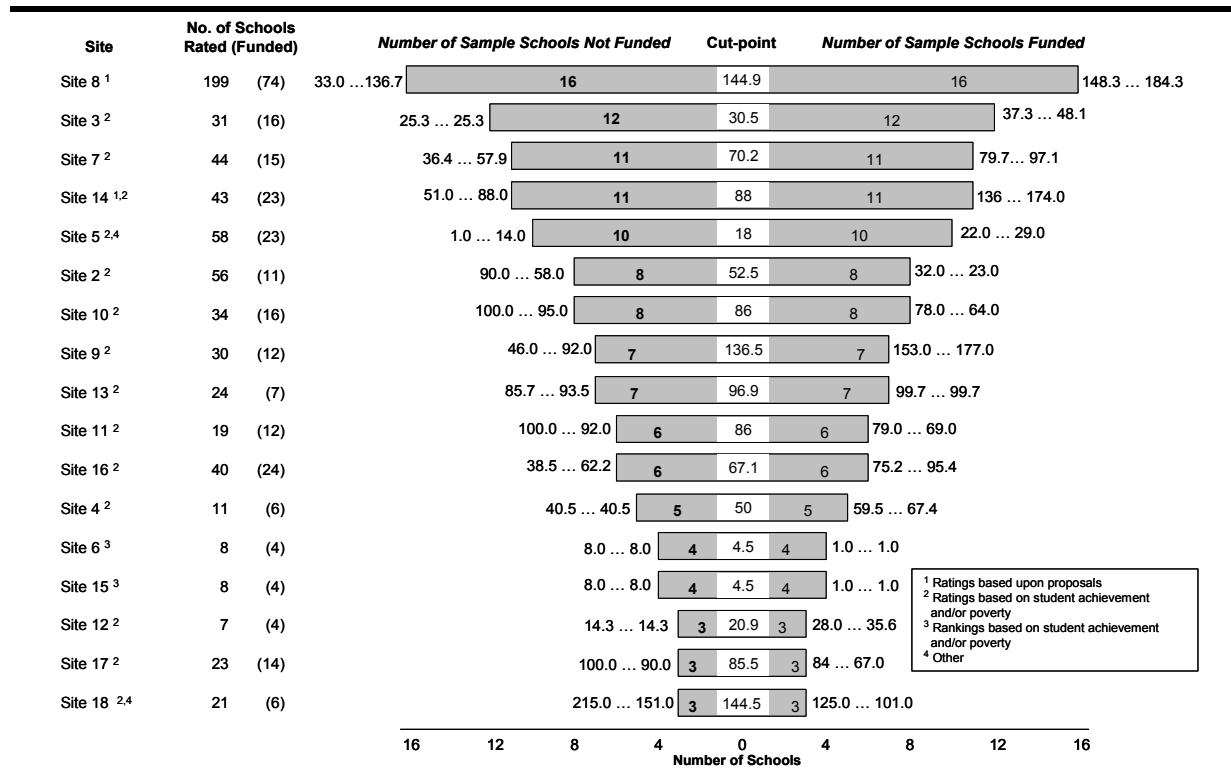
[4] A separate rating coefficient (in the impact estimation model) was specified for each site to account for differences in rating variables and cut-points. These differences enhance the generalizability of the present study because it comprises 17 regression discontinuity analyses from different parts of the United States.

[5] Note that the RDD can be compromised if there is little or no variation on the rating variable within treatment and comparison groups in a given site. As illustrated in Exhibit B.2, however, the schools selected for the study sample were both close to their local cutpoints and varied with respect to the rating variable. Therefore, this potential problem was not present in the study sample.

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

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

Exhibit B.2: Numbers, Ratings, and Cut-points for Selection of Reading First and Reading First Impact Study Schools, by Site (Initial Sample for 17 Sites, Excluding Random Assignment Site)

Site	No. of Schools Rated (Funded)	Number of Sample Schools Not Funded	Cut-point	Number of Sample Schools Funded
Site 8 [1]	199 (74)	33.0 ...136.7 [16]	144.9	[16] 148.3 ... 184.3
Site 3 [2]	31 (16)	25.3 ... 25.3 [12]	30.5	[12] 37.3 ... 48.1
Site 7 [2]	44 (15)	36.4 ... 57.9 [11]	70.2	[11] 79.7... 97.1
Site 14 [1,2]	43 (23)	51.0 ... 88.0 [11]	88	[11] 136 ... 174.0
Site 5 [2,4]	58 (23)	1.0 ... 14.0 [10]	18	[10] 22.0 ... 29.0
Site 2 [2]	56 (11)	90.0 ... 58.0 [8]	52.5	[8] 32.0 ... 23.0
Site 10 [2]	34 (16)	100.0 ... 95.0 [8]	86	[8] 78.0 ... 64.0
Site 9 [2]	30 (12)	46.0 ... 92.0 [7]	136.5	[7] 153.0 ... 177.0
Site 13 [2]	24 (7)	85.7 ... 93.5 [7]	96.9	[7] 99.7 ... 99.7
Site 11 [2]	19 (12)	100.0 ... 92.0 [6]	86	[6] 79.0 ... 69.0
Site 16 [2]	40 (24)	38.5 ... 62.2 [6]	67.1	[6] 75.2 ... 95.4
Site 4 [2]	11 (6)	40.5 ... 40.5 [5]	50	[5] 59.5 ... 67.4
Site 6 [3]	8 (4)	8.0 ... 8.0 [4]	4.5	[4] 1.0 ... 1.0
Site 15 [3]	8 (4)	8.0 ... 8.0 [4]	4.5	[4] 1.0 ... 1.0
Site 12 [2]	7 (4)	14.3 ... 14.3 [3]	20.9	[3] 28.0 ... 35.6
Site 17 [2]	23 (14)	100.0 ... 90.0 [3]	85.5	[3] 84 ... 67.0
Site 18 [2,4]	21 (6)	215.0 ... 151.0 [3]	144.5	[3] 125.0 ... 101.0

[1] Ratings based upon proposals
[2] Ratings based on student achievement and/or poverty
[3] Rankings based on student achievement and/or poverty
[4] Other

Number of Schools axis: 16 12 8 4 0 4 8 12 16

NOTES:

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

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

SOURCES: Interviews with sites' Reading First coordinators in 2004.

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

Sample Size

Although regression discontinuity analysis can provide unbiased impact estimates under the conditions met by this study—and thus is comparable to a true experiment in this regard—the quasi-experimental approach requires a much larger sample of schools to provide the same precision as an experiment because one must include the rating variable in any models to account for the design effect (Bloom, Kemple and Gamse, 2004). The study team conducted analyses of the effect of including the rating of schools as a covariate for a regression discontinuity analysis of program impacts. The team determined that if ratings are used as a covariate, the variance of the impact estimator for a regression discontinuity

analysis will be four times that for a corresponding experiment.[8] Hence, to achieve the same minimum detectable effect the regression discontinuity analysis would need four times as many schools as the experiment.

Based on these analyses and extensive discussions among members of the research team, IES staff, and the project's technical work group , it was decided that a sample of roughly 240 schools was needed, which is four times the sample size planned for the original experimental design. This larger sample size was necessary for the study to achieve a minimum detectable effect size of 0.20 standard deviations. As noted above, initial recruitment efforts produced a sample of 258 schools from one state site and 17 district sites. These 18 sites represent a total of 13 states. Due to refusals, school closings, reconfiguring, or redistricting, 10 schools (4 RF schools and 6 non-RF schools) subsequently dropped out of the study. For results presented in this report, a final analytic sample of 248 schools was used. (See Exhibit B.3 for a flowchart of sample selection from regression discontinuity design target sample to the final analytic sample.)

Specification Tests
In developing the study sample, Reading First schools and non-Reading First schools were selected to be as close as possible to their local cut-points for receipt of Reading First funding. This was done to yield two groups of schools that were as similar as possible. (See Exhibit B.4 for unadjusted baseline characteristics of schools in the study sample.) In addition, program impacts were estimated using a linear regression discontinuity model that controls for values of the ratings used to choose schools for program funding. Furthermore, estimates of impacts on measures of student reading achievement control explicitly for school-level baseline measures of reading achievement. This *combination* of sample design and statistical analysis was expected to provide internally valid estimates of program impacts.

Three sets of specification tests were conducted to assess whether this expectation was met.[9] Although none of these tests by itself can *prove* that internal validity was achieved, in combination they provide evidence that this is most likely the case. The most important such test used a linear regression discontinuity model to compare baseline characteristics of Reading First schools and non-Reading First schools. If a linear regression discontinuity model is an appropriate way to control for all pre-existing differences between the two groups, observable or not, then it should eliminate their observed baseline differences.

Baseline specification tests were conducted using aggregate school-level baseline characteristics.[10] The results of these tests in Exhibit B.5 show that none of the adjusted residual differences between Reading First schools and non-Reading First schools for the selected baseline characteristics were statistically significant. Hence, there is little evidence of residual differences in these school-level baseline characteristics. Results of these tests do not provide statistical evidence of substantial bias in impact estimates for the present report. Also, because impact estimates for student reading comprehension control explicitly for observed differences in school-level mean baseline test scores (typically the strongest predictor of future test scores), they provide further protection against bias.
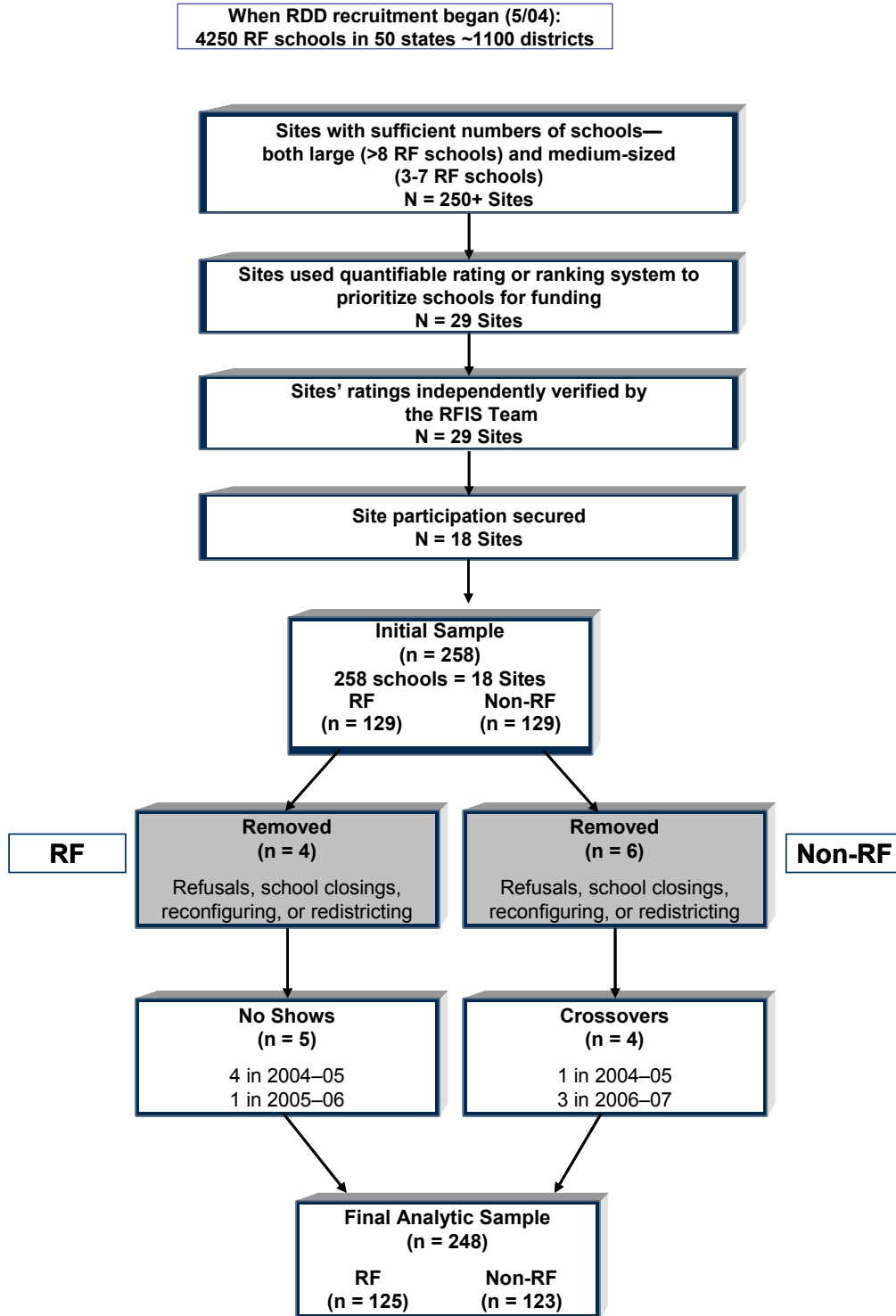
---

[8] See the study's Interim Report Appendix B, Part 5 (Gamse, Bloom, Kemple & Jacob, 2008) for details of these analyses.

[9] See the study's Interim Report Appendix B (Gamse, Bloom, Kemple & Jacob, 2008) for a detailed presentation of the specification tests conducted to assess the study's internal validity.

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

Exhibit B.3: RFIS Sample Selection: From Regression Discontinuity Design Target Sample to Analytic Sample

```
┌─────────────────────────────────────────────┐
│ When RDD recruitment began (5/04):          │
│ 4250 RF schools in 50 states ~1100 districts│
└─────────────────────────────────────────────┘

┌─────────────────────────────────────────────┐
│ Sites with sufficient numbers of schools—   │
│ both large (>8 RF schools) and medium-sized │
│ (3-7 RF schools)                            │
│ N = 250+ Sites                              │
└─────────────────────────────────────────────┘

┌─────────────────────────────────────────────┐
│ Sites used quantifiable rating or ranking   │
│ system to prioritize schools for funding    │
│ N = 29 Sites                                │
└─────────────────────────────────────────────┘

┌─────────────────────────────────────────────┐
│ Sites' ratings independently verified by    │
│ the RFIS Team                               │
│ N = 29 Sites                                │
└─────────────────────────────────────────────┘

┌─────────────────────────────────────────────┐
│ Site participation secured                  │
│ N = 18 Sites                                │
└─────────────────────────────────────────────┘

┌─────────────────────────────────────────────┐
│ Initial Sample                              │
│ (n = 258)                                   │
│ 258 schools = 18 Sites                      │
│ RF              Non-RF                       │
│ (n = 129)       (n = 129)                    │
└─────────────────────────────────────────────┘
```

RF

Removed (n = 4)	Removed (n = 6)
Refusals, school closings, reconfiguring, or redistricting	Refusals, school closings, reconfiguring, or redistricting

Non-RF

No Shows (n = 5)	Crossovers (n = 4)
4 in 2004–05	1 in 2004–05
1 in 2005–06	3 in 2006–07

```
┌─────────────────────────────────────────────┐
│ Final Analytic Sample                       │
│ (n = 248)                                   │
│ RF              Non-RF                       │
│ (n = 125)       (n = 123)                    │
└─────────────────────────────────────────────┘
```

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

Exhibit B.4: Observed Differences in Baseline Characteristics of Schools in the Study Sample: 2002-2003

Characteristic	Actual Mean for Reading First Schools	Actual Mean for Non-Reading First Schools	Difference	Statistical Significance of Difference (p-value)
Students				
Male (%)	52.3	51.6	0.7*	(0.049)
Race (%)				
Asian	3.1	3.3	-0.2	(0.670)
Black	35.6	33.9	1.7	(0.532)
Hispanic	26.7	22.5	4.1*	(0.021)
White	34.2	39.8	-5.6*	(0.006)
American Indian/Alaskan	0.5	0.5	0.0	(0.847)
Free Lunch and Reduced Lunch (%)	74.4	68.9	5.5*	(0.002)
Schools				
Eligible for Title I (%)	97.6	90.7	6.9*	(0.013)
Locale (%)				
Large City	39.2	37.4	1.8	(0.476)
Mid-size City	36.8	34.6	2.2	(0.434)
Other[a]	24.0	28.0	-4.0	(0.286)
Size				
Total Number of Students	474.8	488.7	-13.9	(0.462)
Number of Students in Grade 3	71.6	76.0	-4.4	(0.162)
Student/Teacher Ratio	15.1	15.2	-0.1	(0.613)
Third Grade Reading Performance				
Deviation from State RF Mean Proficiency Rate (%)[b]	-1.3	1.8	-3.0*	(0.019)

NOTES:

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

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

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

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

EXHIBIT READS: On average, , 52.3 percent of students in Reading First schools and 51.6 percent of students in non-Reading First schools were male. The difference on the percent of male students between Reading First and non-Reading First schools was 0.7 percentage points. The difference was statistically significant at the p≤.05 level (p=.049).

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

Exhibit B.5: Estimated Residual Differences in Baseline Characteristics of Schools in the Study Sample: 2002-2003

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

NOTES:

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

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

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

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

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

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

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

Part 2: Estimation Methods

The slightly different statistical models used to estimate the impact of Reading First on the three major outcome domains (student reading achievement, classroom instruction, and student engagement with print), as well as surveys, shared most elements. However, because there were some differences in the models for reading achievement, classroom instruction and student engagement with print and surveys, the approach for each is described separately below.

Impact tables throughout the report and appendices contain: 1) the actual, unadjusted mean outcomes for Reading First schools in the study sample ("Actual Mean with Reading First"), 2) the best estimate of what would have happened in RF schools absent RF funding ("Estimated Mean without Reading First"), 3) the impact estimate, [11] and 4) the effect size of the impact estimate.[12, 13]

Impact Estimation Method for Reading Achievement

The statistical model used to estimate RF impacts on student reading comprehension and decoding is described by (1) below:

$$
Y_{ijkm} = \sum_{mt} \beta_{0m} S_{mk} YR_t + \sum_m \beta_{1m} S_{mk} T_k + \sum_{mt} \beta_{2m} S_{mk} R_k YR_t + \sum_{mt} \beta_{3m} S_{mk} \overline{Y_{-1km}}\, YR_t \tag{1}
$$

$$
+ \sum_t \gamma_t Z_{jk} YR_t + \sum_{nt} \theta_n X_{nijkm} YR_t + \mu_k + \upsilon_{jk} + \varepsilon_{ijk}
$$

where:

Y_{ijkm} = the post-test for student i from classroom j in school k in site m,

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

T_k = one if school k is a treatment school and zero otherwise,

R_k = the rating for school k (standardized and centered by site),

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

YR_t = indicator for follow-up years, 2005, 2006, or 2007[14]

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

X_{nijkm} = demographic characteristic n of student i from classroom j in school k, in site m

μ_k, υ_{jk} and ε_{ijk} = school-level, classroom- level, and student-level random error terms, respectively, assumed to be independently and identically distributed.

---

[11] The estimates of what would have happened in RF schools absent RF funding are calculated by subtracting the impact estimates from the RF schools' actual mean values.

[12] 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 all years for which the outcome was available.

[13] When calculating the effect sizes, standard deviation from the non-Reading First schools were used instead of the pooled standard deviation from Reading First and non-Reading First schools because the treatment could have effected the distribution—hence the standard deviation—of the outcomes in Reading First schools but not in non-Reading First schools. The study team wanted to use a stable standard deviation and non-Reading First schools provided that. It is also important to note that the standard deviations for the student outcomes obtained from non-Reading First schools are very close to those observed in the national norming sample.

[14] For decoding, this indicator is not used because only one year of data is available for this outcome.

The average estimated value of β_{1m} (m = 1, 2, …, 18), weighted by the number of RF schools in each site, is the program impact for the average RF school in the study sample.

The student achievement impact model (Equation 1) used to estimate impacts on reading comprehension and decoding has the following characteristics:

- It is a multi-level model that reflects the nested structure of the data by accounting for three levels of clustering in the estimation of standard errors: clustering of students within classrooms, classrooms within schools, and schools within study sites.

- Baseline covariates are added to the model to improve precision. These covariates include student gender, student age at start of school year,[15] date of the post-test at the classroom level, and a school-level pre-program reading performance measure.[16, 17]

- The rating variable was not included in the model for the one site that assigned schools to Reading First and non-Reading First groups randomly.

- In estimating reading comprehension pooled impacts for the combined sample from 2005, 2006 and 2007, the covariates for site, rating, pretest, test date, and demographic characteristics were interacted with an indicator for follow-up year (2005, 2006 or 2007).

- In estimating decoding impacts, the covariates for site, rating, pretest, test date, and demographic characteristics were not interacted with an indicator for follow-up year because there is only one year of data for this outcome.

Impacts on Classroom Instruction and Student Engagement with Print
The impacts of Reading First on classroom instruction and student engagement with print were estimated using the following three-level model (with observations at level one, classrooms at level two, and schools at level three):

$$Y_{ijkm} = \sum_{mt} \beta_{0m} S_{mk} YR_t + \sum_{m} \beta_{1m} S_{mk} T_k + \sum_{mt} \beta_{2m} S_{mk} R_k YR_t + \mu_k + \nu_{jk} + \varepsilon_{ijk} \qquad (2)$$

Where:

Y_{ijkm} = the outcome measure for observation i from classroom j in school k in site m,
S_{mk} = one if school k is in site m and zero otherwise, (m= 1,2, …, 18),
T_k = one if school k is a treatment school and zero otherwise,
R_k = the rating for school k (standardized and centered by site),
YR_t = indicator for follow-up years, 2005, 2006, or 2007,[18]

---

[15] Age at start of the school year is each student's age as of September 1 of the given year. For example, age as of September 1, 2005 for the 2005-2006 school year.

[16] Different pre-program performance measures were constructed for early and late award sites. For the 10 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, we used 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.

[17] As a robustness test, the analysis was conducted without some or all of these additional covariates and the impact estimates stayed virtually unchanged. *Results for these additional tests are available upon request.*

[18] For the STEP, only two year indicators are included in the model, since STEP data was not collected in the first year.

μ_k, ν_{jk} and ε_{ijk} = school-level, classroom-level, and observation-level random error terms, respectively, assumed to be independently and identically distributed.

The impact estimate is the average estimated value of β_{1m} (m = 1, 2, …, 18) weighted by number of treatment schools in each site.

The impact estimation model for classroom instruction and student engagement with print described by (Equation 2) has the following characteristics:

- It is a multi-level model that reflects the nested structure of the data by accounting for three levels of clustering in the estimation of standard errors: clustering of observation days within classrooms, classrooms within schools, and schools within sites.

- A rating variable was not included in the model for the one site that assigned schools to Reading First and non-Reading First groups randomly.

- In estimating pooled impacts for the combined sample from 2005, 2006 and 2007, the covariates for site and rating were interacted with an indicator for follow-up year (2005, 2006, or 2007).

Estimation Method for Surveys

Data from self-report surveys of teachers, reading coaches, and principals were used to estimate the impact of Reading First on the key components of scientifically based reading instruction.[19] Two models were needed to estimate differences for survey data—one for classroom level data (i.e., from teacher survey) and a second for school level data (i.e., from reading coach survey or principal survey). Differences for classroom level survey data were estimated using the following two-level model (with classrooms at level one and schools at level two):

$$Y_{jkm} = \sum_{mt} \beta_{0m} S_{mk} + \sum_{m} \beta_{1m} S_{mk} T_k + \sum_{mt} \beta_{2m} S_{mk} R_k + \nu_k + \varepsilon_{jk} \qquad (3)$$

Where:

Y_{ijkm} = the outcome measure for classroom j in school k in site m,[20]

S_{mk} = one if school k is in site m and zero otherwise, (m= 1,2, …, 18),

T_k = one if school k is a treatment school and zero otherwise,

R_k = the rating for school k (standardized and centered by site),

ν_k and ε_{jk} = school-level and classroom-level random error terms, respectively, assumed to be independently and identically distributed.

The difference estimate is the average estimated value of β_{1m} (m = 1, 2, …, 18) weighted by number of treatment schools in each site.

The impact estimation model for classroom level survey data described by (Equation 3) has the following characteristics:

---

[19] Only 2007 survey data is included in these analyses due to low survey response rates in 2005.

[20] To maintain parallel structure with other estimation models presented in this appendix, the nomenclature for classroom (j), school (k), and site (m) remains the same even in the absence of student or observation level survey data (i).

- It is a multi-level model that reflects the nested structure of the data by accounting for two levels of clustering in the estimation of standard errors: clustering of classrooms within schools and schools within sites.

- A rating variable was not included in the model for the one site that assigned schools to Reading First and non-Reading First groups randomly.

- Only one year of data were used for survey data, so no interactions with the follow-up year were included in the estimation model.

Differences for school level survey data were estimated using the following ordinary least squares regression model:

$$Y_{km} = \sum_{mt} \beta_{0m} S_{mk} + \sum_{m} \beta_{1m} S_{mk} T_k + \sum_{mt} \beta_{2m} S_{mk} R_k + \varepsilon_k \tag{4}$$

Where:

Y_{ijkm} = the outcome measure for school k in site m,[21]

S_{mk} = one if school k is in site m and zero otherwise, (m= 1,2, …, 18),

T_k = one if school k is a treatment school and zero otherwise,

R_k = the rating for school k (standardized and centered by site),

ε_k = school-level random error term assumed to be independently and identically distributed.

The difference estimate is the average estimated value of β_{1m} (m = 1, 2, …, 18) weighted by number of treatment schools in each site.

The impact estimation model for school level survey data described by (Equation 4) has the following characteristics:

- It is a single-level ordinary least squares regression model that accounts for one level of clustering in the estimation of standard errors: clustering of schools within sites.

- A rating variable was not included in the model for the one site that assigned schools to Reading First and non-Reading First groups randomly.

- Only one year of data were used for survey data, so no interactions with the follow-up year were included in the estimation model.

Part 3: Approach to Multiple Hypothesis Testing

This section addresses the issue of multiple hypothesis testing. It first summarizes the five core principles that were used as a guide for addressing the issue in the current study, and then describes a two-stage approach for operationalizing these principles.

---

[21] To maintain parallel structure with other estimation models presented in this appendix, the nomenclature for school (k) and site (m) remains the same even in the absence of student or observation level survey data (i) or classroom level survey data (j).

Principle #1: Qualify tests instead of adjusting them: The present analysis qualifies specific hypothesis tests using composite tests of pooled hypotheses rather than (1) adjusting significance levels (through Bonferroni methods) or (2) adjusting significance thresholds (through Benjamini and Hochberg methods) of specific tests.

Principle #2: Address multiple testing differently for the central research questions of the study and for supplemental analyses. The analysis specifies two tiers of hypotheses: Tier I comprises a small number of hypotheses about the central research questions of the study, and Tier 2 represents supplemental research questions. Multiple testing is treated separately and differently within the two tiers. Statistical tests of Tier I hypotheses are considered confirmatory. To address the issue of multiplicity within Tier I, the present study tested a reduced set of outcomes by conducting pooled tests of composite hypothesis that represent a set of hypotheses that have been tested separately. The Tier 2 hypothesis tests are allowed to be much larger and less confirmatory. It may or may not be necessary to qualify these findings for multiple testing since they are not confirmatory.

Principle #3: Delineate separate domains that reflect key clusters of constructs represented by the central research questions of a study. Domains comprise broad clusters of outcome constructs that can contain multiple measures, subgroups, or follow-up observations. Domains are defined conceptually, and do not provide narrow "silos" for collecting findings. The central domains for the present study are student reading achievement, classroom reading instruction, and student engagement with print. In addition, survey data is a domain for exploratory analyses of support for scientifically based reading instruction across study schools.

Principle #4: Report analyses to address multiple comparisons in the background of research reports, not in the foreground. For the present study references to the qualifying tests occur in the main text but not in tables.

Principle #5: Use tests for interactions as a composite test (and thus a guide) for focusing on subgroup findings.[22]

Based on the above five principles, the present study uses the following two-stage approach to address multiple hypothesis testing. The first stage involves prioritizing outcomes and subgroups for the impact analysis. The second stage encompasses strategies for conducting composite tests on pooled key outcomes. The core features of each stage are described below.

Stage 1: Creating a Parsimonious List of Outcomes and Subgroups and Prioritizing Key Outcomes

The first stage of the framework involves a process of carefully categorizing and prioritizing the outcomes and subgroups for the impact analysis. The goal of this exercise is to create the shortest possible list of outcomes and subgroups that reflect the most proximal and policy relevant indicators of Reading First's effectiveness. Analytically, the shorter the list, the less likely it is that one would attribute

---

[22] If differences between impacts for subgroups are not statistically significant, then individual subgroup results should be interpreted with caution.

statistical significance to an impact that did not truly occur. These outcomes and subgroups were selected within distinct measurement domains to correspond to key components of the program's theory of action and the key research questions posed by the program's evaluation.

The impact analysis focuses on two components of the Reading First theory of action: 1) aligning teachers' instructional practices and behaviors with the five dimensions of reading instruction,[23] and 2) improving students' reading achievement.[24] The highest priority outcomes within each of these measurement domains would constitute "Tier 1" outcomes for the impact analysis.

Recognizing that a short list of outcomes will almost certainly exclude important policy-relevant indicators of Reading First's effectiveness (a form of Type II error), this first stage of the framework also includes the development of a secondary, or "Tier 2," list of outcomes and subgroups. As discussed below, the present study treats Tier 1 and Tier 2 outcomes and their accompanying subgroups separately, and potentially differently, if or when making adjustments to the standards used for judging statistical significance.

Exhibit B.6 provides a list of the Tier 1 and Tier 2 outcomes defined for each measurement domain for this report. Also displayed are the grade levels and follow-up periods on which the impact analyses focus.

Stage 2: Conducting Composite Tests to Qualify Specific Hypothesis Tests

One approach to qualifying multiple hypothesis tests is to test whether the overall effect of treatment on a family of outcomes is significantly different from zero. For example, a policy maker may be interested in the effect of an intervention on test scores in general, rather than on each subject separately. Measurement of such *overall* effects has its roots in the literature on clinical trials and on meta-analysis (O'Brien, 1984; Logan and Tamhane, 2002; and Hedges and Olkin, 1985). The present analysis constructs summary indices that aggregate information over multiple treatment effect estimates within each domain for Tier 1 outcomes, as well as for survey constructs in Tier 2. See Exhibit B.7.

Reading Comprehension

To qualify the impact estimates for each outcome measure for each grade in the reading comprehension domain, the present analysis ran a composite regression that pooled the sample across grades 1, 2, and 3 and two measures: scaled scores and an indicator of whether or not a student scored at or above grade

---

[23] The RFIS observational instrument, the IPRI, focused primarily on teacher behaviors. In order to ensure that the study also collected some data on student behaviors during observed reading instruction, the RFIS team developed the Student Time-on-Task Engagement with Print (STEP) instrument. Because student engagement with print is an outcome that is distinct from the student reading comprehension or classroom reading instruction domains, it is treated separately.

[24] The Reading First theory of action also includes allocating additional resources for districts and schools to purchase reading curricula, materials, and assessments; exposing teachers to professional development and coaching focused on the five dimensions of effective reading programs; and holding districts and schools accountable for improved reading achievement. The present study was able to measure the impact of Reading First on some of these other elements using survey data.

Exhibit B.6: Outcome Tiers for the Reading First Impact Analysis

| Tier | Domain | Outcome | Impacts Estimate | | | Variation |
			Year	Grade	Sample	
Tier 1	**Student Achievement**	*Reading Comprehension Scaled Score (SAT 10)*	2005, 2006, 2007 Pooled	Separate for Grade 1, 2, 3	Full	**N/A**
		% At or Above Grade Level (SAT 10)	2005, 2006, 2007 Pooled	Separate for Grade 1, 2, 3	Full	
		TOSWRF Standard Score	2007	Separate for Grade 1	Full	
	Instruction	*Time on Five Dimensions*	2005, 2006, 2007 Pooled	Separate for Grade 1, 2	Full	**N/A**
		Highly Explicit Instruction	2005, 2006, 2007 Pooled	Separate for Grade 1, 2	Full	
		High Quality Practice	2005, 2006, 2007 Pooled	Separate for Grade 1, 2	Full	
	Student Engagement with Print	*% Students Engaged with Print*	2006, 2007 Pooled	Separate for Grade 1, 2	Full	**N/A**
Tier 2	**Student Achievement**	*Reading Comprehension Scaled Score (SAT 10)*	2005	Separate for Grade 1, 2, 3	Full, Award Subgroup	• **Variation over time**
			2006	Separate for Grade 1, 2, 3	Full, Award Subgroup	• **Variation across sites**
			2007	Separate for Grade 1, 2, 3	Full, Award Subgroup	• **Variation by Award Subgroup**
			2005, 2006, 2007 Pooled	Separate for Grade 1, 2, 3	Full, Award Subgroup	
			2007	Separate for Grade 3	2005/2007 Stayer Subgroup	• **Variation by Student Exposure**
		% At or Above Grade Level (SAT 10)	2005	Separate for Grade 1, 2, 3	Full	**N/A**
			2006	Separate for Grade 1, 2, 3	Full	
			2007	Separate for Grade 1, 2, 3	Full	
			2005, 2006, 2007 Pooled	Separate for Grade 1, 2, 3	Full, Award Subgroup	
	Instruction	*Time on Five Dimensions (Combined and for Five Dimensions separately)*	2005	Separate for Grade 1, 2	Full, Award Subgroup	• **Variation over time**
			2006	Separate for Grade 1, 2	Full, Award Subgroup	• **Variation across sites**
			2007	Separate for Grade 1, 2	Full, Award Subgroup	• **Variation by Award Subgroup**
			2005, 2006, 2007 Pooled	Separate for Grade 1, 2	Full, Award Subgroup	

Exhibit B.6: Outcome Tiers for the Reading First Impact Analysis (continued)

Tier	Domain	Outcome	Impacts Estimate			Variation
			Year	Grade	Sample	
Tier 2	**Instruction**	*Highly Explicit Instruction*	2005	Separate for Grade 1, 2	Full	**N/A**
			2006	Separate for Grade 1, 2	Full	
			2007	Separate for Grade 1, 2	Full	
			2005, 2006, 2007 Pooled	Separate for Grade 1, 2	Full, Award Subgroup	
		High Quality Practice	2005	Separate for Grade 1, 2	Full	**N/A**
			2006	Separate for Grade 1, 2	Full	
			2007	Separate for Grade 1, 2	Full	
			2005, 2006, 2007 Pooled	Separate for Grade 1, 2	Full, Award Subgroup	
	Student Engagement with Print	*% Students Engaged with Print*	2006	Separate for Grade 1, 2	Full	**N/A**
			2007	Separate for Grade 1, 2	Full	
	Professional Development in SBRI	*Amount of PD in reading received by teachers*	2007	Grades 1, 2, 3 Pooled	Full	**N/A**
		Teacher receipt of PD in the five essential components of reading instruction	2007	Grades 1, 2, 3 Pooled	Full	
		Teacher receipt of coaching	2007	Grades 1, 2, 3 Pooled	Full	
		Amount of time dedicated to serving as K-3 reading coach	2007	Grades 1, 2, 3 Pooled	Full	

Exhibit B.6: Outcome Tiers for the Reading First Impact Analysis (continued)

| Tier | Domain | Outcome | Impacts Estimate | | | Variation |
			Year	Grade	Sample	
Tier 2	**Amount of Reading Instruction**	*Minutes spent on reading instruction per day*	2007	Grades 1, 2, 3 Pooled	Full	**N/A**
	Supports for Struggling Readers	*Availability of differentiated instructional materials for struggling readers*	2007	Grades 1, 2, 3 Pooled	Full	**N/A**
		Provision of extra classroom practice for struggling readers	2007	Grades 1, 2, 3 Pooled	Full	
	Use of Assessments	*Use of assessments to inform classroom practice*	2007	Grades 1, 2, 3 Pooled	Full	**N/A**

Exhibit B.7: Summary of Impacts and Results of Composite Tests

Outcome Measure	Impact (p-value)			Result of Composite Test
	Grade 1	Grade 2	Grade 3	
Reading Comprehension				
• Standard scaled score	4.74 (p=0.083)	1.69 (p=0.462)	0.30 (p=0.887)	p=0.957 for composite test across 3 grades and 2 outcomes
• Percent reading at or above grade level	4.22 (p=0.104)	1.60 (p=0.504)	-0.08 (p=0.973)	
Instruction				
• Minutes of instruction in 5 reading dimensions	6.92* (p=0.005)	9.79* (p<0.001)	--	p<0.001 for composite test across 2 grades and 3 outcomes
• Highly explicit instruction	3.29* (p=0.018)	3.00* (p=0.040)	--	
• High quality student practice	0.82 (p=0.513)	2.94* (p=0.019)	--	
Student Engagement with Print				
• Percent of students engaged with print	5.33 (p=0.070)	-4.75 (p=0.104)	--	p=0.845 for composite test across 2 grades and 1 outcome
Key Components of Scientifically Based Reading Instruction at the School-Level				
• Amount of time dedicated to serving as K-3 reading coach	33.49* (p<0.001)			p=0.009 for composite test across 2 outcomes
• Availability of differentiated instructional materials for struggling readers	0.01 (p=0.661)			
Key Components of Scientifically Based Reading Instruction at the Classroom-Level (aggregated across all three grade levels)				
• Amount of PD in reading received by teachers	12.13* (p<0.001)			p<0.001 for composite test across 3 grades and 6 outcomes
• Teacher receipt of PD in the five essential components of reading instruction	0.55* (p=0.010)			
• Teacher receipt of coaching	0.20* (p<0.001)			
• Minutes of reading instruction per day	18.47 (p<0.001)			
• Provision of extra classroom practice for struggling readers	0.19 (p=0.018)			
• Use of assessments to inform classroom practice	0.18 (p=0.090)			

NOTES:

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

EXHIBIT READS: The result of the composite test for reading comprehension test scores, across three grades and two outcomes, is not statistically significant (p=.957).

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; RFIS Student Time-on-Task and Engagement with Print, fall 2005, spring 2006, fall 2006, and spring 2007, and RFIS Survey administration, spring 2007.

level.[25] To qualify the six multiple hypotheses tests for these outcomes, the RFIS Team created one parsimonious index. The aggregation improves statistical power to detect effects that go in the same direction within a domain. The summary index is defined to be the equally weighted average of z-score outcome components, with the sign of each measure oriented so that more beneficial outcomes have higher scores.[26]

Specifically, the present analysis took the following steps in creating a composite index and conducting the analysis:[27]

1. First, z-scores were created for each outcome component in the reading comprehension domain by subtracting the unadjusted non-RF mean (pooled across years and grade levels) and dividing by its standard deviation (pooled across years and grade levels). Thus, each component of the index has a mean of zero and a standard deviation of one for the non-RF group.

2. If an observation unit has a valid response to at least one component measure of the index, then any missing values of other component measures are imputed as the random assignment group mean. This results in differences between RF and non-RF means of an index being the same as the average of those two groups' means of the components of that index (when the components are divided by their comparison group standard deviation and have no missing value imputation), so that the index can be interpreted as the average of results for separate measures scaled in standard deviation units.[28]

3. The z-scores from each component were averaged to obtain the index and an impact analysis was run on this index using a sample that pooled all years and all grade levels together.

This regression addresses the question whether overall the program "worked" in terms of improving student achievement. This result serves as a "qualifier" to the small number of specific hypothesis tests shown in impact tables.[29]

---

[25] Although decoding is considered to be in the same domain as comprehension, it was not possible to include the TOSWRF scores in the composite because scores are available for only one grade in one year.

[26] An alternative is to use seemingly unrelated regression effects for specific outcomes to estimate the covariance of the effects and then to calculate the mean effect size for groups of estimates in a second step. The average z-score index approach is much simpler to work with. The two approaches yield identical treatment effects when there is no item nonresponse and no regression adjustment (Kling, Liebman, and Katz, 2007).

[27] The discussion and method presented here draw from Kling, Liebman, and Katz (2007).

[28] No data imputation was done in constructing the reading achievement composite index.

[29] Though decoding is included in the student achievement domain, it is not possible to include this outcome measure in the summary index with the reading comprehension outcomes. As noted earlier, the decoding test was administered in only one grade for one year. In addition, 2,158 students who were administered the reading comprehension test in that year and grade were not given the decoding test. Therefore, attempts to calculate a common index for the decoding and reading comprehension measures would have required collapsing the reading comprehension data to a single year and grade and imputing missing decoding scores. This approach was not taken because this process would have resulted in a significant loss of statistical power and would have weakened the usefulness of the index as a qualifier for reading comprehension impacts.

Classroom Instruction

A similar composite analysis was conducted for the instructional domain. To qualify the impact estimates for each outcome measure for each grade in the instructional domain, the analysis ran a composite regression which pooled the sample across grades and used an index constructed from z-scores for all three instructional outcome measures as the dependent variable. The index of instruction averaged together minutes in the five dimensions of reading instruction, percentage of highly explicit instruction, and percentage of high quality student practice.[30]

The results from this analysis help to answer the research question whether *overall* the Reading First program has an impact on instructional practice.

In addition, program impacts for time spent on each of the five dimensions will be reported separately. Since the impact on total time spent on the five dimensions will already have been reported, any additional qualifying test is not necessary for these analyses.

Student Engagement with Print

A similar composite analysis was conducted for the student engagement with print outcome domain. For this domain impacts are reported for the full sample in grades 1 and 2 as the percentage of students engaged with print. To qualify the two multiple hypotheses tests for these outcomes, the RFIS Team reports the result from a composite regression which pools two grades together and represents the outcome measure in one parsimonious index, created in the same way that the composite index for reading comprehension and instruction was created.[31] This regression addresses the question whether overall the program "worked" in terms of having an impact on the percentage of students engaged with print. This result serves as a "qualifier" to the small number of specific hypothesis tests shown in impact tables.

Implementation of Key Components of Scientifically Based Reading Instruction (Surveys)

Because survey data was collected at the school level (i.e., reading coach and principal surveys) and at the classroom level (i.e., teacher surveys), two composite tests *across* domains were conducted, one at the school level and one at the classroom level. To qualify the impact estimates for each *school level* survey outcome measure, the analysis team ran a composite regression where the dependent variable was a single index created from z-scores for each of the two school level outcome variables. The index of key components of scientifically based research instruction at the school level averaged together the following

---

[30] No data imputation was done in constructing the reading instruction composite index.

[31] No data imputation was done in constructing the student engagement with print composite index.

two survey outcomes: amount of time dedicated to serving as K-3 reading coach and availability of differentiated instructional materials for struggling readers.[32]

To qualify the impact estimates for each *classroom* level survey outcome measure, the analysis team ran a composite regression where the dependent variable was a single index created from z-scores for each of the six classroom level outcome variables. The index of key components of scientifically based research instruction at the classroom level averaged together the following six survey outcomes: amount of PD in reading received by teachers, teacher receipt of PD in the five essential components of reading instruction, teacher receipt of coaching, minutes of reading instruction per day, provision of extra classroom practice for struggling readers, and use of assessments to inform classroom practice.[33]

These regression analyses address the question whether overall the program "worked" in terms of having an effect on the implementation of key components of scientifically based reading instruction, in the school and in the classroom. These results serve as a "qualifier" to the small number of specific hypothesis tests shown in impact tables.

Part 4: Statistical Precision

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

Exhibit B.8 lists the minimum detectable effect (or effect size) for full-sample estimates of program impacts on key study outcomes when the data are pooled across the school years for which data are available. These minimum detectable effects are based on the experience of students and schools in the study sample during the follow-up period, and not on the initial assumptions that guided the study design. Hence, the findings in Exhibit B.8 represent the actual precision of the present design as it materialized in the field.[34]

---

[32] Six schools were dropped from these analyses because both survey outcomes were missing. All other schools had a valid response on at least one of the survey outcomes. Missing values on individual survey outcomes ranged from 0% to 7% overall (RF: 0% to 1% and non-RF: 0% to 14%). Missing values were imputed as the random assignment group mean.

[33] All classrooms had a valid response on at least one of the six survey outcomes. Missing values on individual survey outcomes ranged from 0% to 3% overall (RF: 0% to 2% and non-RF: 0% to 5%). Missing values were imputed as the random assignment group mean.

[34] Because for the present full sample the number of degrees of freedom for estimating the standard error of an impact estimator is well beyond 30, the minimum detectable effect of an estimator equals 2.8 times its standard error. For further discussion see Bloom, H. S. (1995) "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs," Evaluation Review, Vol. 19, No. 5, pp. 547–556.

Exhibit B.8: Minimal Detectable Effects for Full Sample Impact Estimates

	Grade Level		
	Grade 1	Grade 2	Grade 3
Panel 1			
Student Achievement			
Reading Comprehension			
Mean Scaled Score	7.62	6.41	5.93
Effect Size	0.16	0.15	0.15
Percent at or above Grade Level	7.22	6.67	6.70
Decoding			
Mean Standard Score	3.14	--	--
Effect Size	0.21	--	--
Panel 2			
Instructional Outcomes			
Instruction in the Five Dimensions Combined			
Minutes	6.84	6.88	--
Effect Size	0.33	0.32	--
Percentage of Intervals in Five Dimensions with Highly Explicit Instruction			
Percentage	3.88	4.06	--
Effect Size	0.22	0.21	--
Percentage of Intervals in Five Dimensions with High Quality Student Practice			
Percentage	3.50	3.49	--
Effect Size	0.21	0.19	--
Panel 3			
Student Engagement with Print			
Percentage of Students Engaged with Print	8.18	8.15	--
Effect Size	0.28	0.28	--

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.

Minimal detectable effects are based on the standard errors of the impact estimates for the full sample pooled across three school years (except for Student Engagement with Print, which is based on two years of data) 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 data.

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

EXHIBIT READS: The minimal detectable effect of the Reading First program on reading comprehension for a mean scaled score in grade 1 is 7.62 scaled score points. The minimal detectable effect of the Reading First program on reading comprehension for a mean scaled score in grade 2 is 6.41 scaled score points. The minimal detectable effect of the Reading First program on reading comprehension for a mean scaled score in grade 3 is 5.93 scaled score points.

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 TOSWRF administration in spring 2007; RFIS Instructional Practice in Reading Inventory, spring 2005, 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.

The three panels in the exhibit present minimum detectable effects for the Tier 1 outcomes of the present study. The three columns in the exhibit present minimum detectable effects for grades one, two, and three separately.

The top panel focuses on measures of student reading comprehension and decoding. Findings in this panel indicate that the present study design and impact estimation model have minimum detectable effects for reading comprehension that range from approximately 6 to 8 scaled score points, which corresponds to 0.15 to 0.16 standard deviations or about 7 percentage points. The minimal detectable effect for decoding was approximately 3 scaled score points, which corresponds to 0.21 standard deviations. These findings indicate that the present study achieved its goal of providing minimum detectable effect sizes that are no larger than 0.20 standard deviations for estimates of the impacts of Reading First on student reading comprehension.[35]

Findings in the second panel of the exhibit indicate that the minimum detectable effect for instructional time spent in the five dimensions of reading instruction in grades 1 and 2 is approximately 7 minutes, which corresponds to 0.32 to 0.33 standard deviations.

Minimum detectable effects for the percentage of instructional intervals in the five dimensions that exhibited highly explicit instruction or that exhibited high quality student practice ranged from approximately 3 to 4 percentage points. The minimum detectable effect on the percentage of students engaged with print was between 10 and 11 percentage points, roughly twice as large as that for the preceding two measures.

On balance, the statistical precision of the present study design and its analytic framework achieve the initial goals of the study's design. The precision is adequate for full-sample impact estimates, which are the primary focus of the present study.

Part 5: Handling Missing Data

This section describes how the study handled missing data for each outcome measure and for covariates used in analyses.

Surveys

The study imputed values for several survey variables. When Reading Coach survey data were missing at the item-level, and the identical questions had been asked of school principals, principal data were used when available. The study used principals' responses for 13 schools (all non-Reading First) without Reading Coach surveys on the "availability of differentiated instructional materials for struggling readers" outcome. The study also imputed values of 0 for the "amount of time dedicated to serving as K-3 reading coach" for six (non-Reading First) schools that the study had confirmed did not have reading coaches. The study imputed data from other RF schools about the presence of a scheduled reading block in grades K-3 and about the length of the K-3 reading block when all other RF schools in the site had reported having a reading block and that the block was the same length (n=2). In one instance, where all other RF

---

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

schools in a site reported using the same core reading program, the study imputed the core reading program for that school using data from other RF schools in that site.

Missing data rates ranged from 0.1 to 3.3 percent for teacher survey outcomes (RF: 0.1 to 1.0 percent; non-RF: 0.0 to 4.9 percent) and 1.3 to 2.8 percent for reading coach and/or principal survey outcomes (RF: 0.0 to 1.6 percent; non-RF: 2.7 to 4.1 percent). Survey constructs (i.e., those outcomes comprised of multiple survey items) were constructed only for observations with complete data, with one qualification: the construct "minutes spent on reading instruction per day" was calculated as the total number of minutes reported for the previous week (across a maximum of 5 days) divided by the number of days with non-missing values. For this construct, teacher surveys with missing data for every day of the previous week were eliminated (0.9 percent).

Classroom Observations: IPRI

No imputations were made for observations that were missing in their entirety. Imputations were made for a small proportion of intervals within observations during data cleaning. When an individual observation record (comprised of 35 successive intervals, on average) contained gaps in time or blank intervals, the record was filled in to make observation interval times internally consistent, using verbatim information provided by the observer on classroom activities. When no such information was available, gaps in time were filled in with intervals coded as non-instructional. When an observation contained more than one blank interval, and the same instructional activity (at the dimension or Part B levels) had been recorded in the intervals immediately before and after the blank interval, the blank intervals were post-coded to be the same as the preceding and successive intervals' activities. If the surrounding intervals were not identical to each other, the blank intervals remained blank. In the pooled sample (spring 2005, fall 2005, spring 2006, fall 2006, spring 2007), imputations were made in 402 classroom observations (2.8 percent). Of those 402 observations, 205 were in Reading First classrooms and 197 were in non-Reading First classrooms, which corresponds to 2.9 percent of all Reading First classrooms and 2.8 percent of non-Reading First classrooms.

Classroom Observations: STEP

No data imputation was done for the STEP. Missing data were handled in the following manner. If classrooms had missing values for all three sweeps, they were dropped from the analytic sample. A sweep was coded as missing if the class was in transition between activities or the entire class was listening to a story. Only 78 classrooms (1 percent) were given a missing code data for all three sweeps. Of the 78 classrooms, 23 were Reading First and 55 were non-Reading First, which corresponds to 1 percent of Reading First classrooms and 2 percent of non-Reading First classrooms in the pooled analytic sample. In cases where classrooms had one or more non-missing values for the three sweeps, the percentage of students engaged with print was calculated by averaging across the number of sweeps available for that classroom. For the pooled analytic dataset (fall 2005, spring 2006, fall 2006, and spring 2007), 70 percent of classrooms had three sweeps of data; 23 percent had two sweeps of data; 5 percent had one sweep of data; and 1 percent were missing all three sweeps.[36]

---

[36] Numbers do not add to 100% due to rounding.

Student Reading Achievement: SAT 10 Reading Comprehension Subtest

No data imputation was done for the SAT 10. Students test scores were coded as missing and excluded from the analytic sample if they were deemed invalid according to SAT 10 scoring guidelines. For the pooled sample (Spring 2005, Spring 2006, Spring 2007), this amounted to 222 student test scores (0.2 percent). Of the missing scores, 92 were Reading First and 130 were non-Reading First, which corresponds to 0.1 percent of Reading First test scores and 0.2 percent of non-Reading First scores.

Data were imputed for three covariates used in student achievement analyses—student age, gender, and date of testing. If missing, student age and gender were imputed by using school-by-grade means, except when data on an entire grade within a school was missing, in which case the district-by-grade mean was used. When an imputed covariate was used in analysis, a dummy variable indicating the imputed observations was also included. For student age, values were imputed for 0.35 percent (0.19 percent Reading First and 0.53 percent non-Reading First) of the pooled sample. For gender, values were imputed for 0.36 percent (0.30 percent Reading First and 0.43 percent non-Reading First) of the pooled sample. For the date of testing covariate, which was recorded at the classroom level, missing values were imputed using the school mean. Values were imputed for 0.79 percent (1.44 percent Reading First and 0.06 percent non-Reading First) of the pooled sample.

Student Reading Achievement: TOSWRF

No data imputation was done for the TOSWRF. Students were coded as missing and were excluded from the analytic sample if birth dates were missing, or out of range (meaning that a standard score could not be computed for those students), or if students did not follow test instructions. Of the 465 test scores (5 percent) excluded from the analytic sample, 233 were Reading First and 232 were non-Reading First, which corresponds to 4 percent of Reading First scores and 4 percent of non-Reading First scores in the analytic sample.

Covariates used in the TOSWRF analyses were imputed in the same manner as described above for the SAT 10 analyses.

Appendix C: Measures

Appendix C, Parts 1 through 5, describe the data collection instruments and assessments used to create measures for all outcome domains assessed in the RFIS. These instruments and assessments include Reading Coach and Teacher Surveys, the Instructional Practice in Reading Inventory (IPRI), the Global Appraisal of Teaching Strategy (GATS), Student Time-on-Task and Engagement with Print (STEP), and assessments of students' reading performance (SAT 10 and TOSWRF). Parts 1 through 5 also include relevant information on properties of instruments, data collection procedures, and response rates.

Appendix C, Part 6 contains copies of each of the RFIS data collection instruments.

Part 1: Reading Coach and Teacher Surveys

Description of the Instruments

The RFIS developed survey instruments for reading coaches and classroom teachers in grades 1 through 3 to learn about how schools were implementing scientifically based reading programs. The surveys for the RFIS and the Reading First Implementation Evaluation[1] contain identical items in order to facilitate comparisons between the purposive RDD sample of the current study and the nationally representative sample of Reading First schools and personnel surveyed by the Implementation Evaluation.

The Reading Coach Survey targeted school level individuals designated as reading coaches, reading specialists, literacy coordinators or Title I/resource teachers. It included items on the coach's background and experience, core and supplemental reading materials, professional development offered to grades 1–3 teachers, specific coaching activities, characteristics of reading instruction in the school, changes that have taken place in reading instruction, and areas needing improvement. See Appendix C, Part 7, Exhibit C.17 for a copy of the Reading Coach Survey.

The Teacher Survey addressed student characteristics, reading instruction (e.g., materials, content, time allocation), assessment, interventions for struggling readers, participation in reading-related professional development, and collaboration and support from other teachers and staff. The surveys provide self-reported information on the instructional emphases across grades. See Appendix C, Part 7, Exhibit C.18 for a copy of the Grade 2-3 Teacher Survey.

Administration Procedures and Response Rates

Surveys were mailed in March 2007 to building level study liaisons who then distributed sealed envelopes to an average of nine classroom teachers (three per grade level), the school's reading coach, and the building principal. Respondents were asked to return the completed surveys within two to three weeks. Follow-up was conducted to encourage potential respondents to complete and return the surveys.

---

[1] The Reading First Implementation Evaluation, commissioned by the Policy and Program Studies Service at the U.S. Department of Education, collected survey data from principals, teachers and reading coaches in nationally representative samples of RF schools and non-RF Title I schoolwide project (SWP) schools in the 2004–05 and 2006–07 school years.

Exhibit C.1 provides the response rate and sample size for each survey. In 2007, 227 reading coaches and 1,792 teachers completed a survey. The effective Reading Coach Survey response rate was 99 percent for RF schools and 89 percent for non-RF schools. The effective Teacher Survey response rate was 87 percent for RF schools and 83 percent for non-RF schools.

Exhibit C.1: Survey Data Collection: School, Reading Coach, and Teacher Sample Information

Reading Coach Survey

RFIS Schools

RF: 125   Non-RF: 123   Total: 248

Reading Coaches

	Spring 2005	*Spring 2007*
RF:	124	124
Non-RF:	109	118
Total:	233	242

Reading Coaches Who Did Not Meet Criteria[1]

	Spring 2005	*Spring 2007*
RF:	1	1
Non-RF:	14	5
Total:	15	6

Reading Coaches Who Responded to Survey

	Spring 2005	*Spring 2007*
RF:	118 (95%)	123 (99%)
Non-RF:	79 (72%)	105 (89%)
Total:	197 (85%)	227 (94%)

Analytic Sample[2]

	Spring 2005	*Spring 2007*
RF:	118 (100%)	123 (100%)
Non-RF:	79 (100%)	105 (100%)
Total:	197 (100%)	228 (100%)

[1] Reading coach respondents who did not meet criteria included individuals who indicated that they did not serve any school(s) as a reading coach who provided ongoing training and support to school staff in delivering effective reading instruction.

[2] All completed surveys were used in the analytic sample. Information on item-level response rates is presented on tables where applicable, and Appendix B, Part 5 describes the overall approach to handling missing survey data.

SOURCE: Reading First Impact Study Reading Coach Surveys, spring 2005 and spring 2007.

Exhibit C.1: Survey Data Collection: School, Reading Coach, and Teacher Sample Information (continued)

Teacher Survey

RFIS Schools

RF: 125   Non-RF: 123   Total: 248

Sampled Teachers

	Spring 2005	Spring 2007
RF:	1,523	1,511
Non-RF:	1,489	1,439
Total:	3,012	1,950

Teachers Who Did Not Meet Criteria[1]

	Spring 2005	Spring 2007
RF:	16	123
Non-RF:	21	81
Total:	37	204

Teachers Who Met Criteria

	Spring 2005	Spring 2007
RF:	1,507	1,388
Non-RF:	1,468	1,358
Total:	2,975	2,746

Eligible Teachers Not Selected

	Spring 2005	Spring 2007
RF:	0	322
Non-RF:	0	317
Total:	0	639

Eligible Teachers Selected

	Spring 2005	Spring 2007
RF:	1,507	1,066
Non-RF:	1,468	1,041
Total:	2,975	2,107

Teachers Who Responded to Survey[2]

	Spring 2005	Spring 2007
RF:	1,076 (71%)	927 (87%)
Non-RF:	961 (65%)	865 (83%)
Total:	2,037 (69%)	1,792 (85%)

Analytic Sample[3]

	Spring 2005	Spring 2007
RF:	1,344 (100%)	1,400 (100%)
Non-RF:	1,315 (100%)	1,342 (100%)
Total:	2,659 (100%)	2,742 (100%)

NOTES:

[1] Respondents who did not meet criteria or were not selected included student teachers, substitute teachers, or teachers whose classrooms were not observed (for grades 1 and 2) or tested (grades 1, 2, and 3).

[2] A total of 23 teachers (15 in 2004-05, 8 in 2006-07) returned surveys but were dropped because they indicated that they did not teach reading or grades 1, 2, or 3.

[3] All completed surveys were used in the analytic sample. Information on item-level response rates is presented on tables where applicable, and Appendix B, Part 5 describes the overall approach to handling missing survey data.

SOURCE: Reading First Impact Study Teacher Surveys, spring 2005 and spring 2007.

Composition, Scale, Internal Consistency and Scientifically Based Research Support

Exhibit C.2 reports the composition, metric, specifications, and internal consistency of the survey outcomes. Exhibit C.3 includes information on the Reading First legislation and guidance that supports the survey outcomes.

Part 2: Classroom Instruction: The Instructional Practice in Reading Inventory (IPRI)

Background

To measure the impact of Reading First on classroom instruction, the RFIS team conducted classroom observations in both Reading First and non-Reading First (non-RF) classrooms. The primary instrument used to evaluate instruction was the Instructional Practice in Reading Inventory (IPRI). The RFIS Team was unable to identify an existing observational instrument that fulfilled all of the study requirements; consequently, the RFIS Team developed the IPRI specifically for the RFIS. The IPRI is designed to measure first- and second-grade teachers' use of instructional behaviors informed by scientifically based reading research (SBRR), as described 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). In particular, the IPRI focuses on instruction in the five dimensions of reading instruction emphasized by SBRR (phonemic awareness, decoding/phonics, fluency, vocabulary, and comprehension). Exhibit C.4 gives specific examples of instructional activities associated with each of the five dimensions.

The development of the IPRI relied on several sources, including (1) research on the components of effective elementary grade reading instruction (e.g., Kamil, 2004; National Institute of Child Health and Human Development, 2000; Snow, Burns and Griffin, 1998; Stahl, 2004); (2) reviews of existing instruments (among the instruments reviewed were the following: *The Instructional Content Emphasis (ICE)* [Edmonds and Briggs, 2003]; *Foorman and Schatschneider direct observation system and instruments from the Center for Academic and Reading Skills (CARS)* [Foorman and Schatschneider, 2003]; *English Language Learner Classroom Observation Instrument (ELLCOI)* [Haager et al., 2003]; *Teachers' Instructional Practice (TIP)* [Carlisle and Scott, 2003]; *Utah's Profile of Scientifically based Reading Research* [Dole, et al., 2001]; *The Classroom Observation Record* [Abt Associates and RMC Research, 2002]; and *Observation Measure of Language and Literacy Instruction (OMLIT)*, developed by Abt Associates as part of the Even Start Classroom Literacy Interventions and Outcomes (CLIO) Study [Goodson et al., 2004]); and (3) research on the development of classroom observation instruments (Vaughn and Briggs, 2003).[2]

---

[2] For a comprehensive description of the development of the IPRI, see Dwyer et al., 2007.

Exhibit C.2: Composition, Metric, Specifications, and Internal Consistency of Survey Outcomes

Domain Survey Outcome …..Individual Items Comprising the Outcome (as applicable)	Survey, Item(s)	Metric	Outcome Specifications[1]	Internal Consistency[2]
Professional Development (PD) in SBRI				
1. Amount of PD in reading received by teachers ..…Attended short, stand-alone training or workshop in reading (half-day or less) …..Attendance longer institute or workshop in reading (more than half-day) …..Attended a conference about reading (might include multiple short offerings)	Teacher, C1: a, b, d	Hours	Sum of hours across 3 items	0.22
2. Teacher receipt of PD in the five essential components of reading instruction …..In phonemic awareness …..In phonics …..In vocabulary …..In fluency …..In comprehension	Teacher, C4: a-s	0-5 scale	Each component (e.g., fluency) was scored dichotomously (1=teacher received PD in at least one of the topics listed, 0= teacher did not receive PD in any of the topics listed). Sum of 5 dichotomously scored components (1=addressed, 0=not addressed).	0.86
3. Teacher receipt of coaching	Teacher, C2: a	Proportion	Dichotomous variable (1=teacher received assistance from reading coach, 0=teacher did not receive assistance from reading coach or not available)	N/A
4. Amount of time dedicated to serving as K-3 reading coach	Reading Coach, B3	Percent	N/A	N/A
Amount of Reading Instruction				
5. Minutes spent on reading instruction per day …..Last week, approximately how many minutes per day did you devote to reading instruction? Reported separately for each of the five weekdays.	Teacher, B1	Minutes	Total number of minutes of reading instruction for last week divided by number of days with non-missing values.	0.99
Supports for Struggling Readers				
6. Availability of differentiated instructional materials for struggling readers[3] …..Use separate program materials in interventions …..Use core reading program with supplemental materials …..Use core reading program only …..Use reading materials written in ELLs' home language …..Use alternative materials designed for ELLs	Reading Coach/ Principal, E1: a-e	Proportion	Dichotomous variable (1=E1: a, b, d, or e are available for struggling readers, 0=only the core reading program is available for struggling readers, E1: c)	N/A

Exhibit C.2: Composition, Metric, Specifications, and Internal Consistency of Survey Outcomes (continued)

Domain Survey Outcome …..Individual Items Comprising the Outcome (as applicable)	Survey, Item(s)	Metric	Outcome Specifications[1]	Internal Consistency[2]
7. Provision of extra classroom practice for struggling readers (over the past month) …..Extra practice in the classroom with phonemic awareness …..Extra practice in the classroom with phonics …..Extra practice in the classroom with fluency …..Extra practice in the classroom with comprehension	Teacher, B8: b-e	0-4 scale	Sum of 4 dichotomously scored items (1=received, 0=did not receive)	0.77
Use of Assessments				
8. Use of assessments to inform classroom practice …..Use test results to organize instructional groups …..Use tests to determine progress on skills …..Use diagnostic tests to identify students who need reading intervention services	Teacher, B5: u, w, y	0-3 scale	Sum of 3 dichotomously scored items (1=central, 0=small or not part of reading instruction)	0.60

NOTES:

[1] Missing data rates ranged from 0.1 to 3.3 percent for teacher survey outcomes (RF: 0.1 to 1.0 percent; non-RF: 0.0 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).

[2] Internal consistency was calculated using Cronbach's raw alpha for survey outcomes other than single dichotomous outcome variables.

[3] This survey item was asked of both reading coaches and principals. In cases where reading coach survey data were not available, the study used principal responses for those schools (n=13).

EXHIBIT READS: The outcome variable "amount of PD in reading received by teachers" consisted of three individual items from the RFIS Teacher Survey, item C1. The sum of these three items represents the total the number of hours of professional development in reading attended by teachers in the form of short trainings/workshops, longer institutes/ workshops, and conferences. The internal consistency reliability was 0.22.

Sources: Reading First Impact Study Reading Coach, Principal, and Teacher Surveys.

Exhibit C.3: Reading First Legislative Support and Guidance for Survey Outcomes

Domain Survey Outcome …..Individual Items Comprising the Outcome (as applicable)	Survey, Item(s)	Reading First Legislation[1] or Guidance[2]
Professional Development (PD) in SBRI		
1. Amount of PD in reading received by teachers ..…Attended short, stand-alone training or workshop in reading (half-day or less) …..Attendance longer institute or workshop in reading (more than half-day) …..Attended a conference about reading (might include multiple short offerings)	Teacher, C1: a, b, d	Legislation, Section 1202(c)(7)(A)(iv) Guidance, p. 7, C-3
2. Teacher receipt of PD in the five essential components of reading instruction …..In phonemic awareness …..In phonics …..In vocabulary …..In fluency …..In comprehension	Teacher, C4: a-s	Legislation, Section 1202(c)(7)(A)(iv)(I) Guidance, p. 7, C-3
3. Teacher receipt of coaching	Teacher, C2: a	Legislation, Section 1202(c)(7)(A)(iv)(III) Guidance, p. 7, C-3
4. Amount of time dedicated to serving as K-3 reading coach	Reading Coach, B3	Legislation, Section 1202(c)(7)(A)(iv)(III) Guidance, p. 7, C-1, C-3
Amount of Reading Instruction		
5. Minutes spent on reading instruction per day	Teacher, B1	Guidance, p. 7, C-1
Supports for Struggling Readers		
6. Availability of differentiated instructional materials for struggling readers …..Use separate program materials in interventions …..Use core reading program with supplemental materials …..Use core reading program only …..Use reading materials written in ELLs' home language …..Use alternative materials designed for ELLs	Reading Coach/ Principal, E1: a-e	Guidance, p. 7, C-2
7. Provision of extra classroom practice for struggling readers …..Extra practice in the classroom with phonemic awareness …..Extra practice in the classroom with phonics …..Extra practice in the classroom with fluency …..Extra practice in the classroom with comprehension	Teacher, B8: b-e	Legislation, Section 1202(c)(7)(A)(ii)(II)(ee)

Exhibit C.3: Reading First Legislative Support and Guidance for Survey Outcomes (continued)

Domain Survey Outcome …..Individual Items Comprising the Outcome (as applicable)	Survey, Item(s)	Reading First Legislation[1] or Guidance[2]
Use of Assessments		
8. Use of assessments to inform classroom practice …..Reading assessments are used to screen students for reading difficulties …..Diagnostic assessments are used to identify strengths and weaknesses of struggling readers …..Reading assessments are used to monitor student progress	Teacher, B5: u, w, y	Legislation, Section 1202(c)(7)(A)(i) Guidance, p. 7, C-1, C-2, C-4

NOTES:

[1] The legislation for Reading First is contained in the No Child Left Behind Act of 2001, ESEA, 2001, Title 1, Part B, Subpart 1.

[2] Guidance for the Reading First Program is provided by the U.S. Department of Education (2002).

EXHIBIT READS: The outcome variable "amount of PD in reading received by teachers", which consists of three individual items from the RFIS Teacher Survey (Question C1), was supported by both the Reading First legislation [No Child Left Behind Act of 2001, ESEA, 2001, Title 1, Part B, Subpart 1, Section 1202(c)(7)(A)(iv)] and the Guidance for the Reading First program (U.S. Department of Education, 2002, p. 7, C-3).

Sources: Reading First Impact Study Reading Coach, Principal, and Teacher Surveys.

Exhibit C.4: Examples of Instruction in the Five Dimensions of Reading Instruction

Phonemic awareness	The teacher is working with a group of four students. The teacher says, "Listen to me. The word is *hat*. If I take away the /h/ sound at the beginning, I have the word *at*. Then if I add a /b/ sound to the beginning I get *bat*. Now you try. The word is *sat*. If we take away the /s/ sound what word do we have?" [students respond orally]. "That's right, *at*. Now add a /k/ sound to the beginning. What word? That's right, *cat*."
	The teacher is working with a pair of students. He asks students to identify the final sound in each of a list of 10 words. The students respond orally to each prompt from the teacher: "*Crack*. What's the last sound in *crack*? [students respond orally]. Good. Ok: *Take*. What's the last sound? [students respond orally]. Ok, next: *kite*. What's the last sound? [students respond orally]. How about *flight*? [students respond orally]. That's right, /t/, /t/ is the last sound in *flight*. "
Phonics	A group of 16 students has assembled in front of the classroom blackboard. The teacher writes the letters *oi* on the board and says, "Ok, now today we're going to be learning about words that have o, i in them. When you see these vowels together, they make the /*oy*/ sound. Here's an example." The teacher writes a sentence on the board: *I want Roger to join my club.* She underlines the letters oi in the word join. "This word is *join*. 'I want Roger to *join* my club. See that *oi*? What sound does *oi* make?" [students respond, some of them incorrectly]. "Ok, listen carefully. Not /*eye*/… no, *oi* makes the /*oy*/ sound. Everyone try that: /*oy*/." [students in unison say /oy/]. "Ok, good, now what's this word [she points to *join*]?" The students pronounce *join* correctly. "Excellent, ok, let's try another one." She writes the word *coin* on the board. "Boys and girls, look at that *oi* in the word. Sound out this word for me."
	Six students are seated with a teacher. Each student has a set of individual magnetic letters and a metal tray. The teacher is asking students to form words that she dictates orally: "Ok, listen to the word, think about the sounds and what letters go with those sounds. Remember that we've been working with the /ō/ sound and its spellings. We know that one way to spell that is with *o, a*. Try to make the word using your letters. The first word is *goat*. Use your letters to make the word *goat*." Students assemble their letters and the teacher checks each student's work. "Good. Everyone used *o, a* to spell *goat*. Ok, let's try another word: *float*." Students form the word with their letters. "Ok, good! You're doing very well. Now, we also know another way to spell some words with the long /ō/ sound. Remember the silent e rule? It makes the vowel say its name. So, to spell the word *tote*, Arthur, tell me how we'd write *tote*?"
Vocabulary	The teacher gives a definition for the word **swift** and uses it in a sentence: "Swiftly? Something that is swift is moving very fast, rapidly. So, remember when we learned about how fast cheetahs can run over land? Well, we might say, 'the cheetah ran *swiftly* across the ground, quickly catching up to the tiger.'"
	As they are reading a story in class, students come across the word **debating**, and the teacher discovers that they do not know what it means. The teacher defines **debating** by contrasting it with more familiar words (**chatting** and **talking**). The teacher says, "When two people are **debating** something, it means that they are talking about the reasons to do something and the reasons not to do something—so in our story, John and Sara are **debating** whether or not to go on a picnic. On the one hand, the weather is nice, but on the other hand they are thinking there may be a lot of ants. So they're **debating** what to do. **Chatting** is different than **debating**. When you're **chatting** with someone, you're usually not trying to decide something, you're just talking about things that aren't too serious. You chat more to enjoy the talking, not really to decide something together."

Exhibit C.4: Examples of Instruction in the Five Dimensions of Reading Instruction (continued)

Fluency	Roberto is reading orally from a passage about parrots and their habitat. When he reaches the end of the second paragraph, the teacher asks Roberto to read that same passage aloud again. When Roberto finishes, the teacher asks him to read the passage out loud a third time.
	The teacher assigns four students to pairs and distributes a page-long excerpt from a story they have been reading in class that week. Each pair of students also has a one minute timer. "Ok, now you each have a partner, and I want you to time your partner reading this passage out loud. Readers, you try to read as far as you can in one-minute. Timers, you keep track of the time and tell your partner to stop reading when time runs out. Then circle the last word the reader got to in the passage."
Comprehension	A teacher pauses in the middle of a story about Shackleton's Antarctic Voyage to ask students to reflect on what they have just read and draw some inferences about how one character might be feeling. "What do you think the captain is feeling? Let's see. The story doesn't tell us exactly, but the story says the ship is starting to break apart. I'd certainly be very worried for myself and my crew if my ship were breaking apart! I bet the captain is really worried. Let's see… the story also says the captain 'furrowed his brow.' That means he made his forehead wrinkle or sort of frown. Some people do that when they're worried. That could be a sign that the captain is worried. He certainly has reason to be worried."
	The teacher introduces a comprehension strategy. "One thing you should always do when you read is constantly ask yourself questions about the story. Asking yourself questions is a strategy to help make sure you understand what you just read. Asking questions also helps you think about what might happen next. We're going to practice using this strategy. At the end of every paragraph today, we're going to come up with some questions and write them up here on the board. Some questions we'll be able to answer right away. But we might have other questions, too, and we'll need to read more of the story before we can find out how to answer those questions."

Overview of the IPRI

The IPRI observation instrument is a booklet containing a series of individual IPRI forms, each of which corresponds to a three-minute observation interval.[3] Observation data for a given reading block are collected via sequentially-ordered IPRI forms that span the entire observation period (e.g., a 60-minute observation would be recorded on 20 sequential forms, one for each successive three-minute interval). During each three-minute interval, observers record any of the teacher's instructional behaviors listed on the IPRI that occur during that interval. At the end of each three-minute interval (signaled by a pre-programmed vibrating wristwatch), observers turn to a new IPRI form and begin another three-minute interval, again recording the presence of targeted behaviors.

Within a given three-minute interval, a particular behavior is coded only once, regardless of how often that behavior occurs within an interval. Recurrences of that same behavior are coded in each subsequent interval. If behavior x occurs in interval n, the observer circles the code for behavior x once during interval n. If behavior x occurs in the next interval, $n+1$, the observer circles the code for behavior x during interval $n+1$.

---

[3] See Exhibit C.19 for a copy of the IPRI instrument.

Structure of the IPRI Instrument

Each IPRI form has four distinct parts: Part A, Part B, Part C, and Part D. Part A is divided into five color-coded sections that correspond to the five dimensions of reading instruction: phonemic awareness, decoding/phonics, fluency, vocabulary, and comprehension, respectively. Within each of these five sections are microcodes, specifically tailored to each of the five dimensions, which denote the following areas of interest:

- the size of the student grouping to which instruction is delivered;

- the use of any instructional support materials (e.g, manipulatives, pictures);

- the teacher's use of explicit instruction;

- the teacher's provision of practice opportunities for students; and

- the teacher's delivery of any corrective feedback or expansion of student responses.

For example, within the phonemic awareness row, the IPRI microcodes for grouping are "whole class, large group, small group, pair, or individual"; for the use of various types of instructional supports, "teacher manipulative or kinesthetic, student manipulatives, kinesthetics"; and for corrective feedback, "teacher pinpoints what student(s) did incorrectly with sound(s) and gives correct response with or without students." For the use of explicit instruction and the provision of practice opportunities for students, these areas of interest are often denoted by the combination of two or more microcodes. So, for example, if a teacher "demonstrates or models oral blending or segmenting with phonemes" *in conjunction with* "gives student(s) chance to practice oral blending or segmenting with phonemes," it would be counted as explicit instruction.

Part B of the IPRI contains codes to capture instruction or other activity outside the five dimensions, including:

- Oral reading by students;[4]
- Oral reading by teacher alone (without student accompaniment);
- Silent reading;
- Spelling;
- Written expression;
- Other language arts;
- Assessment;
- Non-literacy instruction;
- Non-instruction;
- Academic management;
- Transitions between activities;
- Interruptions to instruction for the purpose of managing student behavior.

---

[4] Oral reading under Part B is marked when the teacher has not clearly indicated the instructional purpose of the oral reading. If, however, oral reading is used to advance instruction in one of the five targeted dimensions of reading instruction (e.g., comprehension), then the oral reading is coded within the corresponding row in Part A of the IPRI.

Part C records teachers' instructional errors that are not subsequently self-corrected. Part D records whether the teacher worked with a different small group of students than in any previous part of the observation.[5]

Training and Inter-rater Reliability of Classroom Observers

Prior to each wave of data collection, field staff based in each of the RFIS sites attended a centralized, multi-day training on the IPRI and associated data collection protocols. The training curriculum included extensive practice coding a series of videotaped clips of real-time and unscripted classroom instruction that were filmed in RF and non-RF classrooms. The film clips were created specifically for the RFIS, and were edited to illustrate the codes included on the IPRI. Candidate observers conducted a live observation in a first or second grade classroom during the training session and received ongoing feedback, multiple opportunities for review, tutoring and other support throughout the training.[6]

One component of this training was that observers were required to pass two of three formal inter-rater reliability tests; each videotape used for reliability purposes was approximately 30 minutes in length. To calculate observers' percent agreement with the master coding of each reliability tape, the RFIS Team used a procedure that reduces inflation in inter-rater reliability estimates due to chance agreement (see Kelly, 1977, cited in Suen and Ary, 1989). The inflation due to chance agreement is especially severe when some events (or codes) occur infrequently, as is the case with the IPRI.[7] As a result, observers were credited only for codes that occurred at least once in the reliability tape. In sum, if a behavior occurred at all during a 30-minute tape, observers were credited (or penalized) for correctly coding instances of the behavior and for correctly abstaining from coding behaviors that did not occur. Observers were not credited for abstaining from, nor penalized for, marking behaviors that never occurred throughout the entire reliability tape.

For each potential observer, percent agreement with the master codes was calculated for each code individually; then agreement was aggregated across codes within the five sections in Part A and across codes within Part B. Finally, an aggregate overall percentage agreement across the five sections in Part A and codes within Part B was calculated. A report summarizing all of these measures of agreement (by individual code, by dimension, and overall) was prepared for each potential observer so that s/he (and the study team) could diagnose which codes had proven particularly troubling. Overall percent agreement was used to judge whether or not each observer had met the criterion for employment on the study. Only observers who successfully coded two of three videotaped reliability tests were hired. The mean overall percent agreement for observers was 88 percent (n=155 observers) in spring 2005 (for spring 2005 data collection). The mean overall percent agreement for observers was 90 percent (n=154 observers) in fall 2005 (for fall 2005 and spring 2006 data collection). The mean overall percent agreement for observers was 90 percent (n=130 observers) in fall 2006 (for fall 2006 and spring 2007 data collection).

---

[5] Minor changes were made to the IPRI after the spring 2005 data collection and prior to the fall 2005 wave of data collection; these changes included elaborating upon some micro-behaviors within each of the five dimensions.

[6] For a detailed description of the classroom observer training, see Dixon et al. (2007).

[7] During each observation interval, an IPRI form contains 142 possible codes; typically, only a small subset of the behaviors occur during a given interval. Thus, most of the possible codes are infrequent within a single interval. Including all 142 codes per interval in the calculation of percent agreement severely inflates inter-rater reliability.

Data Collection

Observations were conducted in first- and second-grade classrooms for two consecutive days during each classroom's designated reading block. During the 2004-05 school year, the RFIS conducted two days of classroom observation in spring 2005. In the following study year (2005-06), a second round of observations was added, so that observers conducted observations for two consecutive days in the fall, and then again for two consecutive days in the spring. Again in 2006-07, observers conducted observations for two consecutive days in the fall, and then again for two consecutive days in the spring. The increased number of observations reflects a decision by the National Center for Education Evaluation/Institute of Education Sciences at the Department of Education to collect more data, both in terms of the number of observations and in terms of when during the year data could be collected.

Observation scheduling was arranged by RFIS field supervisors via communication with each participating school's study liaison.[8] Observers coded during the entire scheduled observation period, even when teachers appeared to be offering non-reading-related instruction. In those instances when reading instruction appeared to continue beyond the scheduled reading block, observers observed for up to an additional 30 minutes. Throughout observations, IPRI observers followed the actions and behaviors of classroom teachers. In classrooms with more than one adult present, observers determined beforehand who was the official teacher of record and which adult would be delivering that day's reading instruction. The individuals responsible for delivering instruction were then followed for the observations whether or not they were the official teacher of record. Observations were rescheduled when the classroom teachers were absent or ill, although long-term substitutes replacing a teacher on an extended leave of absence (e.g., maternity, disability) were observed.

The 248 schools in the RFIS study sample allowed for observations in 2,091 classrooms in 2004-2005, 3,997 classrooms in 2005-2006, and 3,985 classrooms in 2006-2007. Of these, 1,917 classrooms met eligibility requirements for classroom observations in 2004-2005, 3,649 in 2005-2006, and 3,676 in 2006-2007. Classrooms were considered eligible to be in the study sample if they were not special education or English as a Second Language classes, if more than 75 percent of the students were in the target grades, and if the class was taught by the regular teacher or a long-term substitute.

Of the eligible classrooms, the RFIS selected a final observation sample of 1,639 classrooms in 2004-2005, 2,770 in 2005-2006, and 2,814 in 2006-2007. Classrooms were sampled within schools, if, within each site as a whole, the number of classrooms exceeded an average of three classrooms per grade. Each classroom in the sample was expected to be observed two times during each of the three waves of data collection. The RFIS completed 96 percent of the expected classroom observations in 2004-2005, and 100 percent in both 2005-2006 and 2006-2007. A flow chart of information on the RFIS IPRI sample and response rates is presented in Exhibit C.5.

---

8 In schools that did not have a designated "reading block," the RFIS Team asked the school's study liaison when observers would be able to see typical reading, literacy, and/or language arts instruction in classrooms. In cases where reading instruction was delivered in two discrete blocks interrupted by other instruction or activities (e.g., lunch, recess, math instruction), field staff observed both blocks.

Exhibit C.5: IPRI Data Collection: School, Classroom, and Observation Sample Information

RFIS Schools

RF: 125   Non-RF: 23   Total: 248

All Grade 1 & 2 Classrooms

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	1,416	1,364	1,378	1,384	1,390
Non-RF:	1,401	1,343	1,355	1,342	1,358
Total:	2,817	2,707	2,733	2,726	2,748

Sampled Grade 1 & 2 Classrooms[1]

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	1,056	1,010	1,022	1,033	1,035
Non-RF:	1,035	979	986	953	964
Total:	2,091	1,989	2,008	1,986	1,999

Classrooms That Did Not Meet Criteria[2]

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	95	90	100	97	95
Non-RF:	79	77	81	58	59
Total:	174	167	181	155	154

Classrooms That Met Criteria

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	961	920	922	936	940
Non-RF:	956	902	905	895	905
Total:	1,917	1,822	1,827	1,831	1,845

Excluded Classrooms[3]

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	145	224	226	221	226
Non-RF:	133	214	215	206	209
Total:	278	438	441	427	435

Final Observation Sample

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	816	696	696	715	714
Non-RF:	823	688	690	689	696
Total:	1,639	1,384	1,386	1,404	1,410

Classrooms Observed

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	788	692	692	711	709
Non-RF:	791	686	688	688	692
Total:	1,579	1,378	1,380	1,399	1,401

Classrooms Completed[4]

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	1,575 (97%)	1,384 (99%)	1,384 (99%)	1,422 (99%)	1,418 (99%)
Non-RF:	1,580 (96%)	1,372(100%)	1,375(100%)	1,375(100%)	1,384 (99%)
Total:	3,155 (96%)	2,756(100%)	2,759(100%)	2,797(100%)	2,802 (99%)

Analytic Sample[5,6]

	Spring 2005	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	1,575(100%)	1,384(100%)	1,384(100%)	1,422(100%)	1,418(100%)
Non-RF:	1,580(100%)	1,372(100%)	1,375(100%)	1,375(100%)	1,384(100%)
Total:	3,155(100%)	2,756(100%)	2,759(100%)	2,797(100%)	2,802(100%)

Notes:

[1] The study conducted observations in all classrooms in schools if across the site as a whole, the average number of classrooms per grade per school was three or less. If for that site as a whole, the average number of classrooms per grade per school exceeded three, the study sampled classrooms for observations within schools.

[2] At the beginning of each data collection wave, the study team contacted schools to obtain classroom rosters and indications of which classrooms were designated as regular classrooms, self-contained special education classrooms, or ESL classrooms. Those classrooms identified as special education or ESL were considered ineligible for observation.

[3] Once the study team began to schedule and conduct observations, classrooms were excluded from the sample if the information that had been provided earlier about special education or ESL status was inaccurate, if the class was taught by someone other than the regular teacher or a long-term substitute, or if fewer than 75 percent of the students were in the target grade.

[4] Each classroom was observed twice during each wave of data collection. Response rates are calculated by dividing the observations completed by two times the number of eligible classrooms selected into the sample.

[5] All IPRI observations were included in the analytic sample. Percentages are calculated by dividing the analytic sample numbers by the number of observations completed.

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

During each data collection wave (spring 2005, fall 2005, spring 2006, fall 2006, and spring 2007), IPRI experts (from the training staff) served as quality control monitors for questions that arose in the field. Quality control monitors visited each site and accompanied a random selection of observers into scheduled classroom observations. The monitors reviewed the observation coding with the observers, addressing coding discrepancies and questions.[9] Throughout the data collection period, observers could direct questions to the monitors and to other RFIS staff. Questions and answers were aggregated and disseminated to all observers via an RFIS observer website and regular mailings.

Creation of Analytic Variables

To test whether or not instruction in RF classrooms differed from that in non-RF classrooms, the study team created eight measures of classroom instruction from the IPRI data. The number of measures was deliberately limited so that the analysis would be parsimonious, and would thereby restrict the number of statistical tests required. The measures were:

- Time spent in instruction in each of the five targeted dimensions of reading instruction separately:
  - phonemic awareness;
  - phonics/decoding;
  - vocabulary;
  - fluency;
  - comprehension;

- Time spent in instruction in the five dimensions combined;

- Proportion of instruction in the five dimensions that was highly explicit—that includes teacher modeling, clear explanations, and the use of examples;

- Proportion of instruction in the five dimensions that provided students with high quality practice opportunities—that includes, for example, teachers giving students the opportunity to practice word learning strategies (e.g., context, word structure, and meanings).

Before describing these measures in more detail, we first describe the transformation of raw interval data into more meaningful metrics.

Transformation of IPRI Observation Intervals Into Minutes
The IPRI contains multiple successive three-minute intervals, each of which could potentially record a large number of instructional behaviors, if the behaviors had indeed been observed. Each behavior on the IPRI is deemed to have occurred or not occurred in each observed interval (e.g., behavior was present [checked or coded] or not [unchecked]). Across the entire set of intervals comprising a classroom observation, the IPRI yields raw data in terms of the number (or proportion) of observed intervals in which a given behavior was observed. The raw data do not directly measure the duration of particular instructional activities or behaviors. In order to describe classroom instruction with a more interpretable metric, raw intervals were transformed into minutes of instruction via the process described below.

---

[9] Study protocols required observers to leave as is any codes marked during the observation. This procedure allowed the RFIS study team to collect a sample of paired observations for use in determining field-based reliability.

For each and every interval, observers recorded instruction in one of the five dimensions (hereafter referred to as "dimensions")[10] or in other activity/instruction not in one of the five dimensions (hereafter referred to as "non-dimension activities"). These latter activities are included in "Part B" described above. Consequently, every observation interval contains *at least* one of the following codes that categorizes the types of instruction the teacher provided during that interval:

- Phonemic awareness;
- Phonics/decoding;
- Vocabulary;
- Fluency;
- Comprehension;
- Oral reading by children;[11]
- Oral reading by teacher;
- Silent reading;
- Spelling;
- Written expression;
- Other language arts;
- Assessment;
- Non-literacy instruction;
- Non-instruction;
- Academic management; and/or
- Transitions between instructional activities.

The allocation of time within the three-minute intervals occurred at the broader level—that is, at the level of dimension and non-dimension activities. When only one dimension or non-dimension activity was observed in an interval, the conversion process was straightforward—all three minutes of the interval were assigned to the dimension or non-dimension activity observed.

When *two* activities were recorded in an interval, however, the process of converting intervals into minutes was less straightforward. From the raw data, there was no direct way to determine the proportion of the three minute interval that the teacher had devoted to each of the two recorded activities. Therefore, the study team developed an estimation process to allocate minutes of that interval to each of the two activities. The RFIS collected supplemental data on the actual duration of instructional activities recorded on the IPRI, and used those supplemental data to inform mathematical simulations of the outcomes of different estimation procedures.

---

[10] For purposes of calculating minutes of instruction in a particular dimension, the micro-level codes corresponding to aspects of instruction *within* each of the five dimensions were collapsed. For example, a teacher who had exhibited two different "phonics" codes within an interval was designated as having delivered phonics instruction within that interval.

[11] Note that the IPRI distinguishes oral reading for its own sake from oral reading in service to a larger instructional purpose. For example, oral reading that occurred to advance a lesson in comprehension was classified as being part of the overarching comprehension instruction and was not counted as oral reading for purpose of analysis. In contrast, oral reading that occurred outside the context of one of the five dimensions of reading instruction was classified for analytic purposes as Oral Reading.

Dividing the minutes of the interval equally. Initially, the RFIS Team considered allocating one-half of the three-minutes of an interval to each of the two activities observed. Under this procedure, if comprehension and phonics were observed in the same interval, for instance, then each would be assigned 1.5 minutes of the three-minute interval. Although this approach provides a good estimate of the true number of minutes spent in the two activities for intervals in which the two observed activities were of similar duration, for intervals in which activities were of unequal duration, however (e.g., one activity was 2.6 minutes and the other .4 minutes), this approach underestimates the amount of time in the longer activity and overestimates the amount of time spent in the shorter one.

Dividing the minutes of the interval according to their relative frequency of occurrence. The study team also explored an estimation method that allocates time to each of two activities within a given interval in direct proportion to the relative frequency with which the two activities occurred, on average, within the school in which the observation had been conducted. If, on average, comprehension was present in 30 percent of the intervals collected across all observations within a school, whereas fluency instruction was present in 10 percent of all intervals collected in the school, then comprehension was three times as likely to occur as fluency instruction. Then for each interval in which comprehension and fluency were the two activities recorded, comprehension would receive 75 percent of the three minutes (or 2.25 minutes) and fluency would receive 25 percent of the three minutes (one-third the amount of time as comprehension, or .75 minutes).

The RFIS Team used supplemental data on the true duration of instructional activities to simulate the precision of this estimate. The simulations suggested that the proportionally-weighted approach provided a close estimate of the true minutes spent in activities for intervals in which two activities were of unequal duration, but, conversely, it produced biased estimates of the true minutes spent in activities for intervals in which the two activities observed were of similar duration. Thus, the strengths and drawbacks of this approach were mirror opposites of those in the first approach (i.e., dividing the minutes equally among the two activities in an interval).

The RFIS Team decided that an average of the two estimations would minimize the biases introduced by using either of the two transformation approaches in isolation.

Dividing the minutes of the interval by taking the average of the equally and proportionally weighted approaches. For each interval with two instructional activities recorded, a three-step estimation process was used:

1. The minutes were allocated *equally* between the two activities (1.5 minutes to each).

2. The minutes were allocated *according to their relative frequency* of occurrence across all observations within school.

3. The *average* of the two estimates produced was calculated for each of the two activities.

Using the example cited above (an interval with only comprehension and fluency instruction, comprehension would be allocated 1.88 minutes, or the mean of the equally weighted and proportionally weighted approach [1.5 and 2.25, respectively]). Fluency would be allocated 1.12 minutes (the mean of 1.5 and .75 minutes).

Three or more activities occurring in the same interval. When three or more instructional activities were observed in a single interval, the three minutes of the interval were divided equally among the activities. This distribution strategy was followed rather than the estimation process used for two-activity intervals because the number of minutes assigned to any given activity type would be limited to one minute or less. Thus, the amount of bias introduced by using this estimation approach was likely to be small.

Analytic Variables
The study team constructed six variables based on the amount of time devoted to instruction in the five dimensions of reading instruction: one variable for the amount of time spent in each of the five dimensions separately, plus a sixth variable for the total amount of time spent in the five dimensions combined.

Also of interest were the degree to which instruction in RF and non-RF schools was highly explicit, and the degree to which instruction offered students meaningful opportunities to practice developing reading skills. To examine these outcomes, two additional variables were constructed: the percentage of instruction in the five dimensions in which at least one instance of highly explicit instruction occurred; and the percentage of instruction in the five dimensions in which at least one instance of high quality student practice occurred. These two variables are defined below.

Percentage of intervals of instruction in the five dimensions that included at least one instance of highly explicit instruction. "Highly explicit instruction" is defined differently in each dimension of reading instruction, based on research published in the National Research council report (Snow, Burns, and Griffin, 1998) as well as more recent research (e.g., Graves, Gerston and Haager, 2004; Gunn et al., 2002 for specific examples of highly explicit instruction in phonemic awareness, and Foorman and Torgesen, 2001, Graves, Gerston and Haager, 2004, for specific examples of highly explicit instruction in phonics). Exhibit C.6 lists the specific citations for examples of highly explicit instructional strategies for each of the five components of reading instruction targeted by the legislation.[12] The specific instructional strategies, or combinations of strategies used together, that were considered to be "highly explicit" are presented in Exhibit C.6. This variable was created by dividing the number of intervals that included one or more "highly explicit" instructional practices by the number of intervals that included instruction in one or more of the five dimensions.

Percentage of intervals of instruction in the five dimensions that included at least one instance of high quality student practice. "High quality student practice" is also defined differently in each dimension of reading instruction, based on research published in the National Reading Panel report (National Institute of Child Health and Human Development, 2000) as well as more recent research (e.g., Armbruster, Lehr and Osborn, 2003 for specific examples of high quality student practice in phonemic awareness, and Rasinski and Oswald, 2005, for specific examples of high quality student practice in phonics). Exhibit C.6 lists the specific citations for examples of high quality student practice for each of the five dimensions of reading instruction targeted by the legislation. The specific instructional strategies,

---

[12] No codes in the fluency dimension were classified as "highly explicit" instruction. Helping beginning readers build fluency inherently rests on providing students high quality practice opportunities, rather than delivering explicit instruction in how to read fluently. As a result, codes in the fluency section were used only in the construction of the high quality student practice variable.

Exhibit C.6: Composite of Classroom Constructs

Minutes spent in instruction in each of the five dimensions of reading instruction

Number of minutes spent in any teacher instruction or student practice activity on the IPRI that was in the five dimensions of reading instruction emphasized in Reading First:

- Phonemic awareness
- Phonics/decoding
- Vocabulary
- Fluency
- Comprehension
- All five dimensions combined

Percentage of observation intervals with instruction in the five dimensions of reading instruction with one or more instance of highly explicit instruction

An observation interval was coded as containing instruction in the five dimensions of reading instruction and at least one instance of highly explicit instruction if one or more of the following teacher activities (or combination of activities) was observed during instruction in one of the four reading dimensions that included highly explicit instructional activities.

Phonemic Awareness:[13]

- Teacher demonstrates or models oral blending or segmenting with phonemes *in conjunction with*:
  - Giving students practice in oral blending or segmenting with phonemes
- Teacher demonstrates or models phoneme isolation *in conjunction with*:
  - Giving students practice in phoneme isolation
- Teacher demonstrates or models phoneme categorization/identity (same/different sounds in words) *in conjunction with*:
  - Giving students practice in phoneme categorization/identity
- Teacher demonstrates or models phoneme deletion, addition, or substitution *in conjunction with*:
  - Giving students practice in phoneme deletion, addition, or substitution
- Teacher contrasts two phonemes to pinpoint a target sound
- Teacher pinpoints what students did incorrectly and gives correct response

Phonics/decoding:[14]

- Teacher identifies words that contrast with or do not follow pattern or rule
- Teacher reminds students of pattern or rule and has students produce or repeat correct response, if a student makes a mistake
- Teacher describes, explains, or identifies, or asks students to describe, explain, or identify a sound-symbol pattern, decoding rule, or a word structure pattern or rule *in conjunction with*:
  - Showing students how to apply a rule or pattern to a whole word example, *and*
  - Giving students chance to practice decoding words
- Teacher describes, explains, or identifies, or asks students to describe, explain, or identify a sound-symbol pattern, decoding rule, or a word structure pattern or rule *in conjunction with*:
  - Showing students how to apply a rule or pattern to a whole word example, and
  - Giving students practice encoding words by manipulating or writing letters

---

[13] Ball and Blachman (1991); Bus and van Ijzendoorn (1999); Foorman et al. (1998); Graves et al. (2004); Gunn et al. (2002); Hatcher et al. (2004); McCutchen et al. (2002); Torgesen et al. (1999).

[14] Foorman et al. (1998); Foorman and Torgesen (2001); Graves et al. (2004).

Exhibit C.6: Composite of Classroom Constructs (continued)

Highly explicit instruction

Vocabulary:[15]
- Teacher goes beyond synonym with definition and/or examples
- Teacher pinpoints word meaning by giving contrasting examples
- Teacher pinpoints word meaning by clarifying or extending a partially correct student response
- Teacher pinpoints word meaning by clarifying or extending a partially correct student response with a synonym, definition, example, or contrasting example

Teacher uses a picture, object, or physical demonstration to illustrate word meaning in conjunction with any other vocabulary instructional behaviors including those above and the following:
- Teacher asks students to give meaning of word
- Teacher gives synonym
- Teacher asks students to apply understanding of word meaning
- Teacher gives students opportunity to practice word learning strategies (e.g., using context, word structure, or root meanings)

Comprehension:[16]
Before, during, or after reading a text passage, teacher describes or explains, or asks students to describe or explain one or more comprehension strategies by specifying:
- What the comprehension strategy is called, *and*
- Why the comprehension strategy is helpful, *and*
- When in the reading process the comprehension strategy is used

During or after reading a text passage, teacher shows how to apply strategy by modeling how to:
- Answer inferential questions based on text
- Make predictions based on text
- Summarize, retell, sequence text, or identify the main idea(s)
- Make text-to-text connections
- Generate own questions about text
- Answer own questions about text
- Review passage to check or clarify understanding
- Check accuracy of prediction or inference
- Work with story or expository structure

If a student response is incorrect or incomplete, teacher assists student in using strategy(ies)

---

[15] Brett et al. (1996); Graves et al. (2004); Kamil (2004); McKeown et al. (1985); Tomesen and Aarnoutse (1998).

[16] Crowe (2005); Kamil (2004); Mason (2004); O'Connor et al. (2002); Rosenshine et al. (1996).

Exhibit C.6: Composite of Classroom Constructs (continued)

Percentage of observation intervals with instruction in the five dimensions of reading instruction with one or more instance of high quality student practice

An observation interval was coded as containing instruction in the five dimensions of reading instruction *and* at least one instance of high quality student practice if one or more of the following teacher activities (or combination of activities) was observed during instruction in one of the five dimensions.

Phonemic Awareness:[17]
- Teacher gives students practice in oral blending or segmenting with phonemes while working with pairs or small groups
- Teacher gives students practice in phoneme isolation while working with pairs or small groups
- Teacher gives students practice in phoneme categorization/identity (same/different sounds in words) while working with pairs or small groups
- Teacher gives students practice in phoneme deletion, addition, or substitution while working with pairs or small groups

Phonics/Decoding:[18]
- Teacher gives students practice encoding words by manipulating or writing letters

Vocabulary:[19]
- Teacher gives students the opportunity to practice word learning strategies (e.g. context, word structure, and root meanings)

Fluency:[20]
- Teacher gives students the opportunity to repeat oral readings with same text that was modeled by a fluent reader

Comprehension:[21]
During or after reading a text passage, teacher gives students practice in applying strategy by having students:
- Generate own questions about text
- Answer own questions about text
- Review passage to check or clarify understanding
- Work with story or expository structure *in conjunction with*:
  - Using a text organizer for support
- Check accuracy of prediction or inference
- Justify their response with evidence

---

[17] Ambruster et al. (2003); National Institute of Child Health and Human Development (2000).

[18] Rasinski and Oswald (2005).

[19] Ambruster et al. (2003); National Institute of Child Health and Human Development (2000).

[20] Graves et al. (2004); O'Connor et al. (2002); Stahl (2004); Therrien (2004).

[21] Kamil (2004); Mason (2004); Reutzek and Hollingsworth (1991); Taylor et al. (2002).

or combinations of strategies used together, that were considered to be "high quality student practice" are presented in Exhibit C.6. This variable was created by dividing the number of intervals that included one or more instance of "high quality student practice" by the number of intervals that included instruction in one or more of the five dimensions.

Field Reliability of the IPRI

In each wave of data collection, experienced IPRI trainers were paired with a random sample of classroom observers to collect data necessary to measure the field-based reliability of the IPRI.[22] In contrast to determining the accuracy of an individual observer for purposes of training and hiring, the purpose of field-based inter-rater reliability (IRR) estimates is to assess the reliability of the instrument itself. Researchers often characterize the reliability of an observation instrument by estimating an intra-class correlation (ICC), defined here as the proportion of variance associated with observers relative to the total variance in the collected data. That is, the team sought to characterize the proportion of variance in the observation data due to each of three sources:

- inter-observer differences
- inter-classroom differences
- random measurement error

The RFIS Team used several approaches to attempt to capture the degree of error that can be attributed to observers themselves (as opposed to random measurement error or other forms of systematic measurement error). These approaches included: (1)(a) calculating a pseudo intraclass correlation (ICC) by running an unconditional Hierarchical Linear Model (HLM), and (b) correlating Observer A's and Observer B's codes across multiple intervals within an observation and then averaging these correlations across pairs of observers, and (2) calculating a generalizability coefficient within the generalizability framework (Brennan, 2001; Cronbach et al., 1972 as cited in Brennan, 2001; and Shavelson and Webb, 1991).

Using a Pseudo Intraclass Correlation to Describe Inter-rater Reliability
In the context of measuring inter-rater reliability of the IPRI based on paired field observations, consider the following model:

$$X_{cr} = \mu + v_c + v_r + v_{cr} \tag{1}$$

In (1), X_{cr} is the outcome measure for classroom c, as rated by observer r; μ is the mean outcome across classrooms; and v_c, v_r, and v_{c4} are independent error terms associated with the variance across classrooms, systematic measurement error introduced by the observers, and random measurement error; each with a mean of 0 and variances of σ_c, σ_r, and σ_{cr}. Using this model, we can define the proportion of the total measurement variance that is due to the systematic measurement error introduced by the observers ρ_1 and the proportion of the true variance across classrooms ρ_2 as follows:

---

[22] Half of the field observers were paired with co-observers in spring 2005. In the second year, field observers were paired with co-observers either in fall 2005 or in spring 2006; the majority of field observers were paired for observation once during the 2005-06 school year. In the third year, approximately half of the observers were paired with co-observers either in fall 2006 or spring 2007.

$$\rho_1 = \frac{\sigma_r^2}{\sigma_c^2 + \sigma_r^2 + \sigma_{cr}^2} \tag{2}$$

$$\rho_2 = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_r^2 + \sigma_{cr}^2} \tag{3}$$

(2) indicates the proportion of error that can be attributed to variation across individual observers (observers may vary in their skill at using the IPRI). An examination of (3) shows that to the extent that variance attributable to observers (σ_r^2) is low, the proportion of variance due to true variance across classrooms is high (assuming that random measurement error is small); as ρ_1 decreases, ρ_2 increases. Thus, the lower the ICC, the higher the reliability of the IPRI.

Ideally, intra-class correlations are calculated using a fully crossed design, such that each of a set of *R* observers observes each of C classrooms. In a fully-crossed design, the variance associated with individual observers can be estimated separately from the systematic error associated with individual classrooms. However, a fully-crossed design was not possible in the context of the RFIS, which used 150 observers to record instruction in approximately 1,400 classrooms during each round of data collection. Instead, joint observations were conducted in a sample of classrooms by two observers, one a master observer and the other a member of the field staff. No individual observed more than a small subset of the total number of classrooms. Thus, these data do not allow separate estimate variation due to rater or classroom alone.

The RFIS Team obtained pseudo-ICC estimates using field IRR samples as if they were fully crossed. Such estimates provide a biased estimate of the actual error due to observers, because they also include some of the error associated with inter-classroom differences; however, the estimates are conservative, attributing *more* error to observers than they would in a fully-crossed design. Therefore, if the pseudo-ICC estimates of inter-observer error are low, despite the fact that they include error associated with the individual classrooms, we can be confident that the true amount of error due to differences between observers is even lower—and thus that the IPRI is a reliable instrument.

Most study classrooms were jointly observed for about 30 three-minute intervals, although some joint observations covered fewer and others covered more intervals. In order to construct a fully balanced sample, for each wave of the field IRR samples, the study team (i) dropped classrooms that were observed for fewer than 25 intervals; and (ii) included only the first 25 observation intervals from classrooms that were observed for more than 25 intervals.[23] As a result, the reliability was calculated with 65 classrooms from spring 2005, 62 classrooms from fall 2005, 36 classrooms from spring 2006, 24 classrooms from fall 2006, and 37 classrooms from spring 2007 data collections to assess field-based IRR. (See Exhibit C.7.)

For each of the analytic variables created from IPRI data, the team calculated reliability estimates by estimating the variance terms in equations 2 and 3 (σ_c, σ_r, and σ_{cr}) and by running an unconditional HLM with the field IRR samples for each observation wave. Each HLM was a two-level model with observer

---

[23] The 25-interval threshold attempts to balance two sometimes competing constraints: (i) minimizing the number of classrooms that would be dropped due to lack of observations and (ii) maximizing the number of observation intervals that could be used to assess IRR.

Exhibit C.7: Unconditional HLM Models to Estimate Pseudo-ICCs (ρ1) and True Variance Across Classrooms (ρ2)

Outcome	Spring 2005 (n=65)		Fall 2005 (n=62)		Spring 2006 (n=36)		Fall 2005 (n=24)		Spring 07 (n=37)	
	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2
Number of Minutes Spent on Phonics	0.046	0.930	0.025	0.959	0.059	0.914	0.175	0.774	0.149	0.805
Number of Minutes Spent on Comprehension	0.049	0.927	0.079	0.888	0.025	0.959	0.080	0.888	0.020	0.966
Number of Minutes Spent on Vocabulary	0.038	0.941	0.067	0.904	0.049	0.926	0.027	0.956	0.107	0.854
Number of Minutes Spent on Phonemic Awareness	0.111	0.849	0.25	0.684	0.030	0.952	0.072	0.897	0.040	0.938
Number of Minutes Spent on Fluency Building	0.170	0.779	0.069	0.901	0.075	0.893	0.031	0.951	0.019	0.968
Number of Minutes Spent on Five Dimensions Combined	0.061	0.912	0.096	0.868	0.058	0.915	0.080	0.888	0.024	0.961
Proportion of Intervals in the 5 Dimensions Containing Highly Explicit Instruction	0.281	0.654	0.327	0.604	0.375	0.551	0.045	0.933	0.395	0.530
Proportion of Intervals in the 5 Dimensions Containing High Quality Student Practice	0.265	0.670	0.274	0.662	0.303	0.632	0.135	0.822	0.439	0.482

NOTE:

The HLM model utilized for this analysis includes an intercept and three independent random error terms that are associated with the variance across classes, systematic measurement error introduced by the raters, and random measurement error. Definitions of ρ_1 and ρ_2 can be found in the text.

EXHIBIT READS: The proportion of variance due to differences between observers for Number of Minutes Spent on Phonics was .046 for the 65 co-observed classrooms from spring 2005. The proportion of variance due to differences between classrooms for Number of Minutes Spent on phonics was .930 for the 65 classrooms from spring 2005.

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

(A or B) nested within classroom, for each classroom that had been co-observed. Next ρ_1 and ρ_2 were calculated using these estimates. Corresponding results are presented in Exhibit C.7 and indicate that ICC-based reliability estimates (ρ_2) are consistent across the five observation waves, ranging from 0.868 to 0.961, for example, for the number of minutes spent on the five dimensions combined.

An alternative way of obtaining a pseudo ICC estimate is by simply correlating the two observers' codes *within* a given observation and *across* the multiple intervals with that observation, and averaging these correlations across the pairs of coders. Similar to the unconditional HLM model, using this method with the co-observation data attributes more error to the observers than it should. This method is also complicated when one observer reports that a specific IPRI code never occurred during an entire observation, but the other observer reports that the same code occurred (at least once); in this case, the correlation coefficient is not defined (these observations were not included in this analysis). In contrast, if both observers agreed that a particular IPRI code never occurred, we imputed the correlation coefficient to be one since these cases could be regarded as perfect agreement. Exhibit C.8 presents estimates of this pseudo ICC with the number of observations used for the calculations. As expected, these results are very similar to the ones from the unconditional HLM model in Exhibit C.7.

Using a Generalizability Coefficient to Measure Inter-rater Reliability
Recall that the previous approach of using a pseudo-ICC to measure the field-based reliability of the IPRI assumes that the field IRR samples are fully crossed. One way to account for the fact that the field IRR samples are not fully crossed and still be able construct an estimate of field based reliability is to calculate a generalizability coefficient using the generalizability framework. The generalizability framework can be defined as a "theory that liberalizes classical theory by employing ANOVA methods that allow an investigator to untangle multiple sources of error" to describe the reliability of a measurement (Cronbach, et al., 1972, as cited in Brennan 2001.)

In field IRR samples, each classroom (c) is observed by a different set of two observers (or raters, [r]) simultaneously during a number of intervals (i). In the generalizability framework, discussed in detail by Brennan (2001), this set-up could be regarded as a **G study (r: c) * i** design with n_c that were observed by 2 observers (n_r=2) for 25 intervals (n_i=25).[24] The main and interaction effects for this model can be depicted as:

Let X_{cri} denote the outcome (an IPRI item) recorded in classroom *c* by rater *r* at interval *i*. Utilizing the effects presented in Exhibit C.9, we can describe this outcome as follows:

$$X_{cri} = \mu + \nu_c + \nu_i + \nu_{r:c} + \nu_{ci} + \nu_{ri:c} \tag{4}$$

---

[24] Note that here an interval is regarded as the object of measurement.

Exhibit C.8: Average Correlation Between Paired Observers' Codes Across Classrooms

Outcome	Spring 2005		Fall 2005		Spring 2006		Fall 2006		Spring 2007	
	Average Correlation	N[1]	Average Correlation	N[1]	Average Correlation	N[1]	Average Correlation	N[1]	Average Correlation	N[1]
Phonics	0.869	65	0.815	60	0.835	32	0.840	23	0.866	35
Comprehension	0.866	65	0.890	62	0.841	35	0.915	23	0.885	35
Vocabulary	0.829	65	0.836	60	0.811	34	0.816	24	0.836	36
Phonemic Awareness	0.990	55	0.976	60	0.963	34	0.859	23	0.942	35
Fluency Building	0.946	50	0.963	55	0.955	36	0.950	23	0.915	30
Five Dimensions Combined	0.845	65	0.836	61	0.807	35	0.845	24	0.864	37
Highly Explicit Instruction	0.579	63	0.649	57	0.590	35	0.705	22	0.688	36
High Quality Student Practice	0.679	60	0.764	52	0.710	24	0.820	23	0.821	28

Notes:

[1] The effective N is shown for the calculation of the average correlation between observer and co-observer codes. Co-observations in which *only* one of the observers reported that the outcome of interest occurred in every interval (or did not occur in any of the intervals) are excluded from the analysis as for such cases, the correlation coefficient could not be calculated. Co-observations in which both of the raters reported that the outcome of interest occurred in every interval (or did not occur in any of the intervals) are included in the analysis with a correlation coefficient of 1.

EXHIBIT READS: The average correlation between paired observers' codes across classrooms for phonics was .869 in spring 2005 (n=65), .815 in fall 2005 (n=60), .835 in spring 2006 (n=32), .84 in fall 2006 (n=23), and .866 in spring 2007 (n=35).

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

Exhibit C.9: Main and Interaction Effects in a (r: c)*i Design

Model	Main Effects	Interaction Effects
*(r: c) * i*	*i, c, r:c*	*ci, ri:c*

Here, μ is the grand mean in the population and ν terms represent the five main and interaction effects listed in Exhibit C.9 (*i, c, r:c, ci, ri:c*). Using (5), one can decompose the total variance observed in the outcome into five independent variance components associated with the effects as follows:

$$\sigma^2(X_{cri}) = \sigma^2(\nu_c) + \sigma^2(\nu_i) + \sigma^2(\nu_{r:c}) + \sigma^2(\nu_{ci}) + \sigma^2(\nu_{ri:c})$$
$$= \sigma^2(c) + \sigma^2(i) + \sigma^2(r:c) + \sigma^2(ci) + \sigma^2(ri:c)$$

(5)

Using this general framework, a measure of the IRR for a single *random* rater ($n_r = 1$) observing a single *fixed* classroom ($n_c = 1$) can be calculated using a **D-*study* (R:C) * i** design. This design is sufficient if one wants to estimate a general IRR across all possible pairs of raters, such that the correlation between a pair of raters estimates the reliability of a single rater, and it is not necessary to generalize across all classrooms. Under a D-study, the IRR estimate is given by the generalizability coefficient, $E\rho^2$, defined in equation (6):

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} = \frac{\sigma_i^2 + \sigma_{ci}^2}{\sigma_i^2 + \sigma_{ci}^2 + \sigma_{ri:c}^2}$$

(6)

In (6), $\sigma^2(\tau)$ and $\sigma^2(\delta)$ denote the universe score variance and variance of the relative error respectively. Exhibit C.10 demonstrates the formulas that could be used to calculate the variance components of the generalizability coefficient $E\rho^2$. Technically, $E\rho^2$ can be interpreted as an intra-class correlation coefficient, which approximates the expected value of the squared correlation between the observed outcome and the universe ("true") outcome for a classroom. In this context, the universe outcome can be defined as the expected value of the mean outcomes for every instance of the measurement procedure (i.e., the mean of the outcomes coded by all possible sets of two observers) of a classroom. Alternatively, $E\rho^2$ can also be seen as the ratio of variance of the universe outcome to the variance of the observed outcome. The difference between the pseudo ICCs described earlier and the generalizability coefficient $E\rho^2$ is that $E\rho^2$ takes into account the fact that each classroom was observed by a different set of two observers during co-observations, whereas the former simply ignores this fact.

Exhibit C.11 presents estimates of the generalizability coefficient calculated using the five waves of the IPRI field IRR data. These estimates of reliability are slightly lower than the reliability estimates determined by calculating pseudo ICC estimates shown in Exhibits C.7 and C.8. One possibility for these estimates being slightly lower is that the generalizability coefficient accounts for the fact that the sample is not fully crossed.

Overall, the various methods of estimating IRR using observation and co-observation data provide consistent results. The reliability estimates for the five dimensions (phonics, comprehension, vocabulary, phonemic awareness, and fluency building) are consistent across all methods. The estimates for highly explicit instruction and high quality student practice measures are lower, a finding that might reflect the fact these measures attempt to capture micro behaviors that are harder for observers to recognize and code accurately.

Exhibit C.10: Calculating Variance Components for a (r: c)*i Design

α	df(α)	T(α)	SS(α)	MS(α)	$\hat{\sigma}^2(\alpha) \equiv \hat{\sigma}_\alpha^2$
i	n_i-1	$n_c n_r \sum_i \overline{X}_i^2$	T(i)- T(μ)		$\dfrac{MS(i) - MS(ci)}{n_c n_r}$
c	n_c-1	$n_i n_r \sum_c \overline{X}_c^2$	T(c)- T(μ)		$\dfrac{MS(c) - MS(r:c) - MS(ci) + MS(ri:c)}{n_i n_r}$
r: c	$n_c(n_r$-1)	$n_i \sum_c \sum_r \overline{X}_{r:c}^2$	T(r: c)- T(c)	$\dfrac{SS(\alpha)}{df(\alpha)}$	$\dfrac{MS(r:c) - MS(ri:c)}{n_i}$
ci	$(n_c$-1)(n_i-1)	$n_r \sum_c \sum_i \overline{X}_{ci}^2$	T(ci)-T(c)-T(i)+ T(μ)		$\dfrac{MS(ci) - MS(ri:c)}{n_r}$
ri: c	$n_c(n_r$-1) (n_i-1)	$\sum_c \sum_r \sum_i X_{cri}^2$	T(ri: c)- T(ci) – T(r:c)+ T(c)		$MS(ri:c)$

Notation:

    α: any of the main and interaction effects

df(α): degrees of freedom for effect α

T (α): sum of squared mean scores for effect α

$T(\mu) = n_r n_c n_i \overline{X}^2$

SS (α): sum of squares for α

MS (α): mean squares for α

$\hat{\sigma}^2(\alpha)$: estimated variance component for effect α

X_{cri} : outcome of interest for class *c* as rated by rater *r* in interval *i*

Exhibit C.11: Generalizability Coefficients Estimated from the Co-Observation Data

Outcome	Spring 2005 (n=65) $\mathrm{E}\rho^2$	Fall 2005 (n=62) $\mathrm{E}\rho^2$	Spring 2006 (n=36) $\mathrm{E}\rho^2$	Fall 2006 (n=24) $\mathrm{E}\rho^2$	Spring 2007 (n=37) $\mathrm{E}\rho^2$
Phonics	.859	.820	.807	.830	.852
Comprehension	.863	.881	.820	.911	.869
Vocabulary	.812	.769	.796	.821	.835
Phonemic Awareness	.802	.822	.792	.723	.582
Fluency Building	.706	.826	.827	.875	.683
Five Dimensions Combined	.841	.843	.799	.848	.886
Highly Explicit Instruction	.577	.610	.545	.691	.668
High Quality Student Practice	.625	.574	.443	.751	.551

EXHIBIT READS: The generalizability coefficients for Phonics are .859 for spring 2005, .820 for fall 2005, .807 for spring 2006, .830 for fall 2006, and .852 for spring 2007.

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

Part 3: Global Appraisal of Teaching Strategies (GATS)

The Global Appraisal of Teaching Strategies (GATS) is a 12-item checklist (see Exhibit C.20 for a copy of the GATS) adapted from the Checklist of Teacher Competencies (Foorman et al., 2006).[25] 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 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.[26]

For the first ten items on the GATS, the observer indicated how often the teacher demonstrated the behaviors targeted by each item, using a five point scale where 1 = all the time, 2 = more than half the time, 3 = half the time, 4 = less than half the time, and 5 = never. (A sixth "not observed" option was included in case the item did not apply during a particular observation; e.g., "takes advantage of opportunities to provide corrective feedback when student makes error" may not have been observed if no student errors occurred during the observation). For items 11 and 12, the observer also rated the teacher on a five point scale, but the descriptors of these scales varied to match the focus of the question.[27]

---

[25] B. Foorman graciously shared the Checklist of Teacher Competencies with the RFIS Study Team but played no role in the development of the GATS. See Foorman & Schatschneider (2003) and Foorman et al., (2006).

[26] The RFIS did not conduct inter-rater reliability analyses for the GATS data, because the study did not include GATS data in impact analyses, and study resources were focused on obtaining reliability data for observational data that were included in impact analyses.

[27] For item 11, the descriptors were 1 = true, 2 = mostly true, 3 = not sure, 4 = mostly not true, and 5 = not true. For item 12, the descriptors were 1 = all students, 2 = more than half of students, 3 = half of students, 4 = less than half of students, and 5 = none.

Factor analyses revealed that all 12 items loaded onto a single factor. Thus, each classroom was assigned a single GATS score, as follows:

- any item marked "not observed" was set to missing.

- the rating for each item was averaged across the two GATS observations for the classroom in each data collection wave.

- the 12 items were averaged together.

Before analyses, the assigned GATS score was subtracted from 6 to reverse the scale so that a higher rating corresponded to more frequent demonstration of the behavior.

Part 4: Student Time-on-Task and Engagement with Print (STEP)

The Student Time-on-Task and Engagement with Print (STEP) instrument was designed to capture information about student engagement during reading instruction as part of the Reading First Impact Study's (RFIS) classroom observation data collection. The STEP is focused on student behavior; it complements the Instructional Practice in Reading Inventory (IPRI) measure, which focuses on teacher behaviors. See Appendix C, Part 7, Exhibit C.21 for a copy of the STEP.

The STEP was designed to collect aggregate, not individual level, data on the percentage of students in classrooms during the scheduled reading block who are on-task and/or interacting with print. The STEP instrument combines a dichotomous "on-task/off-task" rating with additional indicators for student engagement with print.

The data collected with the STEP instrument do not measure the amount of time students are on-task or the amount of time students are engaged with print. Rather, across all students in the classroom, the STEP instrument yields data on the percentage of students who, at a particular point in time, are on-task and engaged with print.

During each wave of classroom observation data collection, one observer per school was assigned to collect student engagement data in each classroom being observed by IPRI observers. STEP observations took place during the reading block in each classroom. While each classroom was observed twice for the IPRI, each classroom was observed once for the STEP.

Each STEP observation consists of data on student engagement from three sweeps of a classroom. Specifically, for each sweep, at an interval of six minutes, an observer classifies every student in the classroom as either on- or off-task, and, if the student is on-task, whether the student is:

a) reading connected text (e.g., a paragraph, story, or longer passage); or

b) reading isolated text (letters, words, or sentences in isolation); and/or

c) writing; or

d) none of the above (i.e., not engaged with print).

A student can be marked as on-task without being engaged with print (but a student cannot be off-task and engaged with print). An on-task student can also be engaged with more than one type of print (e.g., the student is writing on a worksheet that contains isolated text, such as a list of words). The observer records student behavior for each student in each observed classroom three times.

Between sweeps, the observer waits until six minutes have elapsed before beginning the next sweep. After the third sweep, the observer moves on to the next classroom in the sample. The observation protocol is summarized in Exhibit C.12.

Exhibit C.12: Prototypical STEP Observation in One Classroom			
Classroom A	**Duration (minutes)**	**Sample Clock Time**	**Activity**
Rest period 1	6	8:00-8:06	Observer waits for children to acclimate
Sweep 1	3	8:06-8:09	Observer records data on each student in classroom
Rest period 2	3	8:09-8:12	Observer waits
Sweep 2	3	8:12-8:15	Observer records data on each student in classroom
Rest period 3	3	8:15-8:18	Observer waits
Sweep 3	3	8:18-8:21	Observer records data on each student in classroom
Switch classes	6	8:21-8:27	Observer exits Classroom 1 and moves to next classroom
Total time per classroom	**27 min**	Time is approximate (travel time between classrooms may be shorter or longer than 6 minutes)	

NOTE:

The duration of a sweep varies depending on how long it takes the observer to record data on all students in the classroom, but never exceeds three minutes. Exactly six minutes separate the start of one sweep and the start of another.

Under certain circumstances, observers skipped a scheduled sweep. First, if at the time of a scheduled sweep, more than one-half of the students in the classroom were transitioning from one activity to another (e.g., students were rotating between activity "centers"), the observer skipped that sweep. Second, if at the time of a scheduled sweep, the whole class was listening to the teacher read aloud, and the students themselves did not have access to the printed text, the observer skipped that sweep.[28]

Data Collection and Response Rates for Fall 2005, Spring 2006, Fall 2006, and Spring 2007

The STEP was added to the classroom observation data collection battery beginning in fall 2005, reflecting a decision by IES staff (Institute for Education Sciences, U.S. Department of Education) overseeing the RFIS to augment the teacher-focused data collection (using the IPRI) with a student-focused measure. STEP observations were done in grade 1 and 2 classrooms in fall 2005, spring 2006, fall 2006, and spring 2007 by trained field staff who had successfully completed the requirements of the classroom observation training. As described above, during two consecutive days of classroom observations, STEP observations were completed once in each classroom, yielding one STEP record per classroom. For 2005-2006 a total of 2,715 STEP observations were completed in first and second grade

---

[28] These protocols were implemented because pilot-testing of the instrument revealed that on- and off-task judgments were difficult to make reliably under these two circumstances.

classrooms, which represents a 98 percent completion rate for expected observations. In 2006-2007, a total of 2,764 first and second grade observations were made, or 98 percent of expected observations. A flow chart of the sampling process and STEP response rates is presented in Exhibit C.13.

Analytic Variables

The RFIS Team focused on the percentage of students engaged with print as the primary analytic variable derived from the STEP data to be used in impact analyses. This variable was created for each classroom by first summing the number of students in each sweep who were on-task and who were either reading connected text, reading isolated text, or writing. The percentage of students engaged with print for each sweep was then calculated as the number of students engaged with print divided by the total number of students that the observer rated in the sweep (i.e., the number of students in the classroom at the time the sweep was conducted). The percentage of students engaged with print for each sweep was then averaged across the number of sweeps available for that classroom.[29]

STEP Reliability

For reasons of parsimony, results from the fall 2006 STEP training are presented below. Observers were trained on the STEP measure using a combination of still photographs and 3-second video clips of first and second grade students during reading instruction.[30] Trainees viewed five practice sequences, containing both still photographs and short video clips. A sixth sequence of video clips (hereafter, the "test tape") was used to assess the average inter-rater reliability of observers' judgments about student engagement.

The test tape was designed to simulate a single "sweep," and it included three-second clips of 15 first- or second-grade students. Two master coders had viewed and scored the test tape to arrive at a set of master codes for each student on the tape.

Percent agreement was calculated for each trainee with the master codes for each code (i.e., On-Task, Reading Connected Text, Reading Isolated Text, Writing), and then a mean percent agreement was calculated across trainees for each code. Next, overall percent agreement was calculated by aggregating across codes.

---

[29] For the pooled analytic dataset (fall 2005, spring 2006, fall 2006, and spring 2007), 70 percent of classrooms had three sweeps of data; 23 percent had two sweeps of data; 5 percent had one sweep of data; and 1 percent was missing all three sweeps. (Numbers do not add to 100% due to rounding.)

[30] Classroom reading instruction was filmed in both Reading First and non-RF classrooms for the purpose of creating a training resource for the RFIS.

Exhibit C.13: STEP Data Collection: School, Classroom, and Observation Sample Information

RFIS Schools

| RF: 125 | Non-RF: 123 | Total: 248 |

All Grade 1 & 2 Classrooms

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	1,364	1,378	1,384	1,390
Non-RF:	1,343	1,355	1,342	1,358
Total:	2,707	2,733	2,726	2,748

Sampled Grade 1 & 2 Classrooms[1]

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	1,010	1,022	1,033	1,035
Non-RF:	979	986	953	964
Total:	1,989	2,008	1,986	1,999

Classrooms That Did Not Meet Criteria[2]

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	90	100	97	95
Non-RF:	77	81	58	59
Total:	167	181	155	154

Classrooms That Met Criteria

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	920	922	936	940
Non-RF:	902	905	895	905
Total:	1,822	1,827	1,831	1,845

Excluded Classrooms[3]

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	224	226	221	226
Non-RF:	214	215	206	209
Total:	438	441	427	435

Final Observation Sample

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	696	696	715	714
Non-RF:	688	690	689	696
Total:	1,384	1,386	1,404	1,410

Observations Completed[4]

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	683 (98%)	677 (97%)	705 (99%)	702 (98%)
Non-RF:	678 (99%)	677 (98%)	675 (98%)	682 (98%)
Total:	1,361 (98%)	1,354 (98%)	1,380 (98%)	1,384 (98%)

Analytic Sample[5,6]

	Fall 2005	Spring 2006	Fall 2006	Spring 2007
RF:	669 (98%)	675 (99%)	700 (99%)	700 (99%)
Non-RF:	652 (96%)	663 (98%)	666 (99%)	676 (99%)
Total:	1,321 (97%)	1,338 (99%)	1,366 (99%)	1,376 (99%)

Notes:

[1] The study conducted observations in all classrooms in schools if across the site as a whole, the average number of classrooms per grade per school was three or less. If for that site as a whole, the average number of classrooms per grade per school exceeded three, the study sampled classrooms for observations within schools.

[2] At the beginning of each data collection wave, the study team contacted schools to obtain classroom rosters and indications of which classrooms were designated as regular classrooms, self-contained special education classrooms, or ESL classrooms. Those classrooms identified as special education or ESL were considered ineligible for observation.

[3] Once the study team began to schedule and conduct observations, classrooms were excluded from the sample if the information that had been provided earlier about special education or ESL status was inaccurate, if the class was taught by someone other than the regular teacher or a long-term substitute, or if fewer than 75 percent of the students were in the target grade.

[4] Response rates are calculated by dividing the number of observations completed by the eligible classrooms selected into the sample.

[5] Classrooms were dropped from the analytic sample if, for all three sweeps, the class was in transition between activities or the entire class was listening to a story. Of the 78 classrooms (1%) for which this was the case, 23 were Reading First and 55 were non-Reading First, which corresponds to 1% of Reading First classrooms and 2% of non-Reading First classrooms in the pooled analytic sample.

[6] Percentages are calculated by dividing the analytic sample numbers by the number of observations completed.

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

As shown in Exhibit C.14, observers achieved an average of 89 percent agreement across all codes appearing in the test tape. Seventy-five percent of the observers scored at least 86 percent overall agreement. Observers had the lowest average agreement about whether or not a student was Reading Isolated Text (77 percent), and they achieved the highest level of agreement when judging that a student was Writing (96 percent).[31] These differences reflect the fact that the video cameras could zoom in and capture students' expressions more effectively than they could discern the specific types of text with which students were engaged. During actual data collection, observers could move around the classrooms to determine whether students were engaged with specific types of text.

Exhibit C.14: Percent Correct by Code and Overall for STEP Reliability Tape, Fall 2006

| | Percent Agreement | | | | |
| | | Student Is … | | | |
	On Task	Reading Connected Text	Reading Isolated Text	Writing	Overall
Mean	92	92	77	96	89
Minimum	60	67	50	75	73
25th percentile	87	92	67	92	86
50th percentile	93	92	75	100	90
75th percentile	100	92	83	100	92
Maximum	100	100	100	100	100

NOTES:

The number of observers tested on this tape is 130.

EXHIBIT READS: Observers in the fall 2006 training achieved an average of 92 percent agreement on whether a student was on-task; 92 percent agreement on whether a student was reading connected text; 77 percent agreement on whether a student was reading isolated text; 96 percent agreement on whether a student was writing; and 89 percent agreement across all codes appearing in the test tape.

Part 5: Reading Achievement

At the heart of this evaluation is a question about the impact of Reading First on the reading achievement of students. The RFIS had initially planned to use a battery of tests to assess students' reading skill across the components of reading instruction targeted in the legislation (phonemic awareness, phonics, fluency, vocabulary, and comprehension), but when the study's design shifted to RDD, with a much larger number of schools, the planned data collection activities also changed. The RFIS Team, working with its Technical Work Group and staff from the National Center for Education Evaluation/Institute of Education Sciences at the Department of Education, focused its efforts on identifying a single test of reading comprehension. In the spring of 2007, for first graders only, the RFIS added the Test of Silent Word Reading Fluency (TOSWRF, Mather et al., 2004) to assess students' decoding skills, which are distinct from, although related to, reading comprehension.

---

[31] In fall 2005, similar results were obtained from the previous group of observers. They achieved, on average, 87 percent agreement across all codes appearing in the test tape. Seventy-five percent of the trainees scored at least 84 percent overall agreement. Trainees had the lowest average agreement on the Reading Isolated Text code (75 percent), and the highest level of agreement (95 percent) on the Writing code. (The test tape featured only one student who was engaged in Writing.)

Reading Comprehension

Reading Comprehension Instrument Selection

Stanford Achievement Test—10th Edition (SAT 10). The team's priorities in selecting a test for this study included, first, finding a test that directly measured skills related to text comprehension. Other factors included: ease and appropriateness of administration to groups or entire classrooms of students—including appropriateness for fall first grade; modest time demands; use of a norm-referenced test; and consistent reliability and validity. The team also sought a measure that had already been widely used in large-scale studies, and therefore would be more likely to be credible in the research community.

At the outset of the test selection and review process, the team identified 47 assessments of text comprehension that either had been proposed for use by states in their Reading First schools or had been proposed for use in other Department of Education-sponsored evaluations involving preschool and the early elementary grades. From this pool of tests, we identified six test batteries with subtests of reading comprehension that could be group-administered and were valid for fall of first grade.[32] The six test batteries included:

1. ITBS Total Core Battery Reading Subtest;
2. Terra Nova/CTBS Basic Battery Reading Subtest;
3. Gates/MacGinitie Reading Test-3 (GMRT);
4. GRADE (Group Reading Assessment and Diagnostic Evaluation);
5. Stanford Achievement Test—10th Edition (SAT 10); and
6. Stanford Reading First.

Five of the six tests have reliability coefficients reported in published manuals of close to 0.90 for the majority of subtests. Because the reliability for Terra Nova Grade 1 was 0.76, and data were not available for the other grade levels, that test was eliminated from consideration. The Stanford Reading First Test was also eliminated, because it had been normed on a relatively small sample according to a conversation with a Harcourt representative in 2004 (< 400 students across several grade levels), whereas the remaining five tests had been normed on samples of 1,000 or more students.

Next, the team reviewed two related aspects of the tests: the number of items and amount of time required. The number of items varies considerably—from approximately 30 to 80, with fewer items typically required for grade 3 tests (although the amount of time required per item increases by grade level). The tests also vary in amount of time required, from 50 minutes for the Stanford Reading First at all three grade levels to 95 minutes for the GRADE in grade 1. The amount of time required was a consideration, but not the deciding factor. The final consideration was the relative frequency of use for the four remaining assessments in schools in the study sample. Of the states and districts that (in Summer 2004) administered standardized reading assessments to children in grades 1, 2, and 3, more used the SAT 10 than any other test (although none did so in fall of grade 1). The study consequently chose the SAT 10 because it both met all the criteria above and because its use might allow the study to collect extant data, which would reduce the testing burden on students and schools. (Where extant data were not available, the study would administer the SAT 10.)

---

[32] See published manuals (Hoover et al., 2003; CTB/McGraw-Hill, 2003; MacGinitie et al., 2000; Williams, 2001; Harcourt Assessment, Inc., 2004).

The specific properties of the SAT 10 are summarized in Exhibit C.15.

Exhibit C.15: Features of SAT 10: Reading/Listening Comprehension for Spring Administration

	Grade Level		
	Grade 1 Spring (Primary 1)	Grade 2 Spring (Primary 2)	Grade 3 Spring (Primary 3)
Number of Items	40	40	54
Time in Minutes	50	50	60
Test-Retest Reliability*	.91	.91	.93
Concurrent Validity	To SESAT-2:[1] .63 Form A to B: .87	To Primary 1: .69 Form A to B: .85	To Primary 2: .80 Form A to B: .83
N in Norming Sample	3,392	3,558	2,160

*Reliability is test-retest Kuder-Richardson formula 20 (KR 20)

[1] Stanford Early School Achievement Test.

SOURCE: Harcourt Assessment, Inc. (2004)

Decoding

Test of Silent Word Reading Fluency (TOSWRF). In the spring of 2007, the study team augmented the SAT 10 reading comprehension assessment with the TOSWRF for first grade students only. The TOSWRF is a short (three-minute) assessment that measures students' ability to identify words quickly and correctly. It assesses students' decoding skills, which are distinct from, although related to, reading comprehension.

The RFIS added the TOSWRF because decoding is an important reading sub-skill. Based on recommendations from reading researchers, the TOSWRF was selected from among several possible decoding measures. Key factors in selecting an assessment included both pragmatic data collection considerations, including the manner of administration (individual versus group) and length of administration, and psychometric properties of the assessment. It was important that the assessment be not only valid and reliable, but also group administered and brief, in order to limit changes to the data collection protocol in the last year of the study.

The TOSWRF was normed in 2001 and 2002 using a representative sample of 3,592 students residing in 32 states. Raw scores can be converted to standard scores, percentile ranks, and age and grade equivalents. Reliability coefficients for the TOSWRF are .90 or higher across four types of reliability: alternate form, test-retest, alternate form (delayed administration), and scorer differences. Although a relatively new test, initial validity of the TOSWRF has been established by correlating it with other measures of reading fluency, word identification, and comprehension. Corrected coefficients for criterion-prediction validity range from .42 to .78.[33]

---

[33] For specific tests used in validity analyses and for more information on the psychometric properties, see the *TOSWRF Examiner's Manual* (Mather, N., Hammill, D., Allen, E., and Roberts, R., 2004).

Data Collection and Response Rates

In six sites, the RFIS obtained SAT 10 data directly from state and/or district education officials. In 12 sites, the RFIS collected test data directly. In all sites, the RFIS obtained TOSWRF data directly. The SAT 10 student assessments were administered in grades 1, 2, and 3, at four timepoints: fall 2004, spring 2005, spring 2006, and spring 2007, while the TOSWRF was administered in grade 1 only at one timepoint (spring 2007). To conduct the testing, one site assessment coordinator was hired at each district (local), and that coordinator in turn hired a local team of test administrators. Since both the SAT 10 and the TOSWRF are standardized tests, the requirements of the test publisher for administration were followed. Site assessment coordinators also observed each test administrator in the classroom for quality control and technical assistance. In addition, staff from the home office visited districts during the testing for quality control purposes.

The study team collected classroom rosters prior to administration, and used these rosters to pre-label the student test booklets with the student ID and a strippable name label. Once the test booklet was complete, the test administrator stripped the name label from the booklet (for privacy purposes) and adhered it to a receipt sheet. The test administrator then delivered the completed booklets and the receipt sheet to the site assessment coordinator who was responsible for keeping track of who had been tested and who required make-up testing. A computerized field management system allowed the site coordinators to receive the booklets and also to print out a list by school and grade regarding which students needed makeup testing. Once testing was complete in the district, the site coordinator shipped the hardcopy test booklets to be processed.

In fall 2004, there were two main factors in maximizing SAT 10 response rates: obtaining parent permission at more than one timepoint, and administering make-up tests for students who missed the originally scheduled testing sessions. In the initial two weeks of student assessment, the RFIS assessed all students present in the classroom who had returned signed permission slips. Study staff worked with school liaisons prior to the scheduled assessment date to obtain as many permission slips as possible.

For those students who returned permission slips after the scheduled assessment day, or were absent, group make-up sessions were held at each school. Students were not eligible for the assessments if they were excluded from testing in accordance with their own school or district policies (generally because they received instruction primarily in a language other than English), and/or needed special accommodations (particularly an exam writer/scribe). The consent rates and resultant response rates were considerably lower than hoped in fall 2004 (75 and 70 percent, respectively, for Reading First and comparison schools). The RFIS obtained a waiver from participating districts and from the Abt Associates IRB to use passive consent in subsequent testing, which increased the effective SAT 10 response rates to 84 and 83 percent, respectively, for Reading First and comparison schools in spring 2005. During the 2005-06 school year, the effective SAT 10 response rate was 86% for both RF and non-RF schools. During the 2005-06 school year, the effective SAT 10 response rate was 86% for both reading first and non-reading first schools. In 2006-07, the effective SAT 10 response rate was 88% for reading first schools and 85% for non-reading first schools. The TOSWRF response rates were 87% for RF schools and 85% for non-RF schools. A flowchart presenting student assessment sample information by school year in the 12 sites in which the RFIS collected test data directly is presented in Exhibit C.16.

Exhibit C.16: Student Assessment Data Collection: Sample School and Student Information

SAT 10

Panel 1: Data Collected Directly by the RFIS (12 sites) | *Panel 2: Data Collected Directly by the RFIS (12 sites) and Extant Data (6 sites)*

Panel 1

RFIS Schools

RF: 91   Non-RF: 89   Total: 180

All Students in Grades 1-3

	Fall 04	Spring 05	Spring 06	Spring 07
RF:	18,855	19,157	18,379	18,506
Non-RF:	19,409	19,627	19,764	19,901
Total:	38,266	38,784	38,143	38,407

Sample Students in Grades 1-3[1]

	Fall 04	Spring 05	Spring 06	Spring 07
RF:	18,855	19,157	14,616	15,724
Non-RF:	19,409	19,627	14,689	15,794
Total:	38,266	38,784	29,305	31,518

Students Not Eligible to be Tested[2]

	Fall 04	Spring 05	Spring 06	Spring 07
RF:	1,007	1,021	377	440
Non-RF:	1,226	1,247	372	445
Total:	2,303	2,268	749	885

Students Eligible to be Tested

	Fall 04	Spring 05	Spring 06	Spring 07
RF:	17,778	18,136	14,239	15,284
Non-RF:	18,183	18,380	14,317	15,349
Total:	35,961	36,514	28,556	30,633

Eligible Students Not Tested[3]

	Fall 04	Spring 05	Spring 06	Spring 07
RF:	5,052	3,260	2,304	2,220
Non-RF:	5,272	3,734	2,180	2,299
Total:	10,324	6,994	4,484	4,519

Students Eligible to be Tested

	Fall 04	Spring 05	Spring 06	Spring 07
RF:	12,726 (72%)	14,876 (82%)	11,935 (84%)	13,064 (85%)
Non-RF:	12,911 (71%)	14,646 (80%)	12,137 (85%)	13,050 (85%)
Total:	25,637 (71%)	29,522 (81%)	24,072 (84%)	26,114 (85%)

Panel 2

Overall Tested[4,5]

	Spring 05	Spring 06	Spring 07
RF:	22,672 (85%)	19,273 (86%)	19,903 (88%)
Non-RF:	21,097 (82%)	17,227 (86%)	17,272 (85%)
Total:	43,769 (83%)	36,500 (86%)	37,175 (85%)

Analytic Sample[6,7]

	Spring 05	Spring 06	Spring 07
RF:	22,581 (100%)	19,273 (100%)	19,902 (100%)
Non-RF:	20,971 (99%)	17,225 (100%)	17,270 (100%)
Total:	43,552 (100%)	36,498 (100%)	37,172 (100%)

Notes:

The information presented in the top panel represents the 12 sites in which the RFIS collected all test data directly. The information in the bottom panel includes data from those 12 sites plus data from six sites for which the RFIS obtained student test data from state and/or district education officials.

[1] In 2004-05, the study administered SAT 10 tests to all students in grade 1-3 classrooms. In 2005-06 and 2006-07, the study administered SAT 10 tests to all students in grade 1-3 classrooms if across the site as a whole, the average number of classrooms per grade per school was three or less. If for that site as a whole, the average number of classrooms per grade per school exceeded three, the study sampled classrooms within schools and tested students in those classrooms. Students in classrooms that were self-contained special education classrooms or classrooms in which instruction occurred in languages other than English were ineligible for testing.

[2] Students were not eligible for assessments if excluded from testing in accordance with their own school or district policies (e.g., because they received instruction in a language other than English), and/or they needed special accommodations beyond those that could be provided through additional time in a group administered testing situation.

[3] Eligible students were not tested if they were absent at the time the test was given and could not be rescheduled, they had transferred out, they had refused to take the test, or the RFIS did not have consent for them to participate in the study.

[4] Boxes in Panel 2 do not include data from fall 2004 because those data were used only to construct a pretest covariate at the school level.

[5] Response rates for the six sites for which the RFIS obtained student test data from state and/or district education officials were estimated.

[6] Students test scores were excluded from the analytic sample if they were deemed invalid according to SAT 10 scoring guidelines. For the pooled sample (spring 2005, spring 2006, spring 2007), this amounted to 222 student test scores (0.2%). Of the missing scores, 92 were Reading First and 130 were non-Reading First, which corresponds to 0.1% of Reading First test scores and 0.2% of non-Reading First test scores.

[7] Percentages are calculated by dividing the analytic sample numbers by the number of students tested.

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.

TOSWRF (Spring 2007 Only, Grade 1 Only)

RFIS Schools

RF:	125
Non-RF:	123
Total:	248

All Students in Grade 1

RF:	9,415
Non-RF:	9,443
Total:	18,858

Sample Students in Grade 1[1]

RF:	6,502
Non-RF:	6,322
Total:	12,824

Students Not Eligible to be Tested[2]

RF:	133
Non-RF:	152
Total:	285

Students Eligible to be Tested

RF:	6,369
Non-RF:	6,170
Total:	12,539

Eligible Students Not Tested[2]

RF:	849
Non-RF:	898
Total:	1,747

Students Tested

RF:	5,520 (87%)
Non-RF:	5,272 (85%)
Total:	10,792 (86%)

Analytic Sample[4, 5]

RF:	5,287 (96%)
Non-RF:	5,040 (96%)
Total:	10,327 (96%)

Notes:

[1] In 2006-07, the study administered the TOSWRF to students in those grade 1 classrooms where observations were conducted. Students in classrooms that were self-contained special education classrooms or classrooms in which instruction occurred in languages other than English were ineligible to be tested.

[2] Students were not eligible for assessments if they were excluded from testing in accordance with their own school or district policies (generally because they received instruction primarily in a language other than English), and/or they needed special accommodations beyond those that could be provided through additional time in a group administered testing situation.

[3] Eligible students were not tested if absent at the time the test was given and could not be rescheduled, they had transferred out, they had refused to take the test, or the RFIS did not have consent for them to participate in the study.

[4] Student test scores were excluded from the TOSWRF analytic sample if birth dates were missing, or out of range, or if students did not follow test instructions. Of the 465 test scores (5%) excluded from the analytic sample, 233 were Reading First and 232 were non-Reading First, which corresponds to 4% of Reading First scores and 4% of non-Reading First scores in the analytic sample.

[5] Percentages are calculated by dividing the analytic sample numbers by the number of students tested.

SOURCES: RFIS TOSWRF administration, spring 2007.

In the 2004-05 school year, the study team endeavored to test all students within grades 1, 2, and 3 in the participating schools using the SAT 10. However, the fact that some schools had as many as 10 or 12 classrooms per grade level led the study team to sample classrooms within grades in subsequent testing, such that the team assessed an average of three classrooms per grade per school in spring 2006 and spring 2007 (in spring 2007, this approach was also used for the TOSWRF). Note that the RFIS tested all students as required by local policy in those schools that routinely administered the SAT 10 reading comprehension as part of state- or district-standardized assessment. In all sites, SAT 10 and TOSWRF testing procedures were equivalent for Reading First and for comparison schools. Some sites required classroom teachers to administer tests; other sites relied upon RFIS staff to administer assessments. In the latter sites, the RFIS Team worked with district officials to carry out testing in accordance with local guidelines.

In the spring of 2007, the TOSWRF was administered to 10,792 first grade students, with an average response rate of 86% across Reading First and non-Reading First classrooms. Analyses were conducted using standard scores, which are reported in tables. Conversions to percentiles and grade equivalents are also provided in tables for reference, however, the publisher of the TOSWRF discourages the use of grade equivalents due to their unreliability.

Part 6: Data Collection Instruments

Exhibit C.17: Reading Coach Survey

OMB Number: 1850-0797
Expiration Date: 10/31/07

Reading First Impact Study
Reading Coach Survey

Abt Associates has been commissioned by the Institute of Education Sciences (IES) at the U.S. Department of Education to conduct an independent national evaluation of the Reading First program. Your school is one of only 250 elementary schools that have been selected to participate in this study, so your participation is extremely important. The study includes two kinds of schools: half of which have received Reading First funding, and half of which have not received Reading First funding. We know that teaching children how to read is important in every single school, and that is why we are asking teachers, staff, and administrators in both kinds of schools to describe the reading instruction in their schools and classrooms. Your responses will help inform the U.S. Department of Education, Congress, policymakers, practitioners, and researchers about how reading instruction is implemented in schools and what strategies schools use to provide high-quality, evidence-based reading instruction in first, second, and third grades.

The survey will take you approximately 30 minutes to complete. All responses to the survey will be kept confidential. All individual identifying information will be used only by persons on the research team. Information such as school location (state), participants' general job titles, grades they teach, and gender will be included in the study data files to be submitted to the Department of Education. However, participants' names will be stripped from all analysis data files and data files to be submitted to the Department of Education. We will not report any data about individual classrooms—all information will be reported at the grade and school levels. Neither your school nor your district will have access to any of the completed surveys at any time.

Thank you for your cooperation with this survey!

Instructions

Unless otherwise noted, your responses should reflect your experiences during the <u>2006-2007 school year</u> in the school to which this survey was sent.

- Please complete all questions; each question includes directions for recording your answer.

- You are sometimes told to skip over some questions in the survey. When this happens, you will see an arrow with a note that tells you what question to answer like this:

    ❑₁ Yes
    ❑₂ No   Skip to E4

- If you have any questions about how to complete the survey, please call: 1-866-421-6982. This is a free call and will connect you with our expert interviewers who can assist you.

A. Your Background and Experience

> A **reading coach** is a staff member whose primary role is to provide ongoing training and support to school staff in the delivery of effective reading instruction.

	Enter # below…
A1. Including this year, for how many years have you been the K–3 reading coach for this school? (If less than one year, enter 1.)	_____ years 10-11/
A2. Including this year, for how many years have you worked at this school in any capacity? (If less than one year, enter 1.)	_____ years 12-13/
A3. Including this year, how many years of classroom experience do you have, as either a teacher and/or reading coach? (If less than one year, enter 1.)	
a. Number of years experience as a reading coach	_____ years 14-15/
b. Number of years experience as a teacher	_____ years 16-17/

B. Coach Responsibilities

	Enter # below…
B1. This school year, for how many schools do you serve as the reading coach (including this school)?	_____ Schools 18-19/
B2. This school year, for how many teachers do you serve as the reading coach (include all teachers in all schools)?	_____ Teachers 20-21/
B3. Approximately what percentage of your time do you spend as the K–3 reading coach for this school?	_____ % 22-24/

B4.   How central is each of the following activities to your work this year (since July 1st) at this school?

Please rate the activity a "1" if you do not do the activity or if it is not at all central to your role as the literacy coach. Rate the activity a "5" if it is absolutely central or critical to your work.

Activity	Do not do or not at all central		Somewhat central		Absolutely central
	Check only one box for each item…				
a. Administering/coordinating reading assessments	☐1	☐2	☐3	☐4	☐5 25/
b. Compiling reading assessment data for teachers	☐1	☐2	☐3	☐4	☐5 26/
c. Facilitating grade level meetings	☐1	☐2	☐3	☐4	☐5 27/
d. Participating in school leadership team meetings	☐1	☐2	☐3	☐4	☐5 28/
e. Facilitating or coordinating family literacy activities	☐1	☐2	☐3	☐4	☐5 29/
f. Ordering/managing reading instruction materials	☐1	☐2	☐3	☐4	☐5 30/
g. Participating in professional development provided by the district, state or other consultants	☐1	☐2	☐3	☐4	☐5 31/
h. Providing sub time for teachers to observe other more experienced teachers	☐1	☐2	☐3	☐4	☐5 32/
i. Providing direct reading instruction to students	☐1	☐2	☐3	☐4	☐5 33/
j. Providing training/professional development in reading materials, strategies, and assessments	☐1	☐2	☐3	☐4	☐5 34/
k. Coaching staff on a range of topics (note: specific coaching activities are asked about in the next item)	☐1	☐2	☐3	☐4	☐5 35/
l. Organizing professional development for K-3 teachers	☐1	☐2	☐3	☐4	☐5 36/
m. Coordinating activities and meetings between classroom and special education teachers	☐1	☐2	☐3	☐4	☐5 37/
n. Coordinating activities and meetings between classroom teachers and English Language Learner (ELL) staff	☐1	☐2	☐3	☐4	☐5 38/
→ Check if no ELLs ☐1 39/					
o. Other (Please specify): _____ 41-42/	☐1	☐2	☐3	☐4	☐5 40/

English language learner (ELL) indicates a student who is in the process of acquiring English and **has a first language other than English**. Other common related terms include language minority or limited English proficient (LEP) students, students in English as a second language (ESL), or students in classes for English for speakers of other languages (ESOL).

B5.   When you **coach** K-3 staff, how central has each of the following activities been to your work this year (since July 1st)?

Please rate the activity a "1" if you do not do the activity or if it is not at all central to your role as the literacy coach. Rate the activity a "5" if it is absolutely central or critical to your work.

Coaching Activity	Do not do or not at all central		Somewhat central		Absolutely central
	Check only one box for each item…				
a. Giving demonstration lessons using core or supplemental materials	☐₁	☐₂	☐₃	☐₄	☐₅ 43/
b. Assisting teachers in using the core program	☐₁	☐₂	☐₃	☐₄	☐₅ 44/
c. Observing and providing feedback to teachers	☐₁	☐₂	☐₃	☐₄	☐₅ 45/
d. Assisting teachers in forming instructional groups	☐₁	☐₂	☐₃	☐₄	☐₅ 46/
e. Assisting teachers in designing strategies for addressing the needs of struggling readers	☐₁	☐₂	☐₃	☐₄	☐₅ 47/
f. Assisting teachers with monitoring the effectiveness of strategies addressing the needs of struggling readers	☐₁	☐₂	☐₃	☐₄	☐₅ 48/
g. Giving demonstrations on assessment administration and scoring	☐₁	☐₂	☐₃	☐₄	☐₅ 49/
h. Planning reading instruction with teachers	☐₁	☐₂	☐₃	☐₄	☐₅ 50/
i. Reviewing teachers' lesson plans and providing feedback	☐₁	☐₂	☐₃	☐₄	☐₅ 51/
j. Assisting teachers in interpreting assessment results	☐₁	☐₂	☐₃	☐₄	☐₅ 52/
k. Assisting teachers in designing strategies for addressing the needs of special education students	☐₁	☐₂	☐₃	☐₄	☐₅ 53/
l. Assisting teachers in designing strategies for addressing the needs of ELLs (see page 2 for definition of ELL)	☐₁	☐₂	☐₃	☐₄	☐₅ 54/
⟶ Check if no ELLs ☐₁ 55/					

C. Reading Instructional Materials

C1. Which **core reading program** is being used to teach reading in each of grades K-3 **at this school**? We have provided a partial list of core reading programs below. Please check the core reading program used in each grade. If the core reading program you use is **not** included below, please check "other," and write in the name of the program and the publisher and/or developer (if applicable).

Note: Inclusion of a core reading program on the list below does not constitute an endorsement

by the U.S. Department of Education.

A **core reading program** is one that provides a comprehensive program of instruction on a daily basis in all aspects of reading.

Publisher/ Developer	Core reading program	Check one per grade			
		K 56-57/	1 58-59/	2 60-61/	3 62-63/
Addison Wesley	Unspecified or other (Please specify): _____ 64-65/	$\square_1$	$\square_1$	$\square_1$	$\square_1$
Harcourt	Collections	$\square_2$	$\square_2$	$\square_2$	$\square_2$
	Rigby Reading	$\square_3$	$\square_3$	$\square_3$	$\square_3$
	Signatures	$\square_4$	$\square_4$	$\square_4$	$\square_4$
	Trophies	$\square_5$	$\square_5$	$\square_5$	$\square_5$
	Unspecified or other (Please specify): _____ 66-67/	$\square_6$	$\square_6$	$\square_6$	$\square_6$
Heinemann	Fountas Pinnel units of study	$\square_7$	$\square_7$	$\square_7$	$\square_7$
	Unspecified or other (Please specify): _____ 68-69/	$\square_8$	$\square_8$	$\square_8$	$\square_8$
Houghton Mifflin	Horizons	$\square_9$	$\square_9$	$\square_9$	$\square_9$
	Invitation to Literacy	$\square_{10}$	$\square_{10}$	$\square_{10}$	$\square_{10}$
	Lectura	$\square_{11}$	$\square_{11}$	$\square_{11}$	$\square_{11}$
	Legacy of Literacy	$\square_{12}$	$\square_{12}$	$\square_{12}$	$\square_{12}$
	Nation's Choice	$\square_{13}$	$\square_{13}$	$\square_{13}$	$\square_{13}$
	Reading	$\square_{14}$	$\square_{14}$	$\square_{14}$	$\square_{14}$
	State Specific Edition	$\square_{15}$	$\square_{15}$	$\square_{15}$	$\square_{15}$
	Unspecified or other (Please specify): _____ 70-71/	$\square_{16}$	$\square_{16}$	$\square_{16}$	$\square_{16}$
McGraw-Hill	Open Court	$\square_{17}$	$\square_{17}$	$\square_{17}$	$\square_{17}$
	Reading	$\square_{18}$	$\square_{18}$	$\square_{18}$	$\square_{18}$
	Reading Mastery	$\square_{19}$	$\square_{19}$	$\square_{19}$	$\square_{19}$
	Spotlight on Literacy	$\square_{20}$	$\square_{20}$	$\square_{20}$	$\square_{20}$
	Unspecified or other (Please specify): _____ 72-73/	$\square_{21}$	$\square_{21}$	$\square_{21}$	$\square_{21}$
Saxon	Saxon Phonics	$\square_{22}$	$\square_{22}$	$\square_{22}$	$\square_{22}$
	Unspecified or other (Please specify): _____ 74-75/	$\square_{23}$	$\square_{23}$	$\square_{23}$	$\square_{23}$

C1. CONTINUED. Which **core reading program** is being used to teach reading in each of grades K-3 **at this school**? We have provided a partial list of core reading programs below. Please check the core reading program used in each grade. If the core reading program you use is **not** included below, please check "other," and write in the name of the program and the publisher and/or developer (if applicable).

Note: Inclusion of a core reading program on the list below does not constitute an endorsement

by the U.S. Department of Education.

Publisher/ Developer	Core reading program	Check one per grade			
		K	1	2	3
Scholastic	Literacy Place	❑24	❑24	❑24	❑24
	Unspecified or other (Please specify): _____ _76-77/	❑25	❑25	❑25	❑25
Scott Foresman	Literacy Works	❑26	❑26	❑26	❑26
	Reading	❑27	❑27	❑27	❑27
	State Specific Edition	❑28	❑28	❑28	❑28
	Unspecified or other (Please specify): _____ _78-79/	❑29	❑29	❑29	❑29
Sopris	Read Well	❑30	❑30	❑30	❑30
	Unspecified or other (Please specify): _____ _80-81/	❑31	❑31	❑31	❑31
Success for All	Success for All	❑32	❑32	❑32	❑32
	Unspecified or other (Please specify): _____ _82-83/	❑33	❑33	❑33	❑33
Voyager	Universal Literacy	❑34	❑34	❑34	❑34
	Unspecified or other (Please specify): _____ _84-85/	❑35	❑35	❑35	❑35
Wright Group	Unspecified or other (Please specify): _____ _86-87/	❑36	❑36	❑36	❑36
Other	Publisher/Developer: _88-89/ Program Title: _90-91/	❑37	❑37	❑37	❑37
	Publisher/Developer: _92-93/ Program Title: _94-95/	❑38	❑38	❑38	❑38
I use a core reading program developed by teachers or other school personnel		❑39	❑39	❑39	❑39
I do not use a core reading program		❑40	❑40	❑40	❑40

C2.   Has your school made any of the following **changes** to your reading program that took effect at the beginning of the current school year (2006-2007)?

Indicate 'yes' or 'no' for each item …

	Yes	No
a. Adopted a new core reading program	☐1	☐2 96/
b. Added a new intervention program for struggling readers	☐1	☐2 97/
c. Added new supplementary materials	☐1	☐2 98/
d. Added new materials for ELLs (see definition of ELL on page 2)	☐1	☐2 99/
└───────→ Check if no ELLs ☐1 100/		
e. Adopted new reading assessments	☐1	☐2 101/

D.   Instructional Time

D1.   Please indicate for which grades your school has a scheduled reading block.

> A **reading block** is the time period that is formally scheduled for teaching reading.

If <u>yes</u>, please indicate for how many minutes the reading block is scheduled.
Does your school have a reading block in:

	Yes	Scheduled number of minutes	No	
a. Kindergarten	☐1	───→ ☐	☐2	102/ 103-105/
b. First grade	☐1	───→ ☐	☐2	106/ 107-109/
c. Second grade	☐1	───→ ☐	☐2	110/ 111-113/
d. Third grade	☐1	───→ ☐	☐2	114/ 115-117/

E. Reading Interventions for Struggling Readers

E1. What methods has your school used to meet the needs of at-risk or **struggling readers**? For each method listed below, please check whether or not you use the method at your school.

> A **reading intervention** is a program designed **for struggling readers** to be used only with struggling readers in addition to the core-reading program.

Methods for meeting needs of struggling readers		Check one for each item …	
		Use this method	Not used
Materials	a. Use separate program materials in interventions	☐₁	☐₂ 118/
	b. Use core reading program with supplemental materials	☐₁	☐₂ 119/
	c. Use core reading program only	☐₁	☐₂ 120/
	d. Use reading materials written in ELLs' home language (see definition of ELL on page 2)	☐₁	☐₂ 121/
	e. Use alternative materials designed for ELLs	☐₁	☐₂ 122/
	Check if no ELLs ☐₁ 123/		
Staff activities	f. A certified reading specialist provides additional direct instruction to struggling readers, individually or in small groups.	☐₁	☐₂ 124/
	g. The classroom teacher provides additional direct instruction to struggling readers, individually or in small groups.	☐₁	☐₂ 125/
	h. The classroom teacher provides additional opportunities for reading skill practice for struggling readers (e.g., partner reading, peer tutors, audio tapes, computer programs).	☐₁	☐₂ 126/
	i. A certified specialist provides recommendations to classroom teachers on accommodations for struggling readers. (Indicate which type of specialist.)		
	A special education teacher	☐₁	☐₂ 127/
	A bilingual/ESL teacher	☐₁	☐₂ 128/
	Other (Please specify): _____ 130-131/	☐₁	☐₂ 129/
	j. Trained aides or volunteers work with students under the direction of the classroom teacher during the school day.	☐₁	☐₂ 132/
	k. Trained aides or volunteers work with students in a before or after school program.	☐₁	☐₂ 133/
	l. Untrained aides or volunteers work with students under the direction of the classroom teacher during the school day.	☐₁	☐₂ 134/
	m. Untrained aides or volunteers work with students in a before or after school program.	☐₁	☐₂ 135/

F. Professional Development for Reading Coaches

F1. Below is a list of professional development topics for **reading coaches** in which you may have participated.
- In column A, identify any topics that were addressed in **reading coaches'** professional development activities **during the current school year, including summer 2006**.
- In column B, please identify the topics in which you need more professional development, whether or not this school's professional development activities have covered these topics.
- Please <u>check all that apply</u> in columns A and B.

Topics	CHECK ALL THAT APPLY	
	A. Topics addressed in professional development for reading coaches	**B. Topics in which you need more professional development**
a. How to use reading assessment data to guide instruction.	☐₁ 136/	☐₂ 137/
b. What are the types of assessments: screening, diagnostic, progress monitoring, and outcome?	☐₃ 138/	☐₄ 139/
c. How to use assessment data to form instructional groups.	☐₁ 140/	☐₂ 141/
d. How to provide constructive feedback to teachers.	☐₃ 142/	☐₄ 143/
e. How to establish credibility with teachers.	☐₁ 144/	☐₂ 145/
f. Essential components of scientifically based reading instruction.	☐₃ 146/	☐₄ 147/
g. What is the role of the reading coach in fostering change?	☐₁ 148/	☐₂ 149/
h. How to plan instructional interventions for struggling students.	☐₃ 150/	☐₄ 151/
i. Classroom management within the literacy block time.	☐₁ 152/	☐₂ 153/
j. How to conduct effective grade level meetings.	☐₃ 154/	☐₄ 155/
k. How to help teachers identify appropriate instructional materials.	☐₁ 156/	☐₂ 157/
l. How to help teachers make reading instruction systematic and explicit.	☐₃ 158/	☐₄ 159/
m. How to conduct demonstration lessons.	☐₁ 160/	☐₂ 161/
n. How to conduct classroom observations.	☐₃ 162/	☐₄ 163/
o. How to provide onsite professional development.	☐₁ 164/	☐₂ 165/
p. How to provide instructional supports for ELL students learning to read (see definition of ELL on page 2).	☐₃ 166/	☐₄ 167/
Check if no ELLs ☐₁ 168/		

G. Reading Instruction

G1. This item asks you to describe your school using the statements below. Please read each statement, and indicate whether the statement is a good description of your school on a scale from a "Very inaccurate" description of your school to a "Very accurate" description of your school.

In this school …	Very inaccurate				Very accurate
a. K-3 teachers are knowledgeable about scientifically based reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 169/
b. K-3 teachers are motivated to improve reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 170/
c. Reading instruction in K-3 classrooms is aligned with the state reading/language arts content standards.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 171/
d. There is a school-wide focus on reading and language arts.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 172/
e. K-3 teachers are experienced with the core reading program.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 173/
f. K-3 teachers are experienced with supplemental reading materials.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 174/
g. K-3 teachers are experienced with reading intervention materials and strategies.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 175/
h. K-3 classrooms have ample, high quality instructional materials.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 176/
i. Teachers use a variety of instructional materials to fill in gaps in the core program.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 177/
j. The core reading program is aligned with scientifically based reading research.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 178/
k. Supplemental reading materials are aligned with scientifically based reading research.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 179/
l. Reading intervention materials are aligned with scientifically based reading research.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 180/
m. The reading coach has the support of the school principal.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 181/
n. K-3 teachers seek the assistance of the reading coach to improve their reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 182/
o. Sufficient time during the school day is allotted for reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 183/
p. Sufficient time during the school day is allotted for teacher planning.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 184/
q. K-3 teachers collaborate and plan for reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 185/
r. Sufficient time during the school day is allotted for professional development.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 186/
s. Reading assessments are used to screen students for reading difficulties.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 187/
t. Diagnostic assessments are used to identify strengths and weaknesses of struggling readers.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 188/
u. Reading assessments are used to monitor student progress.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 189/
v. Assessment data are used to group students for instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 190/
w. Assessment data are used to guide and/or modify instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 191/
x. The district provides direction concerning reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 192/
y. The state provides direction concerning reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 193/
z. K-3 teachers make an effort to involve parents in their children's reading instruction.	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$ 194/

Check one box for each item…

H.   Out of School Reading

H1.   What school-wide strategies do you use to encourage students to read outside of school?

Strategies to encourage students to read outside of school	Check only one box for each item …	
	YES	**NO**
a. School-wide book-reading contests	☐₁	☐₂ 384/
b. Grade-specific contests	☐₁	☐₂ 385/
c. Book giveaways with parents about reading resources	☐₁	☐₂ 386/
d. Use reading materials written in students' home language	☐₁	☐₂ 387/
e. Relationship with community libraries	☐₁	☐₂ 388/
f. School-wide book clubs	☐₁	☐₂ 389/
g. Partnering students who live near each other to read to each other out of school	☐₁	☐₂ 390/
h. Principal, reading coach or librarian awards/prizes for individual readers	☐₁	☐₂ 391/
i. School-wide book celebrations	☐₁	☐₂ 392/
j. School-wide rewards for reaching reading milestones	☐₁	☐₂ 393/

Thank you for your cooperation and for taking the time to answer these questions. Please place the completed survey in the enclosed envelope, seal the envelope, and return to your school's Reading First Impact Study liaison.

If you have any questions about the survey, please call 1-866-421-6982 and leave a message, or send an e-mail to reading_impact@abtassoc.com

Abt ID // barcode	1-6/

Abt ID / barcode here

Reading First Impact Study
Teacher Survey

Abt Associates has been commissioned by the Institute of Education Sciences (IES) at the U.S. Department of Education to conduct an independent national evaluation of the Reading First program. Your school is one of only 250 elementary schools that have been selected to participate in this study, so your participation is extremely important. The study includes two kinds of schools: half of which have received Reading First funding, and half of which have not received Reading First funding. We know that teaching children how to read is important in every single school, and that is why we are asking teachers, staff, and administrators in both kinds of schools to describe the reading instruction in their schools and classrooms. Your responses will help inform the U.S. Department of Education, Congress, policymakers, practitioners, and researchers about how reading instruction is implemented in schools and what strategies schools use to provide high-quality, evidence-based reading instruction in first, second, and third grades.

Additional Information

The survey will take you approximately 30 minutes to complete. All responses to the survey will be kept confidential. All individual identifying information will be used only by persons on the research team. Information such as school location (state), participants' general job titles, grades they teach, and gender will be included in the study data files to be submitted to the Department of Education. However, participants' names will be stripped from all analysis data files and data files to be submitted to the Department of Education. We will not report any data about individual classrooms—all information will be reported at the grade and school levels. Neither your school nor your district will have access to any of the completed surveys at any time.

Thank you for your cooperation with this survey!

Instructions

Unless otherwise noted, your responses should reflect your experiences during the <u>2006-2007 school year</u> in the school to which this survey was sent.

- Please complete all questions; each question includes directions for recording your answer.

- You are sometimes told to skip over some questions in the survey. When this happens, you will see an arrow with a note that tells you what question to answer like this:

    ❑$_1$ Yes
    ❑$_2$ No   Skip to E4

- If you have any questions about how to complete the survey, please call: 1-866-421-6982. This is a free call and will connect you with our expert interviewers who can assist you.

A.   Your Background and Experience

A1.   Including this year, how long have you been a teacher? (If less than one year, please enter '1')

	Enter # of years below…
a. Total number of years as a teacher	_____ **Years** 10-11/
b. Number of years teaching at this school	_____ **Years** 12-13/

A2.   What grade(s) are you currently teaching? (Check all that apply)

❑₁ Kindergarten   14/
❑₂ First grade   15/
❑₃ Second grade   16/
❑₄ Third grade   17/
❑₉₅ Other (Please specify): _____   18-19/ 20-21/
                                                                                                              22-23/ 24-25/

A3.   How well do you feel your **pre-service teacher training** prepared you to teach each of the following dimensions of reading?

> **Pre-service teacher training** refers to training you received before you became certified and began teaching. For those who began their teacher career through an alternative certification or emergency certification program, and began teaching before they were certified, pre-service teacher training refers to the training you received to become fully certified.

Please choose a '1' if you were 'not at all prepared' to teach the dimension and a '5' if you were 'extremely well prepared.'

	Check only one box for each item…				
	Not at all prepared		Somewhat well prepared		Extremely well prepared
a. Phonemic awareness	❑₁	❑₂	❑₃	❑₄	❑₅ 26/
b. Decoding	❑₁	❑₂	❑₃	❑₄	❑₅ 27/
c. Vocabulary	❑₁	❑₂	❑₃	❑₄	❑₅ 28/
d. Comprehension	❑₁	❑₂	❑₃	❑₄	❑₅ 29/
e. Fluency building	❑₁	❑₂	❑₃	❑₄	❑₅ 30/

B. Instruction and Assessment in Reading

B1. **Last week**, approximately how many minutes per day did you devote to reading instruction? Include only reading instruction and not other language arts such as writing, spelling. Fill in the chart for each day last week with your best estimate of the number of minutes…

Monday	Tuesday	Wednesday	Thursday	Friday
# _____ min. / day	# _____ min. / day	# _____ min. / day	# _____ min. / day	# _____ min. / day
31-33/	34-36/	37-39/	40-42/	43-45/

B2. Has the average number of minutes you spend each day this year teaching reading decreased, remained the same, or increased from last year (2005-2006)? Please check one.

❑₁ **I did not teach reading last year**   46/

❑₂ Decreased

❑₃ Remained the same

❑₄ Increased

B3. How often **during this school year** is time regularly scheduled and formally set aside during the school day for Grade 2-3 teachers to:

	Check only one box for each item …					
	Not at all	1-4 times	5-8 times	Once a month	Once a week or more	Occurs only informally, as needed
a. Collaborate on reading lesson planning and instruction.	❑₁	❑₂	❑₃	❑₄	❑₅	❑₆ 47/
b. Observe reading instruction in other classrooms.	❑₁	❑₂	❑₃	❑₄	❑₅	❑₆ 48/
c. Use assessment data to plan instruction.	❑₁	❑₂	❑₃	❑₄	❑₅	❑₆ 49/
d. Participate in coaching with or be coached about reading by other teachers.	❑₁	❑₂	❑₃	❑₄	❑₅	❑₆ 50/
e. Be coached about my reading instruction by a reading coach (see below).	❑₁	❑₂	❑₃	❑₄	❑₅	❑₆ 51/
f. Coordinate reading interventions for struggling readers with special education staff.	❑₁	❑₂	❑₃	❑₄	❑₅	❑₆ 52/
g. Coordinate reading interventions for struggling readers with English language learner (ELL) staff (see below).	❑₁	❑₂	❑₃	❑₄	❑₅	❑₆ 53/

     ⟶ Check if no ELLs ❑₁   54/

A **reading coach** is a staff member whose primary role is to **provide ongoing training and support to classroom teachers** in the delivery of effective reading instruction. This assistance may include planning instruction, providing demonstration lessons, observing and providing feedback, using assessment results to guide instruction, etc.

English language learner (ELL) indicates a student who is in the process of acquiring English and **has a first language other than English.** Other common related terms include language minority or limited English proficient (LEP) students, students in English as a second language (ESL), or students in classes for English for speakers of other languages (ESOL).

B4. Please describe your use of the following <u>reading instructional activities</u> **this year**.
- **Check column A ONLY** if the instructional activity is one that you use frequently when you teach reading or one on which you rely heavily in your reading instruction.
- **Check column B** if you use the instructional activity, but it is a small part of your teaching, and not one you use frequently. It might be an activity that you use if there is time, but it is not one on which you rely heavily for your reading instruction.
- **Check column C** if the activity is not one you use in your reading instruction.

		Check only one box for each item…		
		A **Central to my reading instruction**	**B** **Small part of my reading instruction**	**C** **Not part of my reading instruction**
Reading Text	a. Students read texts that are easy to decode.	☐₁	☐₂	☐₃ 55/
	b. Students read to locate information.	☐₁	☐₂	☐₃ 56/
	c. Students read aloud unfamiliar text.	☐₁	☐₂	☐₃ 57/
	d. Students reread familiar stories.	☐₁	☐₂	☐₃ 58/
	e. Students read aloud together.	☐₁	☐₂	☐₃ 59/
	f. Students read silently.	☐₁	☐₂	☐₃ 60/
	g. I listen to students read aloud without correcting errors.	☐₁	☐₂	☐₃ 61/
	h. I listen to students read aloud and correct errors immediately.	☐₁	☐₂	☐₃ 62/
	i. Students confirm or revise predictions after reading.	☐₁	☐₂	☐₃ 63/
	j. Students generate their own questions about text material.	☐₁	☐₂	☐₃ 64/
	k. Students identify their comprehension break-downs and use fix-up strategies with a partner.	☐₁	☐₂	☐₃ 65/
	l. Students orally summarize main events in stories and informational texts.	☐₁	☐₂	☐₃ 66/
	m. Students use graphic and semantic organizers to track information.	☐₁	☐₂	☐₃ 67/
Work with sounds and words	n. Students decode multi-syllabic words in isolation.	☐₁	☐₂	☐₃ 84/
	o. I teach decoding skills while reading stories.	☐₁	☐₂	☐₃ 85/
	p. Students memorize sight words.	☐₁	☐₂	☐₃ 86/
	q. Students read irregularly spelled words and non-words.	☐₁	☐₂	☐₃ 87/
	r. Students practice reading high frequency words for automaticity.	☐₁	☐₂	☐₃ 88/
	s. Students use knowledge of root words, prefixes, and suffixes to decode new words.	☐₁	☐₂	☐₃ 89/
	t. Students work with prefixes and suffixes to change the meaning of words.	☐₁	☐₂	☐₃ 90/
	u. I stop students while reading and have them self-correct misidentified words.	☐₁	☐₂	☐₃ 91/
	v. Students use context clues to identify unknown words.	☐₁	☐₂	☐₃ 92/
	w. Students practice writing words as separate syllables.	☐₁	☐₂	☐₃ 93/

68-83/BLANK

94-107/BLANK

B4. CONTINUED. Please describe your use of the following <u>reading instructional activities</u> **this year**.

- **Check column A ONLY** if the instructional activity is one that you use frequently when you teach reading or one on which you rely heavily in your reading instruction.
- **Check column B** if you use the instructional activity, but it is a small part of your teaching, and not one you use frequently. It might be an activity that you use if there is time, but it is not one on which you rely heavily for your reading instruction.
- **Check column C** if the activity is not one you use in your reading instruction.

		Check only one box for each item…		
		A **<u>Central</u> to my reading instruction**	**B** **<u>Small</u> part of my reading instruction**	**C** **<u>Not</u> Part of my reading instruction**
Other Techniques	x. Students learn vocabulary through study of antonyms, synonyms, and homonyms.	☐₁	☐₂	☐₃ 108/
	y. Students learn vocabulary through study of word categories.	☐₁	☐₂	☐₃ 109/
	z. Students write vocabulary words in sentences.	☐₁	☐₂	☐₃ 110/
	aa. Students use dictionaries to find word meanings.	☐₁	☐₂	☐₃ 111/
	bb. I discuss new and unusual words before reading.	☐₁	☐₂	☐₃ 112/
	cc. Students answer questions in writing after reading stories.	☐₁	☐₂	☐₃ 113/
	dd. Students identify story structure and elements.	☐₁	☐₂	☐₃ 114/
	ee. Students read stories they have written to others.	☐₁	☐₂	☐₃ 115/
	ff. Students participate in literature extensions, (e.g., book talks, plays, readers' theaters).	☐₁	☐₂	☐₃ 116/
	gg. Students select books from the library for independent reading.	☐₁	☐₂	☐₃ 117/
	hh. Students are given time to read on their own for enjoyment.	☐₁	☐₂	☐₃ 118/
	ii. Class creates group stories.	☐₁	☐₂	☐₃ 119/

120-134/BLANK

B5.   Please describe your use of the following teaching strategies and materials **this year**.
- **Check column A ONLY** if the item is one that you use frequently or one on which you rely heavily in your reading instruction.
- **Check column B** if you use the item, but it is a small part of your teaching, and not one you use frequently. It may be an approach you use if there is time, but it is not one on which you rely heavily.
- **Check column C** if the item is not one you use in your reading instruction.

		Check only one box for each item…		
		A **Central** to my reading instruction	**B** **Small** part of my reading instruction	**C** **Not** Part of my reading instruction
Instruction	a. Provide time in reading block for skill practice on own.	☐₁	☐₂	☐₃ 135/
	b. Provide materials for at-home practice of skills introduced in class.	☐₁	☐₂	☐₃ 136/
	c. Provide extra reading instructional time for struggling students.	☐₁	☐₂	☐₃ 137/
	d. Include writing opportunities in reading instruction.	☐₁	☐₂	☐₃ 138/
	e. Build spelling practice into reading instruction.	☐₁	☐₂	☐₃ 139/
	f. Develop reading skills through science and social studies.	☐₁	☐₂	☐₃ 140/
Grouping	g. Teach whole class reading lessons.	☐₁	☐₂	☐₃ 141/
	h. Work one-to-one with students on reading.	☐₁	☐₂	☐₃ 142/
	i. Work with small groups of students.	☐₁	☐₂	☐₃ 143/
	j. Group students based on skill levels.	☐₁	☐₂	☐₃ 144/
	k. Group students based on mixed abilities (cooperative groups).	☐₁	☐₂	☐₃ 145/
	l. Pair strong readers with those with weaker skills.	☐₁	☐₂	☐₃ 146/
Reading materials	m. Use core reading series.	☐₁	☐₂	☐₃ 147/
	n. Use supplementary reading materials for instruction in the following areas:			
	1. Phonemic awareness	☐₁	☐₂	☐₃ 148/
	2. Phonics	☐₁	☐₂	☐₃ 149/
	3. Fluency	☐₁	☐₂	☐₃ 150/
	4. Vocabulary	☐₁	☐₂	☐₃ 151/
	5. Comprehension	☐₁	☐₂	☐₃ 152/
	o. Use children's trade books.	☐₁	☐₂	☐₃ 153/
	p. Use books that are easy to decode.	☐₁	☐₂	☐₃ 154/
	q. Use books with patterned predictable language.	☐₁	☐₂	☐₃ 155/
	r. Use separate intervention materials for some students.	☐₁	☐₂	☐₃ 156/
	s. Use reading software/technology.	☐₁	☐₂	☐₃ 157/
	t. Use teacher-made materials.	☐₁	☐₂	☐₃ 158/

Supplementary Reading Materials provide additional instruction in a targeted area of reading to **all** students. **Do not** include materials that are used only with struggling readers. Include teacher-made materials, if applicable.

B5. CONTINUED. Please describe your use of the following teaching strategies and materials **this year**.

- **Check column A ONLY** if the item is one that you use frequently or one on which you rely heavily in your reading instruction.
- **Check column B** if you use the item, but it is a small part of your teaching, and not one you use frequently. It may be an approach you use if there is time, but it is not one on which you rely heavily.
- **Check column C** if the item is not one you use in your reading instruction.

		Check only one box for each item…		
		A **Central to my reading instruction**	**B** **Small part of my reading instruction**	**C** **Not Part of my reading instruction**
Assessments	u. Use test results to organize instructional groups.	□₁	□₂	□₃ 159/
	v. Use informal reading inventories.	□₁	□₂	□₃ 160/
	w. Use tests to determine progress on skills.	□₁	□₂	□₃ 161/
	x. Use tests to determine who can benefit from the core reading series.	□₁	□₂	□₃ 162/
	y. Use diagnostic tests to identify students who need reading intervention services.	□₁	□₂	□₃ 163/
	z. Use screening tests to identify students who need a supplementary reading program.	□₁	□₂	□₃ 164/
	aa. Conduct miscue analysis, analyzing errors students make while reading aloud.	□₁	□₂	□₃ 165/

B6. What specific formal or informal assessments do you primarily use for placing and /or grouping students, determining student mastery of skills, and identifying core deficits of struggling students? If you use more than one assessment, please report only on the one that you use the most often. Please be as specific as possible when naming or describing the assessment(s).

- In <u>column A</u> enter the name of the primary assessment used for each purpose.
- In <u>column B</u> check the number of times the assessment is given during the school year.
- In <u>column C</u> check whether students are usually assessed individually, in small groups, or in a whole class.
- In <u>column D</u> check the average time that it takes to administer the assessment.
- In <u>column E</u> check whether the assessment is used for accountability purposes for the Reading First program, No Child Left Behind (NCLB), or another program. Please check all that apply in this column only.

A. Primary purposes and names of assessments	Check only one box per column for each item…			Check all that apply…
	B. Number of times given per school year *(Check one)*	C. Students are assessed… *(Check one)*	D. Average time it takes to administer assessment *(Check one)*	E. Accountability purposes *(Check all that apply)*
a. Placement and/or grouping students *(Check one):* 166/ ₁☐ Assessment: _____ _____ 167-168/ 169-170/ 171-172/ ₂☐ I do not use an assessment for this purpose	₁☐ 1 ₂☐ 2 ₃☐ 3 ₄☐ 4 ₅☐ 5 or more 173/	₁☐ Individually ₂☐ In small groups ₃☐ In whole class 174/	₁☐ 1-15 *minutes* ₂☐ 16-30 ₃☐ 31-45 ₄☐ 46-60 ₅☐ 61 or more 175/	₁☐ Reading First 176/ ₂☐ NCLB 177/ ₃☐ Other 178/
b. Determining student mastery of skills *(Check one):* 179/ ₁☐ Assessment: _____ _____ 180-181/ 182-183/ 184-185/ ₂☐ I do not use an assessment for this purpose	₁☐ 1 ₂☐ 2 ₃☐ 3 ₄☐ 4 ₅☐ 5 or more 186/	₁☐ Individually ₂☐ In small groups ₃☐ In whole class 187/	₁☐ 1-15 *minutes* ₂☐ 16-30 ₃☐ 31-45 ₄☐ 46-60 ₅☐ 61 or more 188/	₁☐ Reading First 189/ ₂☐ NCLB 190/ ₃☐ Other 191/
c. Identifying the core deficits of struggling students *(Check one):* 192/ ₁☐ Assessment: _____ _____ 193-194/ 195-196/ 197-198/ ₂☐ I do not use an assessment for this purpose	₁☐ 1 ₂☐ 2 ₃☐ 3 ₄☐ 4 ₅☐ 5 or more 199/	₁☐ Individually ₂☐ In small groups ₃☐ In whole class 200/	₁☐ 1-15 *minutes* ₂☐ 16-30 ₃☐ 31-45 ₄☐ 46-60 ₅☐ 61 or more 201/	₁☐ Reading First 202/ ₂☐ NCLB 203/ ₃☐ Other 204/

B7. What materials are used with **English Language Learners (ELLs)** to whom you teach reading (see definition of ELL on page 2)?
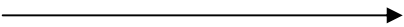(Check all that apply)

❑₁ **Do not teach ELLs →Skip to B8**   205/

❑₂ Core reading program materials in the native language of the ELL   206/

❑₃ ELL students use the same materials as other students   207/

❑₄ Core reading program materials, plus supplementary/intervention resources written in
     the ELL's native language   208/

❑₅ Core reading program materials, plus supplementary/intervention resources written in
     English especially for ELLs   209/

❑₆ Alternative core reading program materials in English geared toward the instructional
     level of the ELL   210/

B8. What additional supports have students who are struggling readers received **in the last month**? Check whether or not your students who are struggling readers received each of the supports during the past month.

		Check only one box for each item…		
Supports for Struggling Readers		**Received**	**Did not receive**	
a.	Diagnostic assessment to determine core deficits.	□₁	□₂	211/
b.	Extra practice in the classroom with phonemic awareness.	□₁	□₂	212/
c.	Extra practice in the classroom with decoding.	□₁	□₂	213/
d.	Extra practice in the classroom with fluency.	□₁	□₂	214/
e.	Extra practice in the classroom with comprehension.	□₁	□₂	215/
f.	Extra instructional time.	□₁	□₂	216/
g.	Placement in materials that supplement the core reading program.	□₁	□₂	217/
h.	Placement in different level of core reading program.	□₁	□₂	218/
i.	Placement in separate core reading program.	□₁	□₂	219/
j.	Placement in special intervention program.	□₁	□₂	220/
k.	Work with tutor on one-to-one basis.	□₁	□₂	221/
l.	Work with reading specialist on one-to-one basis.	□₁	□₂	222/
m.	Work with reading specialist in small group.	□₁	□₂	223/
n.	Work with more advanced peer.	□₁	□₂	224/
o.	Special materials for parents to provide practice.	□₁	□₂	225/

B9. What additional supports have students who are struggling readers *and* ELLs received **in the last month**? Check whether or not your students who are struggling readers and ELLs received each of the supports during the past month (see definition of ELL on page 2).

Check if no ELLs ⟶ □₁ Skip to C1   226/

	Check only one box for each item…		
Supports for Struggling Readers who are ELLs	**Received**	**Did not receive**	
a. English as a Second Language instruction.	□₁	□₂	227/
b. Provide reading instruction in home language.	□₁	□₂	228/
c. In classroom help in reading from ELL teacher.	□₁	□₂	229/

C. Professional Development in Reading for K-3 Teachers

C1.   During the current school year, including summer 2006, did you participate in any professional development activities in reading?

Yes   □₁
No   □₂   ⟶   Skip to C2

If yes, in how many of each of the following types of professional development activities **in reading** have you participated? Please count each activity only once. What is the total number of hours you spent in these activities?

First, write in the <u>number</u> of activities of each type in which you have been engaged. Then, write the total number of <u>hours</u> you spent in these activities. Mark 0 if you participated in none.

	Enter # and hours below…	
	# of Different workshops	Total hours
a. Attended short, stand-alone training or workshop in reading (half-day or less)	#_____ 231-232/	_____**Hours** 233-235/
b. Attended longer institute or workshop in reading (more than half-day)	#_____ 236-237/	_____**Hours** 238-240/
c. Attended a college course in reading (include any courses you are currently attending)	#_____ 241-242/	_____**Hours** 243-245/
d. Attended a conference about reading (might include multiple short offerings)	#_____ 246-247/	_____**Hours** 248-250/

C2. Below is a list of professional development activities that are often used to provide ongoing, direct support to teachers for teaching reading.
- In the first column, please indicate whether you have received any of the following types of assistance/support for teaching **during the current school year, including summer 2006**.
- If you did not receive that type of support, please indicate whether the support was available, but you did not receive it (column 2), or if it was not available at your school (column 3).

	Check only one box for each item …		
	Types of assistance I received this year	**Available, but I did not receive**	**Not available at my school**
a. Coaching or mentoring by reading coach in programs, materials, or strategies.	□₁	□₂	□₃ 251/
b. Coaching or mentoring from fellow teacher.	□₁	□₂	□₃ 252/
c. Peer study group or collegial circle for group study.	□₁	□₂	□₃ 253/
d. Demonstrations in my classroom.	□₁	□₂	□₃ 254/
e. Observations of other teachers.	□₁	□₂	□₃ 255/
f. Diagnostic testing help from a reading coach or specialist for individual students.	□₁	□₂	□₃ 256/
g. Intervention service help from a reading coach or specialist for individual students.	□₁	□₂	□₃ 257/
h. Interpretation of assessment data.	□₁	□₂	□₃ 258/
i. Grade level meetings devoted to reading.	□₁	□₂	□₃ 259/
j. Using assessment data to determine topics that require additional instruction or practice.	□₁	□₂	□₃ 260/

C3. During the current school year, including summer 2006, **approximately** how many of the **reading professional development activities** for **K-3 teachers**: (Please choose the category that most closely describes your professional development.)

	Check only one box for each item…				
	None	**One-Quarter**	**One-Half**	**Three-Quarters**	**All**
a. were also attended by the principal?	□₁	□₂	□₃	□₄	□₅ 261/
b. provide teachers options among which to choose?	□₁	□₂	□₃	□₄	□₅ 262/
c. provide a stipend?	□₁	□₂	□₃	□₄	□₅ 263/
d. provide follow-up activities?	□₁	□₂	□₃	□₄	□₅ 264/
e. include release time for participating teachers?	□₁	□₂	□₃	□₄	□₅ 265/
f. offer graduate college credits?	□₁	□₂	□₃	□₄	□₅ 266/
g. are held in a convenient location (e.g., activities held at school)?	□₁	□₂	□₃	□₄	□₅ 267/
h. use a team-based approach (joint training of people who work together)?	□₁	□₂	□₃	□₄	□₅ 268/
i. are given by trainers or facilitators who have a well-established reputation?	□₁	□₂	□₃	□₄	□₅ 269/

C4. Below is a list of topics that are often covered in professional development activities designed to provide teachers with new information about the **content of reading instruction**.
- In column A, identify the topics that were addressed in professional development activities in which you participated **during the current school year, including summer 2006**.
- In column B, please identify the topics in which you need more professional development, whether or not this school's professional development activities have covered these topics.
- Please check all that apply in columns A and B.

Professional development is defined as any activity in which a teacher has learned about reading or reading instruction. This includes school-based workshops, meetings with reading coaches, and meetings with a study group of other teachers.

		CHECK ALL THAT APPLY	
		A. Topics addressed in professional development	**B.** Topics in which I need more professional development
Phonemic Awareness	a. Building phonological awareness, e.g. rhymes, dividing spoken language into sentences, words, syllables	$\square_1$ 270/	$\square_2$ 271/
	b. Identifying, adding, deleting sounds in spoken words	$\square_3$ 272/	$\square_4$ 273/
	c. Blending phonemes to form words	$\square_1$ 274/	$\square_2$ 275/
Decoding	d. Teaching letter-sound correspondence	$\square_3$ 276/	$\square_4$ 277/
	e. Teaching letter patterns (blends, digraphs, diphthongs)	$\square_1$ 278/	$\square_2$ 279/
	f. Using syllable patterns to read words	$\square_3$ 280/	$\square_4$ 281/
	g. Teaching component parts: roots, prefixes, suffixes	$\square_1$ 282/	$\square_2$ 283/
Vocabulary	h. Teaching use of dictionary, thesaurus	$\square_3$ 284/	$\square_4$ 285/
	i. Direct teaching of vocabulary words and their meaning	$\square_1$ 286/	$\square_2$ 287/
	j. Antonyms and synonyms	$\square_3$ 288/	$\square_4$ 289/
Fluency	k. Teaching sight words	$\square_1$ 290/	$\square_2$ 291/
	l. Guided oral reading	$\square_3$ 292/	$\square_4$ 293/
	m. Encouraging expression while reading	$\square_1$ 294/	$\square_2$ 295/
Comprehension	n. Setting motivation/asking prediction/preview questions	$\square_3$ 296/	$\square_4$ 297/
	o. Constructing information about character, setting, and main events	$\square_1$ 298/	$\square_2$ 299/
	p. Summarizing main ideas in narrative and informational text	$\square_3$ 300/	$\square_4$ 301/
	q. Self-monitoring strategies	$\square_1$ 302/	$\square_2$ 303/
	r. Asking questions at different levels (literal, inferential)	$\square_3$ 304/	$\square_4$ 305/
	s. Strategies for organizing text structure, e.g. story maps	$\square_1$ 306/	$\square_2$ 307/

308-309/BLANK

C5. Below is a list of topics that are often covered in professional development activities that are designed to provide teachers with new information about **teaching strategies used during reading instruction**.

- In column A, identify the topics that were addressed in professional development activities in which you participated **during the current school year, including summer 2006**.
- In column B, please identify the topics in which you need more professional development, whether or not this school's professional development activities have covered these topics.
- Please check all that apply in columns A and B.

Topic		A. Topics addressed in professional development.	B. Topics in which I need more professional development
		CHECK ALL THAT APPLY	
Teaching Strategies	a. How to use the core reading program	☐₁ 310/	☐₂ 311/
	b. How to use children's literature to teach reading	☐₃ 312/	☐₄ 313/
	c. How to use reading research to guide content of instruction	☐₁ 314/	☐₂ 315/
	d. How the core reading program incorporates research principles	☐₃ 316/	☐₄ 317/
	e. How to use the supplemental reading program(s)	☐₁ 318/	☐₂ 319/
	f. How to integrate reading and writing instruction	☐₃ 320/	☐₄ 321/
	g. Strategies for teaching reading to ELLs (see definition of ELL on page 2)	☐₁ 322/	☐₂ 323/
Grouping	h. Learning styles	☐₃ 324/	☐₄ 325/
	i. How to organize small group instruction	☐₁ 326/	☐₂ 327/
Assessment	j. How to diagnose reading problems	☐₃ 328/	☐₄ 329/
	k. How to administer assessments	☐₁ 330/	☐₂ 331/
	l. How to interpret and use assessment data to guide instruction	☐₃ 332/	☐₄ 333/
Struggling Readers	m. How to help struggling readers with decoding	☐₁ 334/	☐₂ 335/
	n. How to help struggling readers with vocabulary	☐₃ 336/	☐₄ 337/
	o. How to help struggling readers with comprehension	☐₁ 338/	☐₂ 339/
	p. How to motivate readers	☐₃ 340/	☐₄ 341/
	q. Strategies for teaching reading to students with diagnosed learning disabilities	☐₁ 342/	☐₂ 343/
Organization/ planning	r. How to use state/district content standards for curriculum planning and teaching	☐₃ 344/	☐₄ 345/
	s. How to align reading curriculum and instruction with state/district assessments	☐₁ 346/	☐₂ 347/
	t. How to work with parents	☐₃ 348/	☐₄ 349/
	u. Classroom management	☐₁ 350/	☐₂ 351/

C6. How well do you feel the **professional development activities in which you participated during the current school year (including summer 2006)** prepared you to teach each of the following dimensions of reading? Please choose a '1' if you were 'not at all prepared' to teach the dimension and a '5' if you were 'extremely well prepared.'

	Check only one box for each item …				
	Not at all prepared		**Somewhat well prepared**		**Extremely well prepared**
a. Phonemic awareness	☐1	☐2	☐3	☐4	☐5 352/
b. Decoding	☐1	☐2	☐3	☐4	☐5 353/
c. Vocabulary	☐1	☐2	☐3	☐4	☐5 354/
d. Comprehension	☐1	☐2	☐3	☐4	☐5 355/
e. Fluency building	☐1	☐2	☐3	☐4	☐5 356/

D. Support for Teaching Reading

D1. The next set of statements is about your reading program. Please indicate the extent to which you agree or disagree with each statement.

	Check one box for each item …			
	Strongly Agree	**Agree**	**Disagree**	**Strongly Disagree**
a. I feel I need to make changes in the methods I use to teach children to read.	☐1	☐2	☐3	☐4 357/
b. Other faculty/staff members have helped me to understand the difficulties that some children have in learning to read.	☐1	☐2	☐3	☐4 358/
c. I have benefited from opportunities to learn more about methods for teaching reading.	☐1	☐2	☐3	☐4 359/
d. The children in my class are making satisfactory progress in learning to read.	☐1	☐2	☐3	☐4 360/
e. I do not have sufficient materials to teach reading effectively.	☐1	☐2	☐3	☐4 361/
f. I do not understand why some children learn to read easily while other children struggle to learn basic reading skills.	☐1	☐2	☐3	☐4 362/
g. The reading coach supports my efforts to teach reading effectively.	☐1	☐2	☐3	☐4 363/
h. I have a good understanding of how children acquire language and literacy skills.	☐1	☐2	☐3	☐4 364/
i. I wish I had more opportunities to discuss how to teach reading with other teachers.	☐1	☐2	☐3	☐4 365/
j. I know the current reading skill levels of all my students.	☐1	☐2	☐3	☐4 366/
k. I know how to assess the progress of my students in reading.	☐1	☐2	☐3	☐4 367/
l. I have changed my methods of teaching reading as a result of professional development in reading.	☐1	☐2	☐3	☐4 368/

E. Out of School Reading

E1.   What strategies do you use in your classroom to encourage students to read outside of school?

Strategies to encourage students to read outside of school	Check only one box for each item…	
	Yes	**No**
a. Classroom-wide reading contests	$\square_1$	$\square_2$ 389/
b. School-wide reading contests	$\square_1$	$\square_2$ 390/
c. Book giveaways or auctions	$\square_1$	$\square_2$ 391/
d. Individual rewards (e.g., coupons, prizes) for reading a certain number of books or pages outside of school	$\square_1$	$\square_2$ 392/
e. Class rewards (e.g., pizza party, "choice time") for reaching a class goal for out-of-school reading	$\square_1$	$\square_2$ 393/
f. Communication with parents about reading resources	$\square_1$	$\square_2$ 394/
g. Classroom relationship with community libraries	$\square_1$	$\square_2$ 395/
h. Use of books or book series students can continue to read at home	$\square_1$	$\square_2$ 396/
i. Partnering students who live near each other to read to each other out of school	$\square_1$	$\square_2$ 397/
j. Classroom book clubs discussing books read outside of school	$\square_1$	$\square_2$ 398/
k. Tape-reading students reading favorite parts of books	$\square_1$	$\square_2$ 399/
l. Inviting community members to speak of their recreational reading	$\square_1$	$\square_2$ 400/
m. Asking students to read to siblings and/or parents at home	$\square_1$	$\square_2$ 401/
n. Using classroom resources for students to buy books through children's book publisher	$\square_1$	$\square_2$ 402/
o. Having students write and illustrate their own books they bring home to read	$\square_1$	$\square_2$ 403/
p. Sharing your own recreational reading with the class	$\square_1$	$\square_2$ 404/
q. Providing time for students to retell stories they've read	$\square_1$	$\square_2$ 405/
r. Providing opportunities for students to act out or role-play stories they've read	$\square_1$	$\square_2$ 406/

E2.   Do you assess students' out-of-school reading?
$\square_1$ Yes→**Skip to E2a**   410/
$\square_2$ No

E2a.   If yes, how do you assess students' out-of school reading? (Check all that apply)
$\square_1$ Reading logs   411/
$\square_2$ Book reports   412/
$\square_3$ Parent verficiation   413/
$\square_{95}$ Other (Please specify): _____   414-415/

Thank you for your cooperation and for taking the time to answer these questions. Please place the completed survey in the enclosed envelope, seal the envelope, and return to your school's Reading First Impact Study liaison.

If you have any questions about the survey, please call 1-866-421-6982 and leave a message, or send an e-mail to reading_impact@abtassoc.com

Exhibit C.19: Instructional Practice in Reading Inventory (IPRI)

Instructional Practice in Reading Inventory (IPRI)

Part A. Dimensions of Reading		Interval 1 Time: ___ : ___ AM PM

DP — Decoding with PRINT (Phonics)

Grouping		T. Instruction: The Teacher. . .	S. Student Practice: The Teacher . . .
W	Whole class	T1 Describes, explains, identifies or asks S(s) to describe, explain or identify	S1 Gives S(s) practice decoding words (PRINT TO SOUND)
L	Lg grp	• *sound-symbol pattern*	S2 Gives S(s) practice encoding words (manipulating letters or writing, not copying)
S	Sm grp	• *decoding rule*	(SOUND TO PRINT)
P	Pair	• *word structure pattern or rule*	S3 Asks student(s) to orally spell word(s)
I	Individual		
		T2 Shows how to apply a rule or pattern to whole word example(s)	
Support		T3 Identifies word(s) that contrast with or do not follow pattern or rule	
TM	Tchr Manipulatives	T4 ***If S makes mistake:***	
V	Picture, Object	Reminds S of pattern or rule and has S produce or repeat correct response	
ST	Sentence(s)/Text		GH **Grammar or Handwriting is part of lesson**
SM	Std Manipulatives	PA **Includes Phonemic Awareness example**	OR **Student Oral Reading is part of lesson**

CP — Comprehension of Connected Text

Grouping		T. Instruction: The Teacher. . .	S. Student Practice: The Teacher . . .
W	Whole class	***Before reading text:***	***During or after reading text passage:***
L	Lg grp	T1 Conducts pre-reading activity(ies)	S1 Asks students to answer literal recall questions about specific details in the text
S	Sm grp	T1a Previews vocabulary prior to lesson (Go to Vocab)	
P	Pair	***Before, during, or after reading text passage:***	
I	Individual	T2 Describes or explains—or asks Ss to describe or explain—one or more comprehension strategies	S2 Asks students to identify or describe genre
		T2a Specifies what the strategy is called	
Support		T2b Specifies why the strategy is helpful	
		T2c Specifies when in the reading process the strategy is used	
V	Picture, Object	***During or after reading text passage:*** Shows how to apply strategy	***During or after reading text passage:*** Gives S(s) practice applying strategy
O	Text Organizer	Using cues to support interpretation or make predictions about text:	Using cues to support interpretation or make predictions about text:
		T3a1 Pictures	S3a1 Pictures
C	Connected text	T3a2 Text cues (e.g., headers, captions)	S3a2 Text cues (e.g., headers, captions)
E	Expository	T3b Answer inferential questions based on text	S3b Answer inferential questions based on text
N	Narrative	T3c Make predictions based on text	S3c Make predictions based on text
CD	Can't determine	T3d Summarize, retell, sequence text or identify main idea(s)	S3d Summarize, retell, or sequence text or identify main idea(s)
		T3e Make text-text connections	S3e Make text-text connections
		T3f Work with story or expository structure	S3f Work with story or expository structure
		T3g Use mental imagery to support interpretation of text	S3g Use mental imagery to support interpretation of text
		T3h Generate own questions about text	S3h Generate own questions about text
		T3i Answer own questions about text	S3i Answer own questions about text
		T3j Review passage to check or clarify understanding	S3j Review passage to check or clarify understanding
		T3k Check accuracy of prediction or inference	S3k Check accuracy of prediction or inference
		T4 Teaches vocabulary during or after reading text passage (Go to Vocab)	S4 Asks S(s) to justify their response with evidence
		T5 ***If S response is incorrect or incomplete:*** Assists S in using strategy(ies)	S5 Sets up independent practice for Ss to apply comprehension strategy(ies) (student work product or response required)

HD **Helps student(s) Decode word(s)**		GH **Grammar or Handwriting is part of lesson**	
OR **Student Oral Reading is part of lesson**	SR **Student Silent Reading is part of lesson**	TOR **Teacher orally reads as students listen**	

Interval 1

VD	**Vocabulary Development**		

Grouping		**T. Instruction: The Teacher. . .**		**S. Student Practice: The Teacher . . .**	
W	Whole class	T1	Asks S(s) to give meaning of word	S1	Asks S(s) to apply understanding of word meaning
L	Lg grp	T2	Gives synonym	S2	Gives S(s) opportunity to practice word learning strategy(ies)
S	Sm grp	T3	Goes beyond synonym with definition and/or examples		(e.g., context, word structure, root meanings)
P	Pair	T4	Pinpoints word meaning with contrasting examples		
I	Individual	T5	Pinpoints word meaning by clarifying or extending Ss' partially correct response:	**List vocabulary words:**		
Support						
V	Picture, Object	T5a	Extension/clarification includes synonym			
P	Physical Demo.	T5b	Extension/clarification includes definition and/or example			
M	Word Map	T5c	Extension/clarification includes contrasting example	**GH**	**Grammar or Handwriting is part of lesson**	
		HD	**Helps student(s) Decode word(s)**	**OR**	**Student Oral Reading is part of lesson**	

PA	**Phonemic / Phonological Awareness (Sounds, NO PRINT)**		

Grouping		**T.**	**Instruction: The Teacher. . .**	**S.**	**Student Practice: The Teacher . . .**
W	Whole class		*Demonstrates or models*:		*Gives S(s) chance to practice:*
L	Lg grp	T1a	Oral work with syllables	S1a	Oral work with syllables
S	Sm grp	T1b	Oral blending or segmenting with onset-rimes	S1b	Oral blending or segmenting with onset-rimes
P	Pair	T1c	Oral blending or segmenting with phonemes	S1c	Oral blending or segmenting with phonemes
I	Individual	T1d	Phoneme Isolation	S1d	Phoneme isolation
		T1e	Phoneme categorization/identity (same/different sound in words)	S1e	Phoneme categorization/identity (same/different sound in words)
		T1f	Phoneme deletion, addition, substitution	S1f	Phoneme deletion, addition, substitution
Support		T2	Contrasts two phonemes to pinpoint target sound		
MK	Tchr Manip or Kin	T3	Pinpoints what S(s) did incorrectly with sound(s) and gives correct response with or		
SM	Manipulatives		without students		
SK	Kinesthetic	T4	Introduces printed letters corresponding to sounds		

FB	**Fluency Building With Connected Text**	

Grouping		**T.**	**Instruction & Student Practice: The Teacher. . .**
W	Whole class	T1	Sets up or prompts S(s) to practice repeated or timed readings with a listener
L	Lg grp		
S	Sm grp		*Listens to Ss practice repeated oral readings:*
P	Pair	T2a	With text that was not modeled
I	Individual	T2b	With same text that was modeled by fluent reader
Support			*Listens to Ss practice timed oral readings:*
Ⓒ	Connected text	T3a	With text that was not modeled
W	Written record	T3b	With same text that was modeled by fluent reader

Part B. Other Instruction	OR	Oral Reading	SP	Spelling	AS	Assessment	TR	Transition	MB: Managing Behavior:
	SR	Silent Reading	WE	Written Expression			AM	Academic Mgmt	① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩
	TOR	Teacher Oral Reading	OL	Other Language Arts	NL	Non-literacy instruction	NI	Non-instructional	

Part C. Instructional Errors		**Part D.** Small group changes	T working with new small group YES

Exhibit C.20: Global Appraisal of Teaching Strategies

Classroom ID#:_____   AbtClassID#: _____

Item The teacher: Check one for each item:	All the time	More than half the time	Half the time	Less than half the time	Never	Not observed
1 **Seems to be organized and has all the materials necessary for instruction easily accessible** Locates and introduces new materials smoothly; comfortable using equipment and materials	❑	❑	❑	❑	❑	❑
2 **Presents and delivers the lesson effectively (makes eye contact, varies voice, engaged)** Makes eye contact with individual students; uses voice for emphasis, to hold student attention; engaged in delivering instruction, neither bored nor distracted; not a distraction from focus on content	❑	❑	❑	❑	❑	❑
3 **Distributes opportunities for students to participate equally and broadly** Solicits responses from a variety of students not just the most active volunteers or same subset	❑	❑	❑	❑	❑	❑
4 **Manages and coordinates student responses effectively** Has students listen to each other, take turns, cooperate and be courteous	❑	❑	❑	❑	❑	❑
5 **Pays attention, acknowledges and responds to student responses or input** Listens actively. Offers verbal/nonverbal acknowledgment, takes opportunities to build on student response	❑	❑	❑	❑	❑	❑
6 **Takes advantage of opportunities to provide corrective feedback when student makes error** Notices when student makes error and takes the opportunity to correct it	❑	❑	❑	❑	❑	❑
7 **Provides feedback in a positive manner** Feedback is not condescending or harsh; feedback may be firm but is neutral or encouraging	❑	❑	❑	❑	❑	❑
8 **Manages student behavior effectively in order to avoid disruptions and to provide productive learning opportunities** Provides enough structure to minimize disruptions; quickly and firmly intervenes to redirect off-task behavior	❑	❑	❑	❑	❑	❑
9 **Maintains a classroom environment that minimizes distractions and is appropriate for learning** Structures activities so groups do not interfere with each other; monitors noise level, insures that students have sufficient physical space (e.g., seating) and resources to attend to instruction	❑	❑	❑	❑	❑	❑
10 **Clearly communicates classroom procedure(s) when student breaks a rule or does not follow proper procedure** If needed, reminds student of relevant class rule or procedure; clear about what student is expected to do	❑	❑	❑	❑	❑	❑

The teacher: Check one:	True	Mostly true	Not sure	Mostly not true	Not true	
11 **Is generally motivated and keeps students actively involved** Appears goal-directed, motivated to deliver instruction, interested in holding students' attention	❑	❑	❑	❑	❑	

The teacher: Check one:	All Ss	More than half of Ss	Half of Ss	Less than Half of Ss	None	
12 **Checks on or monitors individual students' progress** Examines students' independent work one by one, and/or checks on individual students' skill level via one-on-one conference (can be brief), or formal/informal assessment	❑	❑	❑	❑	❑	

Exhibit C.21: Student Time-on-Task and Engagement with Print (STEP) Instrument

BARCODE HERE   BARCODE HERE

Classroom ID		AbtClassID	
School ID		School Name	
Observer ID		Observer Name	
Date			
Grade		Room number	
Total # of students			
Time observation began		Time observation ended	

Anything unusual?

Classroom combined with another? Check here if YES ❑ and complete below:	
❑ Both classrooms part of the study	❑ Only one classroom is part of the study

ClassID observed		ClassID observed	
ClassID not observed			

Sources consulted:

Foorman, B.R. & Schatschneider, C. (2003). Measurement of teaching practices during reading/language arts instruction and its relationship to student achievement. In S. Vaughn and K.L. Briggs (Eds.), <u>Reading in the classroom: Systems for the Observation of Teaching and Learning</u> (pp. 1-30). Baltimore, MD: Paul H. Brooks.

Kim, A., Briggs, K.L., & Vaughn, S. (2003). The classroom climate scale. In S. Vaughn and K.L. Briggs (Eds.), <u>Reading in the classroom: Systems for the Observation of Teaching and Learning</u> (pp. 83-109). Baltimore, MD: Paul H. Brooks.

Vaughn, S. (2005). Personal communication (May, 2005).

Shanahan, T. (2005). Personal communication (May, 2005)

Version 3.2 8/15/06 RFIS 2006-2007 Field Version

Instructions:
STEP observations are focused on the students not on the teacher.

Enter the classroom and start your countdown watch (set for 3:00 minute intervals). Begin walking casually around the room looking at what students are doing during the acclimation phase. Let students become accustomed to you walking around the room. Do not impede teacher or student movement but do not sit in one location during the acclimation period. Do not block IPRI observer's view of the teacher. Do not interact with students. If a student makes eye contact with you, look away – do not smile or otherwise acknowledge the student. If a student talks to you, smile and respond firmly, "I'm working, I can't talk with you right now."

After 6 minutes (beginning with the second buzz of your watch), begin **Sweep 1**:
If the more than ½ the students in the class are in transition, circle Y at the top of the Sweep 1 column and wait for Sweep 2;
If the whole class is listening the to the teacher read aloud a story (not following along in their own texts), circle Y at the top of Sweep 1 column and wait for Sweep 2. Otherwise:
   a)   Select a student to be Pupil 1. Record whether or not P1 is on or off task.
   b)   If P1 is On Task, record whether or not P1 is Reading Connected Text, Reading Isolated Text, or Writing (or none of these three); **If P1 is Off-Task do not fill in any of the three remaining columns for that student.**
   c)   Go on to the next student (P2). Repeat above steps for P2, P3, etc.
   d)   When you have finished with all students in the classroom, draw a solid line underneath the last pupil. Only students who are in the classroom should be counted in any Sweep.

Six minutes (two watch buzzes) after the start of Sweep 1, begin **Sweep 2**. Do not try to observe children in the same order that you observed during Sweep 1. That is, do not try to locate the same individual student who was P1 in Sweep 1. Instead, move around the room in a systematic fashion to observe each child once per sweep.

Repeat for **Sweep 3**.

After Sweep 3 is complete, go to the next classroom. Be sure to use a new STEP for each classroom observed.

On-task behavior is any behavior in which a child appears to be:
   •   Engaged in some independent, paired, or group learning activity (regardless of what teacher has asked student to do);
   •   Paying attention to the teacher (if teacher is delivering instruction) or attending to work that is in front of him/her;
   •   Talking with other students about an instructional activity in which both students are engaged (e.g. working productively on a group project);
   •   Participating in a whole class routine such as the pledge of allegiance

Off-task behaviors include:
   •   Not paying attention to the teacher when appropriate
   •   Looking around or gazing at an activity in which student is not engaged; "blank stares"
   •   Crying or head down with eyes covered on desk.
   •   Going to get new materials or put old materials away; or, wandering aimlessly
   •   Conflict with students or teacher
   •   Playing, teasing, roughhousing with other students, distracting other students
   •   Play behavior (playing with non-literacy related board games, blocks, dolls, action figures, legos, etc.)
   •   Snack/meal times or transitions (e.g., lining up to use the rest room)

Reading Connected Text:
Eyes are on a book, story, passage, or child is turning to next page of story. Even if student is momentarily looking at pictures that accompany a story, code as reading connected text unless book has no text. However, if student is flipping quickly through a book without pausing to read words on the page, do not code as Reading Connected Text, even though student may be coded as OnTask.

Reading Isolated Text:
Working with flashcards, looking at letters or words in isolation, completing a worksheet with isolated letters, words, sentences. Reading isolated sentences not part of a coherent, connected passage.

Writing:
Student has pen, pencil, crayon or other writing implement in hand and is writing or copying text, either isolated or connected. If student is drawing pictures do not code "Writing."

Version 3.2 8/15/06 RFIS 2006-2007 Field Version

Student Time-on-Task and Engagement with Print (STEP)

SWEEP 1			
Circle one:			
Half or more of Class in Transition?	Y	N	**If Yes, SKIP this SWEEP**
Whole Class Listening to Story?	Y	N	**If Yes, SKIP this SWEEP**

		Engagement with Print		
On Task?	**On Task?**	**Reading Connected Text**	**Reading Isolated Text**	**Writing**
01	Y N	❑	❑	❑
02	Y N	❑	❑	❑
03	Y N	❑	❑	❑
04	Y N	❑	❑	❑
05	Y N	❑	❑	❑
06	Y N	❑	❑	❑
07	Y N	❑	❑	❑
08	Y N	❑	❑	❑
09	Y N	❑	❑	❑
10	Y N	❑	❑	❑
11	Y N	❑	❑	❑
12	Y N	❑	❑	❑
13	Y N	❑	❑	❑
14	Y N	❑	❑	❑
15	Y N	❑	❑	❑
16	Y N	❑	❑	❑
17	Y N	❑	❑	❑
18	Y N	❑	❑	❑
19	Y N	❑	❑	❑
20	Y N	❑	❑	❑
21	Y N	❑	❑	❑
22	Y N	❑	❑	❑
23	Y N	❑	❑	❑
24	Y N	❑	❑	❑
25	Y N	❑	❑	❑
26	Y N	❑	❑	❑
27	Y N	❑	❑	❑
28	Y N	❑	❑	❑
29	Y N	❑	❑	❑
30	Y N	❑	❑	❑

Student Time-on-Task and Engagement with Print (STEP)

SWEEP 2				
	Circle one:			
Half or more of Class in Transition?	Y	N	**If Yes, SKIP this SWEEP**	
Whole Class Listening to Story?	Y	N	**If Yes, SKIP this SWEEP**	

Pupil	On Task?	**Engagement with Print**		
		Reading Connected Text	Reading Isolated Text	Writing
01	Y N	❑	❑	❑
02	Y N	❑	❑	❑
03	Y N	❑	❑	❑
04	Y N	❑	❑	❑
05	Y N	❑	❑	❑
06	Y N	❑	❑	❑
07	Y N	❑	❑	❑
08	Y N	❑	❑	❑
09	Y N	❑	❑	❑
10	Y N	❑	❑	❑
11	Y N	❑	❑	❑
12	Y N	❑	❑	❑
13	Y N	❑	❑	❑
14	Y N	❑	❑	❑
15	Y N	❑	❑	❑
16	Y N	❑	❑	❑
17	Y N	❑	❑	❑
18	Y N	❑	❑	❑
19	Y N	❑	❑	❑
20	Y N	❑	❑	❑
21	Y N	❑	❑	❑
22	Y N	❑	❑	❑
23	Y N	❑	❑	❑
24	Y N	❑	❑	❑
25	Y N	❑	❑	❑
26	Y N	❑	❑	❑
27	Y N	❑	❑	❑
28	Y N	❑	❑	❑
29	Y N	❑	❑	❑
30	Y N	❑	❑	❑

Student Time-on-Task and Engagement with Print (STEP)

SWEEP 3				
	Circle one:			
Half or more of Class in Transition?		Y	N	**If Yes, SKIP this SWEEP**
Whole Class Listening to Story?		Y	N	**If Yes, SKIP this SWEEP**

		Engagement with Print		
Pupil	**On Task?**	**Reading Connected Text**	**Reading Isolated Text**	**Writing**
01	Y　N	❏	❏	❏
02	Y　N	❏	❏	❏
03	Y　N	❏	❏	❏
04	Y　N	❏	❏	❏
05	Y　N	❏	❏	❏
06	Y　N	❏	❏	❏
07	Y　N	❏	❏	❏
08	Y　N	❏	❏	❏
09	Y　N	❏	❏	❏
10	Y　N	❏	❏	❏
11	Y　N	❏	❏	❏
12	Y　N	❏	❏	❏
13	Y　N	❏	❏	❏
14	Y　N	❏	❏	❏
15	Y　N	❏	❏	❏
16	Y　N	❏	❏	❏
17	Y　N	❏	❏	❏
18	Y　N	❏	❏	❏
19	Y　N	❏	❏	❏
20	Y　N	❏	❏	❏
21	Y　N	❏	❏	❏
22	Y　N	❏	❏	❏
23	Y　N	❏	❏	❏
24	Y　N	❏	❏	❏
25	Y　N	❏	❏	❏
26	Y　N	❏	❏	❏
27	Y　N	❏	❏	❏
28	Y　N	❏	❏	❏
29	Y　N	❏	❏	❏
30	Y　N	❏	❏	❏

Appendix D: Confidence Intervals

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, instructional outcomes, and student engagement with print. Data are reported across these areas for pertinent study years.

Exhibit D.1: Confidence Intervals for Estimated Impacts on Reading Comprehension and Decoding Skills: Spring 2005, 2006, and 2007*

	Metric			Effect Size		
	Impact	Standard Error	Confidence Interval	Impact	Standard Error	Confidence Interval
Reading Comprehension (SAT 10)						
Scaled Scores						
Grade 1	4.74	2.72	-0.63 – 10.11	0.10	0.06	-0.01 – 0.21
Grade 2	1.69	2.29	-2.83 – 6.21	0.04	0.05	-0.07 – 0.15
Grade 3	0.30	2.12	-3.88 – 4.48	0.01	0.05	-0.10 – 0.11
Percent reading at or above grade level						
Grade 1	4.22	2.58	-0.87 – 9.31	N/A	N/A	N/A
Grade 2	1.60	2.38	-3.11 – 6.30	N/A	N/A	N/A
Grade 3	-0.08	2.40	-4.81 – 4.65	N/A	N/A	N/A
Decoding Skills (TOSWRF)						
Standard Score						
Grade 1	2.51*	1.11	0.32 – 4.69	0.17*	0.07	0.02 – 0.31

NOTES:

*The TOWSRF was administered in spring 2007 only.

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.

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

A 95% confidence interval was used.

EXHIBIT READS: The estimated impact of the Reading First program for grade 1 on reading comprehension scaled scores was 4.74 points with a standard error of 2.72 scaled score points. The 95% confidence interval for the estimated impact ranged from -0.63 points to 10.11 points.

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 TOSWRF administration in the spring of 2007.

Exhibit D.2: Confidence Intervals for Estimated Impacts on Instructional Outcomes: Spring 2005, Fall 2005, Spring 2006, Fall 2006 and Spring 2007

	Metric			Effect Size		
	Impact	Standard Error	Confidence Interval	Impact	Standard Error	Confidence Interval
Panel 1	*(minutes)*					
Number of minutes of instruction in five dimensions combined						
Grade 1	6.92*	2.44	2.13 - 11.71	0.33*	0.12	0.10 - 0.56
Grade 2	9.79*	2.46	4.97 - 14.60	0.46*	0.12	0.23 - 0.69
Panel 2	*(percent)*					
Percentage of intervals in five dimensions with						
Highly Explicit Instruction						
Grade 1	3.29*	1.39	0.57 - 6.01	0.18*	0.08	0.03 - 0.33
Grade 2	3.00*	1.45	0.16 - 5.84	0.16*	0.08	0.01 - 0.30
High Quality Student Practice						
Grade 1	0.82	1.25	-1.63 - 3.27	0.05	0.07	-0.10 - 0.19
Grade 2	2.94*	1.25	0.50 - 5.39	0.16*	0.07	0.03 - 0.30

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.

Impact estimates are statistically adjusted 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 95% confidence interval was used.

EXHIBIT READS: The estimated impact of the Reading First program for grade 1 on the amount of time spent in instruction in the five dimensions was 6.92 minutes with a standard error of 2.44 minutes. The estimated impact was statistically significant. The 95% confidence interval for the estimated impact ranged from 2.13 minutes to 11.71 minutes.

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

Exhibit D.3: Confidence Intervals for Estimated Impacts on Time Spent in Instruction in the Five Dimensions: Spring 2005, Fall 2005, Spring 2006, Fall 2006, and Spring 2007

	Metric			Effect Size		
	Impact	Standard Error	Confidence Interval	Impact	Standard Error	Confidence Interval
Number of minutes of instruction in:						
Phonemic Awareness						
Grade 1	0.61*	0.28	0.06 - 1.16	0.23*	0.10	0.02 - 0.43
Grade 2	0.12	0.12	-0.12 - 0.36	0.10	0.10	-0.09 - 0.29
Phonics						
Grade 1	2.86*	1.44	0.04 - 5.68	0.21*	0.11	0.00 - 0.42
Grade 2	3.27*	1.17	0.97 - 5.57	0.31*	0.11	0.09 - 0.53
Vocabulary						
Grade 1	0.57	0.66	-0.72 - 1.86	0.09	0.10	-0.11 - 0.28
Grade 2	1.73*	0.82	0.13 - 3.34	0.20*	0.09	0.01 - 0.38
Fluency						
Grade 1	1.24*	0.61	0.05 - 2.43	0.20*	0.10	0.01 - 0.40
Grade 2	0.58	0.56	-0.51 - 1.67	0.11	0.10	-0.09 - 0.31
Comprehension						
Grade 1	1.78	1.53	-1.22 - 4.77	0.12	0.11	-0.08 - 0.33
Grade 2	4.01*	1.69	0.70 - 7.32	0.24*	0.10	0.04 - 0.44

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.

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

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

A 95% confidence interval was used.

EXHIBIT READS: The estimated impact of the Reading First program for grade 1 on the amount of time spent in instruction in phonemic awareness was 0.61minutes with a standard error of 0.28 minutes. The estimated impact was statistically significant at the p≤.05 level. The 95% confidence interval for the estimated impact ranged from 0.06 minutes to 1.16 minutes.

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

Exhibit D.4: Confidence Intervals for Estimated Impacts on Student Engagement with Print: Fall 2005, Spring 2006, Fall 2006, and Spring 2007

	Metric			Effect Size		
	Impact (percent)	Standard Error	Confidence Interval	Impact	Standard Error	Confidence Interval
Percentage of student engagement with print						
Grade 1	5.33	2.92	-0.40 – 11.05	0.18	0.10	-0.01 – 0.38
Grade 2	-4.75	2.91	-10.45 – 0.96	-0.17	0.10	-0.37 – 0.03

NOTES:

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

Impact estimates are statistically adjusted 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 95% confidence interval was used.

EXHIBIT READS: The estimated impact of the Reading First program for grade 1 on the percentage of students engaged with print was 5.33 percentage points with a standard error of 2.92 percentage points. The 95% confidence interval for the estimated impact ranged from -0.40 percentage points to 11.05 percentage points.

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

Appendix E: Analyses of Impacts and Trends Over Time

Part 1: Additional Exhibits of Separate Impact Estimates for Each Follow-up Year and Pooled

Exhibits E.1 through E.3 present additional exhibits of separate impact estimates for each grade level for each follow-up year as well as pooled across applicable years, for the number of minutes of instruction in each of the five dimensions of reading and for the percentage of students engaged with print.

Part 2: Student Achievement Trends Over Time

This part of the appendix presents a brief discussion of student achievement results over time.

Part 3: Reading Achievement on State Tests

The final part of the appendix reports on the estimated impact of Reading First on statewide reading tests in the RFIS core sample, and discusses how these results compare to results on the SAT 10 reading comprehension subtest for the RFIS core sample.

Exhibit E.1: Estimated Impacts on the Number of Minutes of Instruction in Each of Five Dimensions of Reading in First Grade: 2005, 2006, and 2007, and 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					
Spring 2005	1.64	0.76	0.88*	0.33*	(0.004)
School year 2006	2.32	1.69	0.62	0.23	(0.102)
School year 2007	2.71	2.19	0.52	0.19	(0.244)
Pooled 3 years (Sp05, Sy06, Sy07)	2.32	1.71	0.61*	0.23*	(0.030)
Phonics					
Spring 2005	21.02	18.05	2.97	0.22	(0.141)
School year 2006	21.56	16.99	4.57*	0.34*	(0.012)
School year 2007	21.27	18.90	2.37	0.18	(0.223)
Pooled 3 years (Sp05, Sy06, Sy07)	21.32	18.45	2.86*	0.21*	(0.048)
Vocabulary					
Spring 2005	7.03	5.48	1.55	0.23	(0.072)
School year 2006	8.22	8.02	0.20	0.03	(0.827)
School year 2007	8.10	7.86	0.24	0.04	(0.794)
Pooled 3 years (Sp05, Sy06, Sy07)	7.92	7.35	0.57	0.09	(0.386)
Fluency					
Spring 2005	5.26	3.72	1.53	0.25	(0.180)
School year 2006	4.13	3.22	0.91	0.15	(0.165)
School year 2007	4.84	3.23	1.61*	0.26*	(0.041)
Pooled 3 years (Sp05, Sy06, Sy07)	4.67	3.43	1.24*	0.20*	(0.043)
Comprehension					
Spring 2005	24.29	22.19	2.10	0.15	(0.349)
School year 2006	23.27	20.85	2.42	0.17	(0.244)
School year 2007	22.01	20.80	1.21	0.08	(0.504)
Pooled 3 years (Sp05, Sy06, Sy07)	23.01	21.23	1.78	0.12	(0.247)

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≤.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 in spring 2005 with Reading First was 1.64 minutes. The estimated mean amount of time without Reading First was 0.76 minutes. The impact of Reading First on the amount of time spent in instruction in phonemic awareness was 0.88 minutes (or 0.33 standard deviations), which was statistically significant (p=.004).

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

Exhibit E.2: Estimated Impacts On the Number of Minutes in Instruction in Each of Five Dimensions of Reading in Second Grade: 2005, 2006, and 2007, and Pooled

Number of minutes of instruction in:	Actual Mean With Reading First	Actual Mean Without Reading First	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Grade 2					
Phonemic Awareness					
Spring 2005	0.43	0.30	0.13	0.10	(0.386)
School year 2006	0.42	0.28	0.14	0.11	(0.271)
School year 2007	0.59	0.49	0.11	0.09	(0.621)
Pooled 3 years (Sp05, Sy06, Sy07)	0.49	0.37	0.12	0.10	(0.319)
Phonics					
Spring 2005	13.99	10.74	3.25*	0.31*	(0.044)
School year 2006	14.02	10.09	3.93*	0.37*	(0.013)
School year 2007	13.87	10.93	2.94	0.28	(0.067)
Pooled 3 years (Sp05, Sy06, Sy07)	13.92	10.65	3.27*	0.31*	(0.006)
Vocabulary					
Spring 2005	10.46	8.69	1.76	0.20	(0.117)
School year 2006	12.27	9.93	2.34	0.27	(0.058)
School year 2007	12.05	10.86	1.19	0.14	(0.275)
Pooled 3 years (Sp05, Sy06, Sy07)	11.79	10.06	1.73*	0.20*	(0.036)
Fluency					
Spring 2005	5.12	2.80	2.32*	0.42*	(0.013)
School year 2006	3.73	4.22	-0.49	-0.09	(0.475)
School year 2007	3.98	3.54	0.44	0.08	(0.523)
Pooled 3 years (Sp05, Sy06, Sy07)	4.14	3.56	0.58	0.11	(0.297)
Comprehension					
Spring 2005	28.33	22.84	5.49*	0.33*	(0.024)
School year 2006	29.69	24.85	4.85*	0.29*	(0.033)
School year 2007	28.08	26.24	1.84	0.11	(0.372)
Pooled 3 years (Sp05, Sy06, Sy07)	28.74	24.73	4.01*	0.24*	(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≤.05 level are indicated by *.

EXHIBIT READS: The observed mean amount of time spent per daily reading block in instruction in phonemic awareness for second grade classrooms in spring 2005 with Reading First was 0.43 minutes. The estimated mean amount of time without Reading First was 0.30 minutes. The impact of Reading First on the amount of time spent in instruction in phonemic awareness was 0.13 minutes (or 0.10 standard deviations), which was not statistically significant (p=.386).

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

Exhibit E.3: Estimated Impacts on the Percentage of Students Engaged with Print: 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					
Percentage of students engaged with print					
School Year 2006	46.92	42.28	4.64	0.16	(0.216)
School Year 2007	48.68	43.24	5.44	0.18	(0.170)
Pooled (SY 2006, SY 2007)	47.84	42.52	5.33	0.18	(0.070)
Grade 2					
Percentage of students engaged with print					
School Year 2006	49.83	58.25	-8.42*	-0.29*	(0.029)
School Year 2007	51.13	52.14	-1.01	-0.04	(0.808)
Pooled (SY 2006, SY 2007)	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 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).

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≤.05 level are indicated by *.

EXHIBIT READS: For the 2006 school year, the actual average percentage of students engaged with print in first grade classrooms with Reading First was 46.92 percent. The estimated average percentage without Reading First was 42.28 percent. The impact of Reading First on the average percentage of student engagement with print was 4.64 percentage points (or 0.16 standard deviations), which was not statistically significant (p=.216).

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

Part 2: Student Achievement Trends Over Time

Exhibits E.4 and E.5 present student achievement trends over time for schools in the RFIS study sample. Data on mean SAT 10 scores are presented at three time points—spring 2005, spring 2006 and spring 2007—separately for Reading First and non-Reading First schools across the 248 schools in the 18 sites in the RFIS study sample.

For each year and grade, three mean scaled score values were calculated. **The Actual Mean with Reading First** value is simply that; it is the actual unadjusted mean for the Reading First schools in the study sample. The **Estimated Mean without Reading First** value represents the best estimate of what would have happened in Reading First schools absent Reading First funding. The Actual Mean for Non-Reading First schools value is the unadjusted mean for the non-Reading First schools in the study sample.[1]

The Estimated Mean without Reading First is the counterfactual and in the absence of Reading First represents the best estimate of what would have happened in the treatment schools—if they had not been selected as Reading First schools. The Actual Mean with Reading First and the Estimated Mean without Reading First values are identical to the values shown in the impact tables in Chapter 3 and Appendix E, Part 1. Calculation of the counterfactual accounts for each school's rating and prior achievement, both of which were generally higher in non-RF schools, as RF grants were awarded to schools with greatest need within each site. The Actual Mean for Non-Reading First schools value does not take into account either (1) the criteria (or rating) used to determine their RF status or (2) any pre-RF differences in student achievement.

In Exhibit E.4, the first row shows mean scaled scores on the SAT 10 for grade 1 in spring 2005. From left to right, the table displays the actual (or unadjusted) mean for RF schools (541.2), then the estimated mean in the absence of RF (538.9), and in the third column, the actual (or unadjusted) mean for non-RF schools, (542.5). Note that this exhibit does not display the estimated impact of Reading First, which is the presented in the main body of the report (i.e., 2.2 scaled score points, representing the difference between the values in columns 1 and 2).

Exhibit E.4 also includes the corresponding grade equivalent and national percentile for each scaled score mean value.[2] The remaining rows in the table show values for grade 2 and grade 3 (spring 2005), grades 1-3 (spring 2006), and grades 1-3 (spring 2007).

The scaled score means displayed in Exhibit E.4 are graphed in Exhibit E.5. Because the SAT 10 scaled score range is continuous across grades, all values can be shown on a single set of axes. For each grade, the vertical bars represent the average scaled score for RF schools (unadjusted), schools in the absence of RF (estimated), and non-RF schools (unadjusted); the first bar in each set of three represents the mean for spring 2005, the second bar in each set of three represents the mean for spring 2006, and the third bar in

---

[1] All means are weighted by the number of Reading First schools in each site, which is the same weighting scheme used for the impact estimates presented in the main body of this report.

[2] Calculations of mean values were done for scaled scores only. Average scaled scores for Reading First schools and non-Reading First schools were converted to grade equivalents and national percentiles. It is not appropriate to perform arithmetic calculations with grade equivalents or percentiles.

Exhibit E.4: Reading Comprehension Means: Spring 2005, Spring 2006, and Spring 2007

	Actual Mean with Reading First	Estimated Mean without Reading First	Actual Mean for Non-Reading First Schools
All Sites			
Spring 2005			
Grade 1			
Scaled Score	541.2	538.9	542.5
Corresponding Grade Equivalent	*1.7*	*1.7*	*1.7*
Corresponding Percentile	*43*	*41*	*44*
Grade 2			
Scaled Score	583.5	582.4	586.7
Corresponding Grade Equivalent	*2.5*	*2.4*	*2.5*
Corresponding Percentile	*38*	*38*	*41*
Grade 3			
Scaled Score	607.4	609.9	610.7
Corresponding Grade Equivalent	*3.2*	*3.3*	*3.4*
Corresponding Percentile	*38*	*39*	*40*
Spring 2006			
Grade 1			
Scaled Score	545.7	540.4	545.8
Corresponding Grade Equivalent	*1.8*	*1.7*	*1.8*
Corresponding Percentile	*46*	*42*	*46*
Grade 2			
Scaled Score	585.3	583.7	586.0
Corresponding Grade Equivalent	*2.5*	*2.5*	*2.5*
Corresponding Percentile	*40*	*38*	*40*
Grade 3			
Scaled Score	609.5	610.0	613.9
Corresponding Grade Equivalent	*3.3*	*3.3*	*3.5*
Corresponding Percentile	*39*	*39*	*43*
Spring 2007			
Grade 1			
Scaled Score	545.3	537.8	545.8
Corresponding Grade Equivalent	*1.8*	*1.7*	*1.8*
Corresponding Percentile	*46*	*40*	*46*
Grade 2			
Scaled Score	584.8	582.3	585.9
Corresponding Grade Equivalent	*2.5*	*2.4*	*2.5*
Corresponding Percentile	*39*	*38*	*40*
Grade 3			
Scaled Score	610.6	605.1	611.7
Corresponding Grade Equivalent	*3.4*	*3.1*	*3.4*
Corresponding Percentile	*40*	*36*	*41*

NOTES:

The complete Reading First Impact Study (RFIS) sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test scores were not available. For grade 3, in 2007, one RF school could not be included in the analysis because test score data were not available.

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. The actual mean for non-Reading First schools is the observed average for non-Reading First schools in the study sample.

EXHIBIT READS: On average, for first-graders in the spring of 2005, the observed mean reading comprehension score with Reading First was 541.2 scaled score points. The estimated mean without Reading First was 538.9 scaled score points. The observed mean in non-Reading First schools was 542.5 scaled score points.

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).

■ Spring 2005 ■ Spring 2006 □ Spring 2007

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 scores 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.

For each grade, the vertical bars represent the average scaled score for RF schools (unadjusted), schools in the absence of RF (estimated), and non-RF schools (unadjusted).

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).

each set of three represents the mean for spring 2007. Mean values for grade one are the first group of vertical bars, mean values for grade two are the middle group of bars, and mean values for grade three are the last group of bars.

Part 3: Reading Achievement on State Tests

The Reading First Impact Study (RFIS) also examined reading achievement patterns for the RFIS sample schools on state-mandated reading assessments in grade 3 only (the first grade level for which state assessments are required). The study team used extant data from the 2006 administration of state-specific assessments to address the following two questions:

1) What is the impact of Reading First on statewide reading tests in the RFIS core sample?

2) How do results for the RFIS core sample on the SAT 10 reading comprehension subtest compare with results for the RFIS core sample on statewide tests?

Data

The RFIS team used school-level reading performance test score data for elementary grades in all study schools. Most states use criterion-referenced tests and report performance as the percent of students (in a given grade at a given school) who scored at state-defined levels of proficiency (e.g., percent proficient, percent advanced). The extant data available (at the time of these analyses) are current through school year 2005-06, and include data from as early as 1997 for some states. These extant data were merged with the Common Core of Data (CCD), which includes salient demographic and other school-level information.

Analysis

As was the case for the main impact findings reported in Chapter 3, the analyses described below used the regression discontinuity design (RDD) to model achievement outcomes. To be eligible for these analyses, a state's data had to include: (1) at least one year of pre-Reading First reading/English Language Arts (ELA) assessment data, so that a pretest measure could be included in the analyses, (2) 2006 grade 3 reading/ELA assessment scores, and (3) school-level scores based on a percent proficient (or percent advanced, etc.) metric. Two of the 18 sites were excluded from these analyses because their states' data did not meet these three criteria. A total of 210 schools in 16 sites in 11 states were included in these analyses.

The outcome measure for these analyses is percent performing at grade level on the SAT 10. For the statewide reading tests, the outcome measure was percent proficient (on statewide tests); in some states, however, the percent proficient metric indicated ceiling effects, and therefore an alternative benchmark was used (e.g., percent advanced) to allow more meaningful comparisons to the SAT 10 metric.

Results

Exhibit E.6 shows the RDD impacts on state test scores and percent performing at grade level on the SAT 10 for the RFIS core sample. In both cases, the impact estimates are negative (-2.11 percentage points for the state test, -.70 percentage points for the SAT 10) and are not statistically significant.[3]

---

[3] In 2006, the 3rd grade impact estimate on the SAT 10 percent at grade level metric for the *entire* RFIS core sample (244 schools and 18 sites) was -0.9, p=0.80. (See Exhibit 3.3 for year by year SAT 10 results).

Exhibit E.6: Estimated Impacts of Reading First on Grade 3 State Reading/ELA Tests and SAT 10 Reading Comprehension Subtest: 2006

	State Reading/ELA Test *Percent Proficient*	SAT 10 Reading Comprehension *Percent at Grade Level*
Estimate	-2.11	-0.70
Standard Error	2.51	3.51
P-value	(0.401)	(0.843)

NOTES:

The sample includes 210 schools from 16 sites. 104 schools are Reading First and 106 are non-Reading First schools.

Site-by-site impact estimates are weighted by the number of RF schools in each site's core sample to yield the overall impact estimates.

Models include site indicators, interactions of the site indicators with the rating and pre-test variables, and percentage of male students.

EXHIBIT READS: The estimated average impact of Reading First on the percentage of grade 3 students performing at the proficient level on 2006 state reading assessments was -2.11 percentage points with a standard error of 2.51, which was not statistically significant (p=.401).
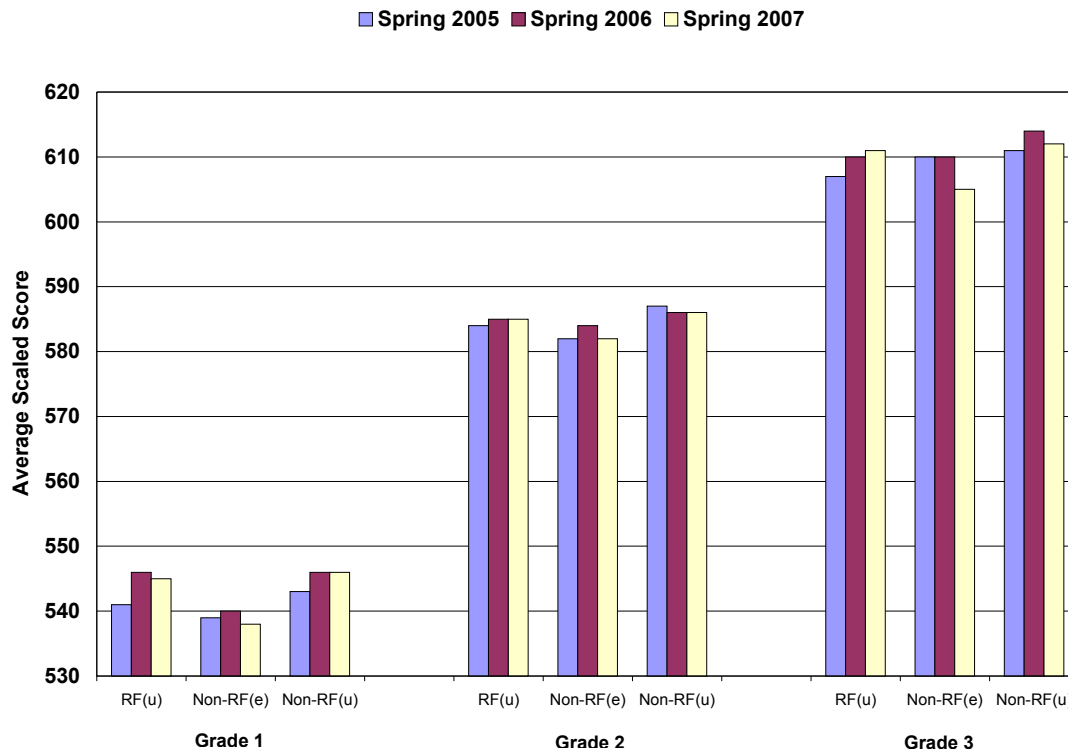
SOURCES: State reading/ELA test data. RFIS SAT 10 administration in the spring of 2006, as well as from state/district education agencies in those sites that already used the SAT 10 for their standardized testing (i.e., FL, KS, MD, OR).

Appendix F: Analysis of Student Exposure to Reading First

Variation in Impacts on Reading Comprehension Based on Student Exposure

Reading First is intended to provide schools with a complete instructional program from kindergarten through grade three. However, some students have not had the opportunity to be exposed to the full range of Reading First instructional practices and support services due to student mobility. As a result, the impact of the program for any one grade level may be muted by the mix of students with varying levels of exposure to the program.

To explore the potential effect of a full three years of exposure to Reading First, the analysis focuses on students in the study sample who were likely to have been enrolled in a Reading First school in grades one, two and three, specifically, those 3rd grade students tested by the study in both 2007 and 2005 who were enrolled in Reading First schools in both 2007 and in 2005. Those students' reading achievement scores are then compared with their 3rd grade counterparts in non-Reading First schools.[1] Impact estimates from this sample of students may provide an indication of the effect of Reading First for students who received three years of exposure to Reading First.

These results should be interpreted with caution, however, because the impact estimates may be biased if Reading First caused a difference either in student mobility rates or in the types of students who move from or stay at the same school. For example, if Reading First induced higher achieving students to remain in the same school while similar students were more likely to leave a non-Reading First school, positive impact estimates may be an artifact of differential mobility patterns rather than real improvements in reading comprehension for the "remaining" students. Consequently, before presenting the test score impact findings for students with three years of exposure to Reading First, these analytic issues must be discussed.

Analysis of the proportion of 3rd grade students from the 2007 data collection who were in the study sample in 2005 provides evidence that Reading First did not appear to affect overall mobility rates. Exhibit F.1 shows the observed percentage of 3rd grade students from the 2007 data collection effort who were in a school with the same treatment status in 2005,[2] by Reading First status and by site.

---

[1] In the spring of 2005, the study tested in all classrooms in study schools; in subsequent waves of testing, the study tested only a subsample of classrooms in those schools with more than an average of three classrooms per grade. Because not all classrooms per grade level were tested in both 2006 and 2007, we cannot use 2006 data to identify how many 3rd graders in our 2007 sample were in study schools in 2006.

[2] The percentages presented in Exhibits F.1 and F.2 do necessarily reflect actual mobility patterns. It is possible that some of these students were enrolled in the same school in 2005 and 2007 but in a different school in 2006. Also, the sample of students included in this analysis are those who were tested as 3rd graders in a given school in 2007 and were also enrolled in a study school with the same program status (Reading First or non-Reading First) in 2005. Because not all 3rd grade students were tested in all of the study schools in 2007, this sample may not encompass all students who remained in the same type of school for three years.

Exhibit F.1: Percentage of Third Graders in Same Treatment Status for Three Years by Site and Treatment Status

	Percentage of 3rd Graders in Reading First Schools in 2007 and in 2005	Percentage of 3rd Graders in Non-Reading First Schools in 2007 and in 2005	Overall Percentage of 3rd Graders in Same Treatment Status in 2007 and 2005
Site 1	64	41	52
Site 2	64	67	66
Site 3	53	54	53
Site 4	73	71	72
Site 5	42	41	42
Site 6	51	75	63
Site 7	59	63	61
Site 8	72	71	72
Site 9	55	61	59
Site 10	67	68	67
Site 11	70	54	61
Site 12	65	59	63
Site 13	51	54	52
Site 14	61	63	62
Site 15	63	74	69
Site 16	56	56	56
Site 17	66	65	66
Site 18	70	65	67
Overall	60	60	60

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

This exhibit shows unweighted means for each site, unlike exhibits that present impact estimates in which means for each site are weighted by the number of Reading First schools in the site.

The percentages presented above do necessarily reflect actual mobility patterns. It is possible that some of these students were enrolled in the same school in 2005 and 2007 but in a different school in 2006. Also, the sample of students included in this analysis are those who were tested as 3rd graders in a given school in 2007 and were also enrolled in a study school with the same program status (Reading First or non-Reading First) in 2005. Because not all 3rd grade students were tested in all of the study schools in 2007, this sample may not encompass all students who remained in the same type of school for three years.

EXHIBIT READS: In Site 1, 64 percent of third grade students in the study sample in Reading First schools in 2007 were also in Reading First schools in 2005. Forty-one percent of third graders in the study sample in non-Reading First schools in 2007 were in non-Reading First schools in 2005. Overall, in Site 1, 52 percent of third grade students in the study sample in 2007 had the same treatment status in 2005.

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).

The average mobility rate across all sites is 60 percent, and it ranges from 41 to 75 percent. Exhibit F.2 shows the overall estimated differences in the percentage of students who were enrolled in schools of the same treatment status in 2005 and 2007. Panel 1 shows a regression adjusted difference, including adjustments for the regression discontinuity design. Panel 2 is not regression adjusted.

Exhibit F.2: Estimated Regression Adjusted and Unadjusted Impacts of Reading First on the Percent of Students With Three Years of Exposure to the Same Treatment Status, 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)
Panel 1					
Percent of Students With Three Years of the Same Treatment Status (%)					
Regression Adjusted	60.8	56.4	4.4	N/A	(0.278)

	Actual Mean with Reading First	Actual Mean for Non-Reading First Schools	Impact	Effect Size of Impact	Statistical Significance of Impact (p-value)
Panel 2					
Percent of Students With Three Years of the Same Treatment Status (%)					
Unadjusted	60.8	61.5	-0.7	N/A	(0.696)

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.

In panel 1, the value in the "Actual Mean with Reading First" column is the actual, unadjusted value for Reading First schools; the value in the "Estimated Mean without Reading First" column represents the best estimates of what would have happened in RF schools absent RF funding and is calculated by subtracting the impact estimate from the RF schools' actual mean values.

In panel 2, the value in the "Actual Mean with Reading First" column is the actual, unadjusted value for Reading First schools; the "Actual Mean for Non-Reading First Schools" is the actual, unadjusted value for non-Reading First schools in the study sample and is calculated by subtracting the impact estimate from the RF schools' actual mean values.

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

EXHIBIT READS: The observed average percent of third-grade students in Reading First schools that were in a Reading First school two years earlier is 60.8 percent. The estimated average percent of third-grade students in non-Reading First Schools in non-Reading First schools two years earlier is 56.4 percent. The impact of Reading First on the percent of third-grade students with three years of exposure to the same treatment status is 4.4 percentage points, which is not statistically significant (p=.278).

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).

- There was a 4.4 percentage point difference in the number of students who stayed in a Reading First versus a non-Reading First school for three years. The difference was not statistically significant, indicating that Reading First did not have a systematic impact on whether or not a student stayed in a school with the same treatment status over time. Also, although Exhibit F.1 indicates that there was variation in the observed difference between Reading First and non-Reading First schools on this percentage (ranging from +23 percentage points in site 1 to -24 percentage points in site 6), an F-test indicates that the variation is not statistically significant.

It is also important to assess the extent to which the Reading First program may have influenced the compositional mix of students who were enrolled in RF or non-RF schools, among schools whose RF or non-RF status did not change between 2005 and 2007. 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.

Exhibit F.3 (which is identical to Exhibit 3.4) shows the estimated impacts of Reading First on the subsample of students who remained in either RF or non-RF study schools in both 2005 and 2007.

- The results indicate that the impact of the program on grade three students with three years of exposure was 4.3 scaled score points and was not statistically significant.

Exhibit F.3: 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≤.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).

Appendix G: Subgroup Analyses

This appendix updates analyses completed in the RFIS Interim Report on two subgroups of sites that were identified by the study team on the basis of when they received their initial Reading First grants.[1,2] The analyses below mirror those from the Interim Report for the two main outcomes of reading instruction and reading comprehension updated with an additional school year (2006-07) of data.[3]

The intent of assessing impacts separately for these two groups of sites was to explore whether sites with a longer implementation period (those that received Reading First funding earlier in the grant making period) produced larger impacts than sites with a shorter implementation period (those that received Reading First funding later). 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.

Part 1: Subgroup Impacts over Time

Exhibits G.1-G.5 provide impact estimates separately by award group across follow-up years for reading comprehension and minutes in the five dimensions (impact estimates presented in the main body of the report are for the pooled full sample and not by award group).

The first set of analyses in this section explores the year-to-year changes in impacts as Reading First schools gained experience with the program (see Exhibits G.1 and G.2). The study's follow-up periods represent, roughly, years 1 to 3 of program funding for late award sites and years 2 to 4 for early award sites. By separating annualized findings for the two groups, one can see the findings in the context of a specific year of program implementation. Findings indicate that:

- In the early award sites, there is an overall pattern of year-to-year increases in impact estimates for reading comprehension scaled scores (Exhibit G.1). This pattern was statistically significant when grades were pooled, but not for any single grade (Exhibit G.2). In the late award sites, impact estimates increased less consistently over time, and the overall pattern was not statistically significant for the three grades pooled together or for any single grade.

---

[1] See pp. 51-63 in Gamse et al. (2008).

[2] Due to inconsistent availability of information on the date of receipt of RF funding at the school level, all schools within a site were assigned the date at which their site received RF funding from the state.

[3] For this report, analyses were not conducted for the Student Engagement with Print measure.

Exhibit G.1: Estimated Impacts on Reading Comprehension and Minutes in the Five Dimensions, by Implementation Year, Calendar Year, and Award Status

	Implementation Year							
	Year 1		**Year 2**		**Year 3**		**Year 4**	
	Impact	**(p-value)**	**Impact**	**(p-value)**	**Impact**	**(p-value)**	**Impact**	**(p-value)**
Panel 1								
Early Award Sites	**2004**		**2005**		**2006**		**2007**	
Grade 1								
Percent reading at or above grade level (%)	N/A	N/A	-2.6	(0.708)	-1.9	(0.751)	6.8	(0.324)
Instruction in five dimensions (minutes)	N/A	N/A	5.49	(0.376)	4.26	(0.448)	1.00	(0.850)
Grade 2								
Percent reading at or above grade level (%)	N/A	N/A	-8.2	(0.163)	-6.8	(0.303)	-0.2	(0.970)
Instruction in five dimensions (minutes)	N/A	N/A	10.90	(0.079)	3.88	(0.478)	2.13	(0.685)
Grade 3								
Percent reading at or above grade level (%)	N/A	N/A	-9.9	(0.110)	-7.7	(0.225)	3.6	(0.588)
Instruction in five dimensions (minutes)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Panel 2								
Late Award Sites	**2005**		**2006**		**2007**			
Grade 1								
Percent reading at or above grade level (%)	6.3	(0.077)	9.4*	(0.024)	5.8	(0.156)	N/A	N/A
Instruction in five dimensions (minutes)	11.51*	(0.001)	12.21*	(0.003)	9.79*	(0.006)	N/A	N/A
Grade 2								
Percent reading at or above grade level (%)	6.3*	(0.028)	5.7	(0.155)	7.3*	(0.049)	N/A	N/A
Instruction in five dimensions (minutes)	14.84*	(<0.001)	16.26*	(<0.001)	9.94*	(0.004)	N/A	N/A
Grade 3								
Percent reading at or above grade level (%)	1.7	(0.537)	4.2	(0.269)	7.4*	(0.035)	N/A	N/A
Instruction in five dimensions (minutes)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

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.

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

EXHIBIT READS: The estimated impact of Reading First on percent of students reading at or above grade level in grade 1, early award sites in 2005 was -2.6 percentage points, which was not statistically significant (p=.708).

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); and RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, spring 2006, fall 2006 and spring 2007.

Exhibit G.2: Change Over Time in Program Impact on Reading Comprehension and Instruction, By Award Status

Panel 1 Early Award Sites		Reading Comprehension (SAT 10 Scaled Score)	Reading Instruction (min. in 5 Dimensions)
Grade 1	Linear Year-to-Year Change	5.55	-3.98
	SE	3.61	2.17
	p-value	(0.123)	(0.066)
	F-test for overall variation across years	1.537	0.19
	p-value	(0.215)	(0.660)
Grade 2	Linear Year-to-Year Change	0.21	-3.02
	SE	3.28	2.12
	p-value	(0.948)	(0.154)
	F-test for overall variation across years	0.013	2.52
	p-value	(0.987)	(0.113)
Grade 3	Linear Year-to-Year Change	5.94	n.a.
	SE	3.33	n.a.
	p-value	(0.074)	n.a.
	F-test for overall variation across years	1.834	n.a.
	p-value	(0.160)	n.a.
All Available Grades[a]	Linear Year-to-Year Change	3.96*	-3.05*
	SE	1.98	1.52
	p-value	(0.045)	(0.045)
	F-test for overall variation across years	2.708	1.64
	p-value	(0.067)	(0.201)

Exhibit G.2: Change Over Time in Program Impact on Reading Comprehension and Instruction, By Award Status (continued)

Panel 2 Late Award Sites		Reading Comprehension (SAT 10 Scaled Score)	Reading Instruction (min. in 5 Dimensions)
Grade 1	Linear Year-to-Year Change	-0.25	-1.50
	SE	2.34	1.38
	p-value	(0.913)	(0.278)
	F-test for overall variation across years	0.909	1.46
	p-value	(0.403)	(0.228)
Grade 2	Linear Year-to-Year Change	0.28	-2.78
	SE	1.92	1.49
	p-value	(0.884)	(0.063)
	F-test for overall variation across years	0.044	2.49
	p-value	(0.957)	(0.115)
Grade 3	Linear Year-to-Year Change	2.19	n.a.
	SE	1.80	n.a.
	p-value	(0.223)	n.a.
	F-test for overall variation across years	0.833	n.a.
	p-value	(0.435)	n.a.
All Available Grades [a]	Linear Year-to-Year Change	0.87	-1.83
	SE	1.17	1.02
	p-value	(0.458)	(0.073)
	F-test for overall variation across years	0.590	2.80
	p-value	(0.554)	(0.094)

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, and where applicable, statistically significant findings at the p≤.05 level are indicated by *.

EXHIBIT READS: For grade 1, the program impact on reading comprehension in early award sites increases by 5.55 scaled score points per year between 2005 and 2007. This change was not statistically significant (p=.123). The program impact on instruction in the five dimensions of reading instruction decreases by -3.98 minutes per year. This change was not statistically significant (p=.066).

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.

- In the early award sites, there is an overall pattern of year-to-year decreases in the impact estimates for instructional time in the five dimensions (Exhibit G.1). This pattern was statistically significant for the pooled sample of first and second grade teachers, but not for either grade alone (Exhibit G.2). In the late award sites, there was no consistent pattern over time and the year-to-year variation in impacts on instructional time was not statistically significant for both grades pooled together or for any single grade.

The second set of analyses examines the three-year average impacts of Reading First on instructional time in the five dimensions and reading comprehension scaled scores (see Exhibits G.3, G.4 and G.5, which are identical to Exhibits 4.8, 4.9, and 4.10). These analyses explore whether the differences in impacts between the two groups of sites were statistically significant.

- In the early award sites, estimated impacts on instruction in the five dimensions were not statistically significant. In late award sites, estimated impacts on instructional time in the five dimensions were positive and statistically significant. Differences between early and late award sites in their estimates of impacts on instructional time in the five dimensions were not statistically significant.

- In the early award sites, estimated impacts on reading comprehension were not statistically significant. In late award sites, the estimated impacts on reading comprehension scaled scores are positive and statistically significant only for grade two. Differences between early and late award sites in their estimates of impacts on reading comprehension scaled scores were statistically significant for grade two, but not for grade one or grade three.

Part 2: Linear Interactions between Program Impacts and Site Characteristics

Exhibit G.6 presents updated results regarding the change in impact associated with one unit change in three characteristics that distinguish the early award sites from the late award sites: (1) the number of months each site had access to its Reading First grant; (2) the amount of Reading First funding allocated per K-3 student in Reading First schools; and (3) the levels of reading comprehension exhibited by students in non-Reading First schools in fall 2004. Relevant information is presented in the next section below about the site characteristics used in these analyses.

The information in Exhibit G.7 indicates that:

- On average, late award sites allocated more Reading First funding per school and per student than did early award sites. Hence, there may have been a greater concentration of resources to produce change in the late award sites.

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

Exhibit G.3: 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≤.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 G.4: 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≤.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 G.5: 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≤.05 level are indicated by *.

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.

Exhibit G.6: Change in Impact Associated with One Unit of Change In Continuous Dimensions

		Reading Comprehension (SAT 10 scaled score)	Reading Instruction (min. in 5 Dimensions)
Award Date			
Grade 1	Impact	0.12	-0.14
	SE	0.13	0.09
	p-value	(0.375)	(0.096)
Grade 2	Impact	-0.02	-0.20*
	SE	0.11	0.09
	p-value	(0.859)	(0.032)
Grade 3	Impact	0.12	n.a.
	SE	0.11	n.a.
	p-value	(0.281)	n.a.
Fall 2004 Reading Performance of non-RF Schools			
Grade 1	Impact	0.22	0.18
	SE	0.24	0.23
	p-value	(0.348)	(0.431)
Grade 2	Impact	-0.20	0.28
	SE	0.20	0.23
	p-value	(0.303)	(0.226)
Grade 3	Impact	-0.10	n.a.
	SE	0.18	n.a.
	p-value	(0.590)	n.a.
Reading First Funding Per Student			
Grade 1	Impact	0.02*	0.01
	SE	0.01	0.01
	p-value	(0.039)	(0.390)
Grade 2	Impact	0.02*	0.01
	SE	0.01	0.01
	p-value	(0.011)	(0.191)
Grade 3	Impact	0.01	n.a.
	SE	0.01	n.a.
	p-value	(0.275)	n.a.

NOTES:

The complete Reading First Impact Study (RFIS) sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2, one non-RF school could not be included in the analysis because test score data were not available. For grade 3 in 2007, one RF school could not be included in the analysis because test score data were not available. There are 8 late award sites totaling 137 schools and 10 early award sites totaling 111 schools. There are 10 high performance non-RF school sites totaling 120 schools, and 8 low performance non-RF sites totaling 128 schools. There are 8 low RF funding sites totaling 126 schools and 10 high RF funding sites totaling 122 schools.

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

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

EXHIBIT READS: An increase of one month in Reading First award date in grade 1 is associated with an increase of 0.12 scaled score points in reading comprehension and a decrease of -0.14 minutes of instruction in the five dimensions. Neither of these impacts was statistically significant.

SOURCES: RFIS SAT 10 administration in the spring of 2005, 2006, 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); and RFIS Instructional Practice in Reading Inventory, spring 2005, fall 2005, spring 2006, fall 2006, and spring 2007.

Exhibit G.7: Characteristics of Early and Late Award Sites

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

Notes:

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

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

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

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

Part 3: Impact Estimates for Subgroups Defined by Site Characteristics

Award Date

The award date information was obtained from Reading First district coordinators in November 2005. District coordinators were asked to provide the month and year that Reading First money was made available to schools in their respective districts. The continuous award variable was then calculated as the number of months between the month/year the funds became available to each site and January 2003. For example, if the funds became available to a site in April of 2003, the continuous award variable for that site would be 3. When dividing the sample into two groups, the study grouped those sites that received funding April and December 2003 as early award sites, and those sites funded between January and August 2004 as late award sites.

Fall 2004 Reading Performance of the non-Reading First Schools

Fall 2004 reading performance for students in the non-RF schools represents the best approximation of existing student reading proficiency in each site. This variable draws on test score data from fall 2004,

which is up to 16 months after the RF award date in early award sites, and prior to the RF award date in all late award sites. The percent of students in grades 1-3 at or above grade level variable was constructed using students' fall 2004 SAT 10 scaled scores,[4] as well as the test date at each school. Each student's scaled score was compared to corresponding grade equivalency norms to determine whether the student was at or above grade level. The percent of non-Reading First students at or above grade level was created by taking the mean of the student-level at or above grade level variable, across all grades within a school, and averaging across all schools within a site. To create two subgroups of sites, the study team ordered the sites according to non-RF school performance, taking into account the number of RF schools in each site, and then creating two subgroups as equivalent as possible with respect to the number of schools. This resulted in 120 schools in the high performance group (at or above 48.57 percent of students at grade level) and 128 schools in the low performance group (at or below 48.02 percent of students at grade level).

Reading First Funding Per Student

The amount of the Reading First funding per student was constructed using data from the SEDL database[5] (as of October 2004) about award amounts for each site, and the Common Core Data that provided the number of K-3 students within each school. The Reading First funding per pupil was calculated separately for the 2002-2003 and 2003-2004 school years. Since a portion of the funds were set aside for district and state Reading First activities, and therefore not used to directly fund schools, the grant amounts were adjusted to reflect the proportion of funding set aside (3.5 percent of the Reading First grant to each district). The award amount was then divided by the number of students in grades K-3 in all Reading First schools per site. The Reading First funding per pupil for the two school years was then averaged by site to create the Reading First per pupil expenditure variable used in analysis. To create two subgroups of sites, the study team ordered the sites according to funding, taking into account the number of RF schools in each site, and then creating two subgroups as equivalent as possible with respect to number of schools. This resulted in 126 schools in the high funding (sites with a per-pupil Reading First grant amount at or above $635), and 122 schools in the low funding group (sites with a per-pupil Reading First grant amount at or below $513).

For each of the three site characteristics described above, study sites were ordered and then separated into two subgroups of sites that were as balanced as possible, with respect to the number schools in each group. For each characteristic, the order of sites was slightly different. Therefore, the composition of the two subgroups for each moderating factor differed both in the actual sites included and in the total number of schools included. Program impacts were then estimated for one key outcome measure from each of the three domains for the two subgroups. These outcomes included (a) the SAT 10 scaled score for reading comprehension and (b) total minutes in the five dimensions of reading instruction. First, analyses tested the difference between impacts for the two subgroups. Then, to test whether the conclusions were sensitive to the specific cut-point chosen to define the subgroups, average impacts were re-estimated for each subgroup after dropping the two sites closest to the cut-point between the two

---

[4] In the fall of 2004, students' SAT 10 scores were unavailable. For those sites scores from the spring of 2005 were substituted and adjusted by the mean difference of all other students' spring and fall SAT 10 scores, by grade.

[5] Southwest Educational Development Laboratory (SEDL) was contracted to maintain the Reading First Awards database available online at http://www.sedl.org/readingfirst/welcome.html. SEDL lists the amount awarded to each Reading First district in the first year. State Reading First Coordinators are responsible for providing this information to SEDL.

subgroups. This was repeated again after dropping the next two sites closest to the cut-point between the two subgroups.

As reported in the study's Interim Report, over the first two of the study's three data collection years (2004-05 and 2005-06), there were statistically significant differences in impacts between early and late award sites for some outcome variables, but not others. There were not statistically significant differences in estimated impacts between early and late award sites for classroom instruction or student engagement with print. However, there were statistically significant differences in impacts between early and late award sites on reading comprehension for grades two and three (but not for grade one). For more detailed information, see Appendix H of the Interim Report (Gamse, Bloom, Kemple & Jacob, 2008). The final set of exploratory analyses presented below provide a follow up to Interim Report analyses conducted to test possible explanations for the award date differences found in the earlier report (see Exhibits G.8 - G.13).

- The relationship between the numbers of months of access to funding and impact on student achievement are not statistically significant in any grade. The relationship between months of funding access and impact on instruction in the five dimensions of reading was not significant in grade 1 and was significant in grade 2.

- The relationship between fall 2004 reading achievement of students in non-Reading First schools and impacts on student achievement or reading instruction are not statistically significant.

- The results indicate that sites with higher allocations of Reading First funds per K-3 student had larger program impacts on student achievement than did sites with lower allocations. This relationship was statistically significant for grades one and two. The relationship between funding and impacts on reading instruction was not statistically significant.

Exhibit G.8: Estimated Impacts on Reading Comprehension, by Award Status

SAT 10 Scaled Scores		Full	Drop 1 Pair	Drop 2 Pairs
Early Award Sites				
Grade 1	Impact	2.87	2.52	6.37
	SE	5.01	5.23	5.39
	p-value	(0.567)	(0.630)	(0.237)
Grade 2	Impact	-4.36	-5.37	-3.93
	SE	4.05	4.25	4.02
	p-value	(0.282)	(0.207)	(0.328)
Grade 3	Impact	-3.89	-5.09	-3.03
	SE	4.09	4.30	4.11
	p-value	(0.342)	(0.236)	(0.462)
Late Award Sites				
Grade 1	Impact	5.64	5.08	2.24
	SE	2.98	3.20	3.58
	p-value	(0.058)	(0.112)	(0.531)
Grade 2	Impact	6.05*	5.00	5.72
	SE	2.59	2.75	3.32
	p-value	(0.019)	(0.070)	(0.085)
Grade 3	Impact	3.52	2.75	2.75
	SE	2.17	2.25	2.47
	p-value	(0.105)	(0.221)	(0.265)
Difference				
Grade 1	Impact	-2.77	-2.57	4.12
	SE	5.83	6.13	6.47
	p-value	(0.636)	(0.676)	(0.525)
Grade 2	Impact	-10.41*	-10.36*	-9.66
	SE	4.81	5.07	5.21
	p-value	(0.032)	(0.042)	(0.066)
Grade 3	Impact	-7.41	-7.85	-5.78
	SE	4.63	4.85	4.80
	p-value	(0.111)	(0.108)	(0.231)

Notes:

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2 in 2006, one non-RF school could not be included in the analysis because test scores 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. There are 8 late award sites totaling 137 schools and 10 early award sites totaling 111 schools.

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

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

EXHIBIT READS: The impact of the Reading First program in early award sites for grade 1 on reading comprehension was
2.87 scaled score points, on average, for the full sample of 10 early award sites. The impact was not statistically significant (p=.567). The impact of the Reading First program in early award sites for grade 1 on reading comprehension was 2.52 scaled score points, on average, for the sample of 9 early award sites remaining after one pair of sites closest to the cut-point was dropped. The impact was not statistically significant (p=.630). The impact of the Reading First program in early award sites for grade 1 on reading comprehension scaled score was 6.37 scaled score points, on average, for the sample of 8 early award sites remaining after two pairs of sites closest to the cut-point were dropped. The impact was not statistically significant (p=.237).

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).

Exhibit G.9: Estimated Impacts on Reading Instruction, by Award Status

Minutes in Five Dimensions		Full	Drop 1 Pair	Drop 2 Pairs
Early Award Sites				
Grade 1	Impact	2.02	0.82	0.02
	SE	4.30	4.52	4.76
	p-value	(0.640)	(0.856)	(0.996)
Grade 2	Impact	5.55	5.45	6.28
	SE	4.52	4.77	5.07
	p-value	(0.223)	(0.256)	(0.219)
Late Award Sites				
Grade 1	Impact	10.74*	9.65*	6.70
	SE	2.85	3.07	3.40
	p-value	(<0.001)	(0.002)	(0.052)
Grade 2	Impact	13.08*	11.25*	9.99*
	SE	2.73	2.83	3.27
	p-value	(<0.001)	(<0.001)	(0.003)
Difference				
Grade 1	Impact	-8.72	-8.83	-6.68
	SE	5.16	5.46	5.85
	p-value	(0.092)	(0.107)	(0.256)
Grade 2	Impact	-7.53	-5.79	-3.71
	SE	5.28	5.55	6.03
	p-value	(0.155)	(0.297)	(0.540)

NOTES:

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

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

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

EXHIBIT READS: The impact of the Reading First program in early award sites for grade 1 on the number of minutes of instruction in the five dimensions was 2.02 minutes on average for the full sample of 10 early award sites. The impact was not statistically significant (p=.640). The impact of the Reading First program in early award sites for grade 1 on the number of minutes of instruction in the five dimensions was 0.82 minutes on average for the sample of 9 early award sites remaining after one pair of sites closest to the cut-point was dropped. The impact was not statistically significant (p=.856). The impact of the Reading First program in early award sites for grade 1 on the number of minutes of instruction in the five dimensions was 0.02 minutes on average for the sample of 80 early award sites remaining after two pairs of sites closest to the cut-point were dropped. The impact was not statistically significant (p=.996).

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

Exhibit G.10: Estimated Impacts on Reading Comprehension, by Fall 2004 Reading Performance of the non-Reading First Schools

SAT 10 Scaled Scores		Full	Drop 1 Pair	Drop 2 Pairs
High Non-RF School Performance				
Grade 1	Impact	8.26	6.44	9.93*
	SE	4.84	4.00	3.91
	p-value	(0.088)	(0.108)	(0.011)
Grade 2	Impact	-1.85	0.73	2.69
	SE	3.72	3.04	2.72
	p-value	(0.619)	(0.810)	(0.324)
Grade 3	Impact	-0.81	0.20	2.61
	SE	3.73	2.98	2.76
	p-value	(0.828)	(0.946)	(0.344)
Low Non-RF School Performance				
Grade 1	Impact	0.98	2.64	3.27
	SE	3.03	3.32	3.35
	p-value	(0.747)	(0.428)	(0.331)
Grade 2	Impact	5.13	5.54	5.27
	SE	2.76	3.08	3.14
	p-value	(0.063)	(0.075)	(0.096)
Grade 3	Impact	1.21	1.16	3.01
	SE	2.35	2.70	2.75
	p-value	(0.607)	(0.670)	(0.277)
Difference				
Grade 1	Impact	-7.28	-3.79	-6.66
	SE	5.71	5.20	5.15
	p-value	(0.204)	(0.467)	(0.198)
Grade 2	Impact	6.98	4.81	2.59
	SE	4.63	4.33	4.16
	p-value	(0.133)	(0.268)	(0.535)
Grade 3	Impact	2.02	0.95	0.40
	SE	4.41	4.02	3.90
	p-value	(0.648)	(0.813)	(0.919)

NOTES:

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2 in 2006, one non-RF school could not be included in the analysis because test scores 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. There are 10 high non-RF comparison school sites totaling 120 schools and 8 low performance non-RF school sites totaling 128 schools.

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

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

EXHIBIT READS: The impact of the Reading First program in high performance non-RF school sites for grade 1 on reading comprehension was 8.26 scaled score points on average for the full sample of 10 high non-RF comparison school sites. The impact was not statistically significant (p=.088). The impact of the Reading First program in high performance non-RF school sites for grade 1 on reading comprehension was 6.44 scaled score points on average for the sample of 9 high non-RF comparison school sites remaining after one pair of sites closest to the cut-point was dropped. The impact was not statistically significant (p=.108). The impact of the Reading First program in high performance non-RF school sites for grade 1 on average reading comprehension scaled score was 9.93 scaled score points on average for the sample of 8 high non-RF comparison school sites remaining after two pairs of sites closest to the cut-point were dropped. The impact was statistically significant (p=.011).

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).

Exhibit G.11: Estimated Impacts on Reading Instruction, by Fall 2004 Reading Performance of the Non-Reading First Schools

Minutes in Five Dimensions		Full	Drop 1 Pair	Drop 2 Pairs
High Non-RF School Performance				
Grade 1	Impact	9.76*	12.96*	13.60*
	SE	3.93	3.59	3.64
	p-value	(0.015)	(0.001)	(<0.001)
Grade 2	Impact	10.14*	13.16*	14.84*
	SE	3.87	3.41	3.45
	p-value	(0.010)	(<0.001)	(<0.001)
Low Non-RF School Performance				
Grade 1	Impact	4.18	5.23	6.14
	SE	3.05	3.29	3.32
	p-value	(0.173)	(0.115)	(0.068)
Grade 2	Impact	9.36*	8.65*	9.86*
	SE	3.12	3.47	3.55
	p-value	(0.003)	(0.015)	(0.007)
Difference				
Grade 1	Impact	-5.57	-7.73	-7.46
	SE	4.97	4.87	4.92
	p-value	(0.264)	(0.114)	(0.131)
Grade 2	Impact	-0.78	-4.51	-4.98
	SE	4.98	4.87	4.95
	p-value	(0.875)	(0.355)	(0.316)

NOTES:

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. There are 10 high performance non-RF school sites totaling 120 schools, and 8 low performance non-RF school sites totaling 128 schools.

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

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

EXHIBIT READS: The impact of the Reading First program in high performance non-RF school sites for grade 1 on the number of minutes of instruction in the five dimensions was 9.76 minutes on average for the full sample of 10 high non-RF comparison school sites. The impact was statistically significant (p=.015). The impact of the Reading First program in high performance non-RF school sites for grade 1 on the number of minutes of instruction in the five dimensions was 12.96 minutes on average for the sample of 9 high non-RF comparison school sites remaining after one pair of sites closest to the cut-point was dropped. The impact was statistically significant (p=.001). The impact of the Reading First program in high performance non-RF school sites for grade 1 on the number of minutes of instruction in the five dimensions was 13.60 minutes on average for the sample of 8 high non-RF comparison school sites remaining after two pairs of sites closest to the cut-point were dropped. The impact was statistically significant (p<.001).

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

Exhibit G.12: Estimated Impacts on Reading Comprehension, by Reading First Funds Per Student

SAT 10 Scaled Score		Full	Drop 1 Pair	Drop 2 Pairs
Low RF Funding				
Grade 1	Impact	0.05	-1.60	-3.22
	SE	3.87	3.87	3.96
	p-value	(0.990)	(0.680)	(0.416)
Grade 2	Impact	-4.42	-2.33	-3.19
	SE	3.12	3.08	3.28
	p-value	(0.156)	(0.449)	(0.329)
Grade 3	Impact	-2.53	-0.80	-1.89
	SE	3.03	2.97	3.10
	p-value	(0.404)	(0.788)	(0.543)
High RF Funding				
Grade 1	Impact	8.76*	8.95*	5.92
	SE	4.22	4.30	4.73
	p-value	(0.038)	(0.037)	(0.211)
Grade 2	Impact	9.11*	10.06*	9.98*
	SE	3.97	3.96	4.47
	p-value	(0.022)	(0.011)	(0.026)
Grade 3	Impact	0.77	0.72	-0.92
	SE	3.06	2.97	3.27
	p-value	(0.800)	(0.808)	(0.779)
Difference				
Grade 1	Impact	-8.71	-10.54	-9.14
	SE	5.72	5.78	6.17
	p-value	(0.130)	(0.070)	(0.141)
Grade 2	Impact	-13.53*	-12.40*	-13.17*
	SE	5.05	5.02	5.54
	p-value	(0.008)	(0.015)	(0.019)
Grade 3	Impact	-3.30	-1.52	-0.97
	SE	4.30	4.20	4.51
	p-value	(0.444)	(0.718)	(0.830)

NOTES:

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. For grade 2 in 2006, one non-RF school could not be included in the analysis because test scores 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. There are 8 low RF funding sites totaling 126 schools and 10 high RF funding sites totaling 122 schools.

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

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

EXHIBIT READS: The impact of the Reading First program in low Reading First funding sites for grade 1 on reading comprehension was 0.05 scaled score points on average for the full sample of 8 low RF funding sites. The impact was not statistically significant (p=.990). The impact of the Reading First program in low Reading First funding sites for grade 1 reading comprehension scaled score was –1.60 scaled score points on average for the sample of 7 low RF funding sites remaining after one pair of sites closest to the cut-point was dropped. The impact was not statistically significant (p=.680). The impact of the Reading First program in low Reading First funding sites for grade 1 on average reading comprehension scaled score was –3.22 scaled score points on average for the sample of 6 low RF funding sites remaining after two pairs of sites closest to the cut-point were dropped. The impact was not statistically significant (p=.416).

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).

Exhibit G.13: Estimated Impacts on Reading Instruction, by Reading First Funds Per Student

Minutes in Five Dimensions		Full	Drop 1 Pair	Drop 2 Pairs
Low RF Funding				
Grade 1	Impact	3.28	5.82	1.39
	SE	3.56	3.26	3.77
	p-value	(0.359)	(0.077)	(0.714)
Grade 2	Impact	4.86	7.23*	4.51
	SE	3.57	3.15	3.82
	p-value	(0.177)	(0.024)	(0.241)
High RF Funding				
Grade 1	Impact	10.66*	9.35*	10.15*
	SE	3.33	3.70	3.73
	p-value	(0.002)	(0.013)	(0.008)
Grade 2	Impact	14.95*	13.16*	12.36*
	SE	3.37	3.66	3.75
	p-value	(<.001)	(<.001)	(0.001)
Difference				
Grade 1	Impact	-7.38	-3.52	-8.76
	SE	4.88	4.94	5.31
	p-value	(0.132)	(0.476)	(0.101)
Grade 2	Impact	-10.10*	-5.93	-7.84
	SE	4.91	4.83	5.36
	p-value	(0.041)	(0.221)	(0.145)

NOTES:

The complete RF study sample includes 248 schools from 18 sites (17 school districts and 1 state) located in 13 states. 125 schools are Reading First schools and 123 are non-Reading First schools. There are 8 low RF funding sites totaling 126 schools and 10 high RF funding sites totaling 122 schools.

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

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

EXHIBIT READS: The impact of the Reading First program in low Reading First funding sites for grade 1 on the number of minutes of instruction in the five dimensions was 3.28 minutes on average for the full sample of 8 low RF funding sites. The impact was not statistically significant (p=.359). The impact of the Reading First program in low Reading First funding sites for grade 1 on the number of minutes of instruction in the five dimensions was 5.82 minutes on average for the sample of 7 low RF funding sites remaining after one pair of sites closest to the cut-point was dropped. The impact was not statistically significant (p=.077). The impact of the Reading First program in low Reading First funding sites for grade 1 on the number of minutes of instruction in the five dimensions was 1.39 minutes on average for the sample of 6 low RF funding sites remaining after two pairs of sites closest to the cut-point were dropped. The impact was not statistically significant (p=.714).

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

References

Abt Associates and RMC Research (2002). Classroom Observation Record. Unpublished instrument.

Armbruster, B.B., Lehr, F., and Osborn, J. (2003). *Put reading first: The research building blocks for teaching children to read* (2nd Ed.). Developed by the Center for the Improvement of Early Reading Achievement (CIERA). Washington, DC: The National Institute for Literacy (NIFL).

Ball, E.W., and Blachman, B.A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly, 26*, 49–66.

Beck, I.L., Perfetti, C.A., and McKeown, M.G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*, 506-521.

Bloom, H., Kemple, J.K., and Gamse, B.C. (2004). Using regression discontinuity analysis to measure the impacts of Reading First. Paper prepared for the Institute of Education Sciences.

Bloom, H.S. (2001). *Measuring the impacts of whole-school reforms: Methodological lessons from an evaluation of accelerated schools.* New York: MDRC.

Bloom, H.S. (1995). "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs," *Evaluation Review*, Vol. 19, No. 5, pp. 547–556.

Brennan, R.L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Brett, A., Rothlein, L., and Hurley, M. (1996). Vocabulary acquisition from listening to stories and explanations of target words. *Elementary School Journal, 96*, 415–422.

Bus, A.G., and van Ijzendoorn, M.H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 91*, 403–414.

Cain, G. (1975). Regression and Selection Models to Improve Nonexperimental Comparisons. In C. A. Bennet and A.A. Lumsdaine (Eds.), *Evaluation and Experiment* (pp.297-317). New York: Academic Press. 297-317.

Cappelleri, J.C.; Trochim, W.M.K.; Stanley, T.D.; Reichardt, C.S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review*, *15*:395-419.

Carlisle, J., and Scott, S. (2003). Teachers' Instructional Practice. Unpublished instrument.

Cook, T.D. (2008). "Waiting for life to arrive:" A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics, 142(2)*, 636-654.

Crowe, L.K. (2005). Comparison of two oral reading feedback strategies in improving reading comprehension of school-age children with low reading ability. *Remedial and Special Education, 26*, 32–42.

CTB/McGraw-Hill. (2003). *Terranova second edition: California Achievement Tests technical report*. Monterey, CA: Author.

Cullen, J., Jacob, B. and Levitt, S. (2006). The Effect of School Choice on Student Outcomes: Evidence from Randomized Lotteries. *Econometrica*. 74(5): 1191-1230.

Dixon, L.Q., Smith, W.C., Dwyer, M.C., Peabody, B.K., and Gamse, B.C. (2007, April). Training observers to use the Instructional Practice in Reading Inventory. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL

Dole, E., Nelson, D., Fox, D., and Gardner, J. (2001). Utah's Profile of Scientifically based Reading Instruction. Salt Lake City, Utah: The Institute for Behavioral Research in Creativity.

Dwyer, M.C., Smith, W.C., Dixon, L.Q., and Gamse, B.C. (2007, April). The development of the Instructional Practice in Reading Inventory. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Edmonds, M.S., and Briggs, K.L. (2003). The instructional content emphasis instrument: Observations of reading instruction. In S. Vaughn and K. L. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 31–52). Baltimore: Paul Brookes.

Foorman, B.R., Francis, D.J., Fletcher, J.M., Schatschneider, C. and Mehta, P. (1998). "The Role of Instruction in Learning to Read: Preventing Reading Failure in At-Risk Children", *Journal of Educational Psychology*, Vol. 90, No. 1, pp. 37-55.

Foorman, B.R., and Torgesen, J. (2001). Critical Elements of classroom and small-group instruction promote reading success in all children. *Learning Disabilities Research & Practice*, *16*(4), 203–212.

Foorman, B. and Schatschneider, C. (2003). Measuring teaching practice during reading/language arts instruction and its relation to student achievement. In S. R. Vaughn and K. L. Briggs (Eds.), *Reading in the classroom: Systems for observing teaching and learning*. Baltimore: Paul Brookes.

Foorman, B.R., Schatschneider, C., Eakin, M.N., Fletcher, J.M., Moats, L.C., & Francis, D.J. (2006). The impact of instructional practices in grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology, 31*, 1-29.

Gamse, B.C., Bloom, H., Celebuski, C., Dwyer, M.C., Gersten, R., Leos-Urbel, J., et al. (2004). *Reading First Impact Study: Revised study design*. Cambridge, MA: Abt Associates Inc.

Gamse, B.C., Bloom, H.S., Kemple, J.J., Jacob, R.T., (2008). Reading First Impact Study: Interim Report (NCEE 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gelman, A. and Stern, H. (2006). The difference between "Significant" and "Not Significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328—331.

Goldberger, A.S. (1972). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*. Institute for Research on Poverty: Madison, WI, Discussion Paper 123.

Goodson, B., Layzer, J., Smith, C., and Rimdzius, T. (2004). *Observation Measure of Language and Literacy Instruction (OMLIT)*. Developed as part of the Even Start Classroom Literacy Interventions and Outcomes (CLIO) Study, under contract ED-01-CO-0120 as administered by the Institute of Education Sciences, U.S. Department of Education.

Graves, A.W., Gerston, R., and Haager, D. (2004). Literacy instruction in multiple-language first-grade classrooms: Linking student outcomes to observed instructional practice. *Learning Disabilities Research & Practice*, *19*(4), 262–272.

Gunn, B., Smolkowski, K., Biglan, A., and Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up study. *Journal of Special Education, 36*(2), 69–79.

Haager, D., Gersten, R., Baker, S., and Graves, A. (2003). The English-Language Learner Classroom Observation Instrument: Observations of beginning reading instruction in urban schools. In S. R. Vaughn and K. L. Briggs (Eds.), *Reading in the Classroom: Systems for Observing Teaching and Learning*. Baltimore: Paul Brookes.

Hahn, H., Todd, P., and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(3), 201–209.

Harcourt Assessment, Inc. (2004). *Stanford Achievement Test series, Tenth Edition technical data report.* San Antonio, TX: Author.

Hatcher, P. J., Hulme, C. & Snowling, M.J. (2004). Explicit phoneme training combined with phonic reading instruction helps young children at risk of reading failure. *Journal of Child Psychology & Psychiatry, 45*, 338-358.

Hedges, L.V., and Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hernandez, A.V., Boersma, E., Murray, G.D., Habbema, J. D, Steyerberg, E.W. (2006). Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? *American Heart Journal. 151*(2):257-264.

Hoover, H.D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., Lewis. J.C., Qualls, A, L., Mengeling, M.A., and Shannon, G.P. (2003). The Iowa Tests guide to research and development, Forms A and B. Riverside Publishing, Itasca, IL.

Kamil, M.L. (2004). Vocabulary and comprehension instruction: Summary and implications of national reading panel findings. In P. McCardle and V. Chhabra (Eds.), *The voice of evidence in reading research*. Baltimore: Paul Brookes.

Kling, J.R., Liebman, J.B., and Katz, L.F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119.

Logan, A.C. and Tamhane, B. (2002). Accurate critical constants for the one-sided approximate likelihood ratio test of a normal mean vector when the covariance matrix is estimated. *Biometrics*, *58*(3), 650–656.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., and Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests–Manual for scoring and interpretation.* Itasca, IL: Riverside.

Mason, L.H. (2004). Explicit self-regulated strategy development versus reciprocal questioning: Effects on expository reading comprehension among struggling readers. *Journal of Educational Psychology, 96*, 283–296.

Mather, N., Hammill, D.D., Allen, E.A., and Roberts, R. (2004). Test of Silent Word Reading Fluency: Examiner's Manual. Austin, TX: PRO-ED, Inc.

McCutchen, D., Abbott, R.D., Green, L.B., Beretvas, S. N., Cox, S., Potter, N.S., Quiroga, T., & Gray, A. L. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning, *Journal of Learning Disabilities, 35*, 69-86.

McKeown, M.G., Beck, I.L., and Omanson, R.C. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly*, *20*, 522–535.

McKeown, M.G., Beck, I.L., and Omanson, R.C., and Perfetti, C.A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior, 15*, 3-18.

Mohr, L.B. (1995). *Impact Analysis for Program Evaluation*. Second Edition. Thousand Oaks, CA: Sage Publications.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No 00-4769 and 00-4754). Washington, DC: U.S. Government Printing Office.

No Child Left Behind Act of 2001, ESEA, 2001, Title 1, Part B, Subpart 1.

O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics, 40*(4), 1079–1087.

O'Connor, R.E., Bell, K.M., Harty, K.R., Larkin, L.K., Sackor, S.M., and Zigmond, N. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology*, *94*, 474–485.

Rasinski, T. and Oswald, R. (2005). Making and writing words: Constructivist word learning in a second-grade classroom. *Reading & Writing Quarterly*, 21, 151–163.

Reading First Advisory Committee. (2008) *Advisory Committee's Response to the Reading First Impact Study Interim Report*, retrieved 10/27/2008 from U.S. Department of Education website: http://www.ed.gov/programs/readingfirst/statement.pdf.

Reichardt, C.S., Trochim, W.M.K., and Cappelleri, J.C. (1995). Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review*, 19:39-63.

Reutzel, D.R. and Hollingsworth, P.M. (1993). Effects of fluency training on second graders' reading comprehension. *Journal of Educational Research, 86*, 325-331.

Reutzel, D.R. and Hollingsworth, P.M. (1991). Using literature webbing for books with predictable narrative: Improving young readers' prediction, comprehension, and story structure knowledge. *Reading Psychology*, *12*(4), 319–333.

Rosenshine, B., Meister, C., and Chapman, S. (1996). Teaching students to generate questions: A review of the intervention strategies. *Review of Educational Research*, *66*(2), 181–221.

Rothwell, P.M (2005). Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. Lancet. 365:176-86.

Schochet, Peter Z. (2008). *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561-84.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability Theory: A Primer.* Newbury Park, CA: Sage Publications.

Snow, C.E., Burns, M.S., and Griffin, P. (Eds.) (1998). Preventing reading difficulties in young children (Report of the Committee on the Prevention of Reading Difficulties in Young Children, National Research Council). Washington, DC: National Academies Press.

Southwest Educational Development Laboratory. Summary Report: Awards. Retrieved April 20, 2007 from http://www.sedl.org/readingfirst/report-awards.html.

Stahl, S.A. (2004). What do we know about fluency? Findings of the National Reading Panel. In P. McCardle and V. Chhabra (Eds.), *The Voice of Evidence in Reading Research* (pp. 187–210). Baltimore: Paul Brookes.

Suen, H.K., and Ary, D. (1989). Reliability: Conventional methods. In H. K. Suen and D. Ary (Eds.), *Analyzing Quantitative behavioral observation data* (pp. 99–129). Hillsdale, NJ: Lawrence Erlbaum.

Taylor, L.K., Alber, S.R., and Walker, D.W. (2002). The comparative effects of a modified self-questioning strategy and story mapping on the reading comprehension of elementary students with learning disabilities. *Journal of Behavioral Education*, *11*(2), 69–87.

Therrien, W.J. (2004). Fluency and comprehension gains as a result of repeated reading: A meta-analysis. *Remedial and Special Education, 25*, 252–261.

Thistlethwaite, D.L. and Campbell, D. T. (1960). Regression discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51*(6): 309–317.

Tomesen, M. and Aarnoutse, C. (1998). Effects of an instructional program for deriving word meanings. *Educational Studies, 24*, 107–128.

Torgesen, J.K., Wagner, R.K., Rashotte, C.A., Rose, E., Lindamood, P., Conway, T., et al. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*, *91*, 579–593.

Trochim, W.M.K. (1990). The Regression Discontinuity Design. In L. Sechrest, P. Perrin and J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (pp. 119-130). Rockville, MD: Agency for Health Care Policy and Research, U.S. Public Health Service.

U.S. Department of Education, Office of Elementary and Secondary Education. (2002). Guidance for the Reading First Program. Washington, DC.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. (2006) *Reading First Implementation Evaluation: Interim Report*, Washington, D.C.

Vaughn, S. and Briggs, K.L. (2003). Reading in the classroom: System for the observation of teaching and learning. Baltimore: Brookes Publishing.

Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J., & Drazen, J.M. (2007). Statistics in medicine—Reporting of subgroup analyses in clinical trials. *New England Journal of Medicine, 357*, 2189-2194.

Williams, K.T. (2001). Group Reading Assessment and Diagnostic Evaluation: Technical Manual. Circle Pines, MN: American Guidance Service, Inc.