NCEE 2010-4015

U.S. DEPARTMENT OF EDUCATION

Effectiveness of Selected Supplemental Reading Comprehension Interventions: Findings From Two Student Cohorts

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

Effectiveness of Selected Supplemental Reading Comprehension Interventions: Findings From Two Student Cohorts

May 2010

Susanne James-Burdumy
John Deke
Julieta Lugo-Gil
Nancy Carey
Alan Hershey
Mathematica Policy Research

Russell Gersten
Rebecca Newman-Gonchar
Joseph Dimino
Kelly Haymond
RG Research Group

Bonnie Faddis
RMC Research Corporation

Audrey Pendleton
Project Officer
Institute of Education Sciences

NCEE 2010-4015
U.S. DEPARTMENT OF EDUCATION

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

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
John Q. Easton
Acting Commissioner

May 2010

The report was prepared for the Institute of Education Sciences under Contract No. ED-01-C0039/0010. The project officer is Audrey Pendleton in the National Center for Education Evaluation and Regional Assistance.

To order copies of this report,

- Write to ED Pubs, U.S. Department of Education, P.O. Box 22207, Alexandria, VA 22304.
- 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 703-605-6794.
- Order online at www.edpubs.gov.

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

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.

ACKNOWLEDGMENTS

Many individuals, organizations, and agencies contributed to the Reading Comprehension Evaluation. The members of the evaluation's Technical Work Group—Donna Alvermann, Isabel Beck, Mark Berends, Thomas Cook, David Francis, Larry Hedges, Timothy Shanahan, Joseph Torgesen, and Joanna Williams—imparted valuable input at critical junctures.

At Mathematica Policy Research, important contributions were made by Annette Luyegu, Valerie Williams, Melissa Dugger, Irene Crawley, Sue Golden, and Season Bedell-Boyle, who helped manage the data collection activities; Arabinda Hazarika, Mark Beardsley, and Neil DeLeon, who developed and maintained the data collection databases; Mason DeCamillis, Elizabeth Petraglia, Ravaris Moore, Carol Razafindrakoto, and Maricar Mabutas, who programmed the impact models; Sally Atkins-Burnett and Aaron Douglas, who gave crucial psychometric assistance; Sonya Vartivarian, Sue Ahmed, and Amang Sukasih, who provided statistics support; and Cindy George, William Garrett, and Jill Miller, who were instrumental in editing and producing the report. We acknowledge the support and advice of David Myers and Jerry West, former study directors.

At RMC Research, we thank Steve Murray for his leadership in managing the competition to select the reading interventions and facilitating the study's pilot year, Wendy Graham for overseeing the team of RMC observers, and Margaret Beam for serving as a liaison to developers during the pilot and implementation years and for reviewing program training, classroom instruction, and materials. We are grateful to Lauren Liang at the University of Utah, who had a major role in developing the fidelity and observation measures and contributing to the development of the ETS assessments. At the University of Texas, we benefited from Sharon Vaughn's help in recruiting schools and addressing reading issues throughout the study, and Meaghan Edmond's assistance in creating the observation forms and in conducting the observation training. We thank Greg Roberts of Evaluation Research Services and Mary Jo Taylor at RG Research Group for their support in recruiting schools.

We appreciate the willingness of reading developers to engage in a large-scale, rigorous evaluation and to contribute their perspectives and insights during interviews. We could not have conducted this study without the districts, schools, and teachers who agreed to participate in the study, use the reading curricula, permit observation of their classroom instruction, and share their views.

This page is intentionally left blank.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

The research team for this evaluation consists of a prime contractor, Mathematica Policy Research, and two major subcontractors: RG Research Group and RMC Research Corporation. None of these organizations or their key staff members have financial interests that could be affected by findings from the study. None of the members of the Technical Working Group, convened by the research team to provide advice and guidance, have financial interests that could be affected by findings from the study.

This page is intentionally left blank.

CONTENTS

Chapter **Page**

Chapter **Page**

This page is intentionally left blank.

TABLES

Table **Page**

Table **Page**

Table **Page**

Table **Page**

Table **Page**

Table **Page**

This page is intentionally left blank.

FIGURES

Figure **Page**

EXECUTIVE SUMMARY

EFFECTIVENESS OF SELECTED SUPPLEMENTAL READING COMPREHENSION INTERVENTIONS: FINDINGS FROM TWO STUDENT COHORTS

Improving the ability of disadvantaged students to read and comprehend text is an important element in federal education policy aimed at closing the achievement gap. Title I of the No Child Left Behind Act (NCLB) calls on educators to close the gap between low- and high-achieving students using approaches that scientifically based research has shown to be effective. Such rigorous research is relatively scarce, however, so it is difficult for educators to determine how best to use Title I funds to improve student outcomes. Identifying interventions that improve reading comprehension is part of this challenge.

There are increasing cognitive demands on student knowledge in middle elementary grades where students become primarily engaged in reading to learn, rather than learning to read (Chall 1983). Children from disadvantaged backgrounds often lack general vocabulary, as well as vocabulary related to academic concepts that enable them to comprehend what they are reading and acquire content knowledge (Hart and Risley 1995). They also often do not know how to use strategies to organize and acquire knowledge from informational text in content areas such as science and social studies (Snow and Biancarosa 2003). Instructional approaches for improving comprehension are not as well developed as those for decoding and fluency (Snow 2002). Although multiple techniques for direct instruction of comprehension in narrative text have been well demonstrated in small studies, there is not as much evidence on the effectiveness of teaching reading comprehension within content areas (National Institute of Child Health and Human Development 2000).

The Institute of Education Sciences (IES) of the U.S. Department of Education (ED) has undertaken a rigorous evaluation of curricula designed to improve reading comprehension as one step toward meeting that research gap. In 2004, ED contracted with Mathematica Policy Research and its subcontractors to conduct the study.[1] The study team worked with ED to refine the study design and select the curricula to be tested, and then recruited districts and schools, collected data on implementation and outcomes in two consecutive school years, and analyzed the data. The study was conducted based on a rigorous experimental design for assessing the effects of four reading comprehension curricula on reading comprehension in selected districts across the country, where schools were randomly assigned to use one of the four treatment curricula in their fifth-grade classrooms or to a control group. The four curricula included in the study are: (1) Project CRISS, developed by CRISS (Santa et al. 2004), (2) ReadAbout, developed by Scholastic (Scholastic 2005), (3) Read for Real, developed by Chapman University and Zaner-Bloser (Crawford et al. 2005), and (4) Reading for Knowledge, developed by the Success for All Foundation (Madden and Crenson 2006).

[1]These subcontractors were RMC Research Corporation, RG Research Group, the Vaughn Gross Center for Reading and Language Arts at the University of Texas at Austin, the University of Utah, and Evaluation Research Services.

The experimental design ensures a valid basis for answering the study's key research questions:

1. What is the impact of the reading comprehension curricula as a whole on reading comprehension, and how do the impacts of the individual curricula compare to one another?
2. How are student, teacher, and school characteristics related to impacts of the curricula?
3. Which instructional practices are related to impacts of the curricula?
4. What is the impact of the curricula on students one year after the end of the intervention implementation?
5. Are impacts larger after schools and teachers have had one year of experience with the curricula?

The study's first report—based on the first year of data collected in 2006-2007 for the first cohort of fifth-grade students and released in May 2009 (James-Burdumy et al. 2009)—focused on the first three research questions. The findings indicated that, after one school year, there were no statistically significant positive impacts of the interventions, based on comparisons of fifth-grade student test scores in schools that were randomly assigned to use the interventions and schools that were randomly assigned to not use the interventions. Four statistically significant negative impacts of the curricula were observed. There was no clear pattern to the relationship between student, teacher, and school characteristics and the effectiveness of the interventions.

SECOND YEAR STUDY COMPONENTS AT A GLANCE

- **Fifth-grade component** – In this component, a second cohort of fifth-grade students from a subset of the study's original schools was added to the study, maintaining the original treatment assignments. Fifth-grade teachers in treatment schools implemented their assigned interventions and fifth-grade teachers in control schools continued teaching reading using methods they would have used in the absence of the study. Pre-tests and post-tests administered to students were used to assess the impact of the interventions on the second cohort of students. The rationale for including this component in the study is that impacts may be larger after schools and teachers have had one year of experience using the curricula.
- **Sixth-grade component** – In this component, the first cohort of students (all but 64 of whom were in sixth grade in the study's second year) was tracked for one additional year and follow-up tests were administered at the end of the school year to assess whether the interventions had statistically significant impacts one year after the end of their implementation. Fourteen sixth-grade students (0.2 percent) had the same teacher in sixth grade as in fifth grade, but the study interventions were *not* implemented in the second year when first cohort students were in sixth grade. There are two main rationales for including this component in the study: (1) it is possible that impacts of the interventions could emerge in the second year even after the intervention implementation has ended and (2) to examine whether the negative effects of Reading for Knowledge observed in the first year continued into the second year.

This report focuses on the fourth and fifth research questions, based on a second year of data collected for the study. The second year of the study focuses on (1) the impact of the interventions on Cohort 2 fifth-graders after one school year of implementation and (2) the impact of the interventions on Cohort 1 sixth graders one year *after* the *end* of the intervention implementation. In particular, it presents findings related to whether the curricula had an impact on students one year after the end of the intervention implementation based on follow-up student assessment data collected in spring 2008 for the first cohort of students (enrolled in the study in the 2006-2007 school year). The component of the study addressing this research question is referred to as the sixth-grade component of the second year of the study throughout the report (see box). This report also presents findings related to whether impacts are larger after *teachers* and *schools* had one year of experience using the curricula (the distinction between teachers and schools is due to mobility of teachers – some teachers in the second year are new to the study schools, but they might still benefit from the experience of their colleagues who had previously implemented the curricula). These findings are based on data collected for a second cohort of fifth-grade students (enrolled in the study in the 2007-2008 school year, after treatment schools, and some treatment teachers, had one year of experience using the curricula). The component addressing this research question is referred to as the fifth-grade component of the second year of the study throughout the report.

The main findings are:

- **The curricula did not have an impact on students one year after the end of their implementation.** In the second year, after the first cohort of students was no longer using the interventions, there were no statistically significant impacts of any of the four curricula.

- **Impacts were not statistically significantly larger after *schools* had one year of experience using the curricula.** Impacts for the second cohort of students were not statistically significantly different from zero or from the impacts for the first cohort of students. (Treatment students in the *second* cohort attended schools that had one prior year of experience using the study curricula, while treatment students in the *first* cohort attended schools with no prior experience using the study curricula. Reading for Knowledge was not implemented with the second cohort of students.)

- **The impact of one of the curricula (ReadAbout) was statistically significantly larger after *teachers* had one year of experience using the curricula.** There was a positive, statistically significant impact of ReadAbout on the social studies reading comprehension assessment for second-cohort students taught by teachers who were in the study both years (effect size: 0.22). This impact was statistically significantly larger than the impact for first-cohort students taught by the same teachers in the first year of the study.

In summary, our findings do not support the hypothesis that these four supplemental reading comprehension curricula improve students' reading comprehension, except when ReadAbout teachers have had one prior year of experience using the ReadAbout curriculum.

Curricula Included in the Second Study Year

The curricula included in the two second-year study components differed. The design of the study did not call for the interventions to be implemented in the sixth-grade component of the study, and, indeed, the interventions were not implemented in that component.[2] Rather, the design called for following first-cohort students for one additional year after the *end* of the implementation of the interventions in the study's first year, to assess whether implementation in the study's first year had longer-term effects on students' outcomes (measured in the study's second year when first-cohort students were in sixth grade). Therefore, the sixth-grade component focused on examining the impacts of the interventions implemented in the study's first year, which include Project CRISS (developed by CRISS) (Santa et al. 2004), ReadAbout (developed by Scholastic) (Scholastic 2005), Read for Real (developed by Chapman University and Zaner-Bloser) (Crawford et al. 2005), and Reading for Knowledge (developed by the Success for All Foundation) (Madden and Crenson 2006).

Three of the four curricula (Project CRISS, ReadAbout, and Read for Real) were included in the fifth-grade component of the second year, which involves a new cohort of fifth-grade students. Reading for Knowledge was not included in this component because 9 of the 18 schools that had been assigned to implement Reading for Knowledge elected not to continue implementing the intervention in the second year.

Study Design

The study's second year (2007-2008) design builds on the study's first year design (2006-2007). Before the start of the first year, schools in districts that agreed to participate were randomly assigned to one of the five study arms (four intervention groups and one control group). In both years of the study, fifth-grade teachers in schools assigned to an intervention group developed their own strategies for incorporating the assigned reading comprehension curriculum into their daily schedules and their core reading instruction. (The curricula being evaluated in this study were designed to supplement—not replace—the core curriculum being used by each teacher.) Teachers in control group schools continued to teach reading using whatever methods they had been using before the study began. Due to the experimental design, differences in outcomes of students in the treatment and control groups are attributable to the curricula being tested.[3]

---

[2]Thirty percent of sixth-grade students attended the same school in sixth grade as they did in fifth grade (because their school's grade structure included sixth grade). Very few sixth-grade students (0.2 percent) had the same teacher in sixth grade as in fifth grade. As noted above, none of the sixth-grade students received instruction in the study interventions in sixth grade.

[3]The study design just discussed is also described in James-Burdumy et al. (2006). Early study design proposals are laid out in Glazerman and Myers (2004).

SUMMARY OF FIRST- AND SECOND-YEAR EVALUATION DESIGN

Intervention:

- **First Year:** Four reading comprehension curricula (Project CRISS, ReadAbout, Read for Real, and Reading for Knowledge) were implemented with first-cohort students.
- **Second Year:**
  - **First-cohort students:** Interventions were *not* implemented with first-cohort students.
  - **Second-cohort students:** Due to attrition of schools assigned to the Reading for Knowledge group, only three curricula (Project CRISS, ReadAbout, and Read for Real) were implemented with second-cohort students.

Participants:

- **First Year:** 10 districts, 89 schools, 268 teachers, and 6,349 fifth-grade students in the study's first cohort. Districts were recruited from among those with at least 12 Title I schools, and schools were recruited only if they did not already use any of the four selected curricula. Students in those schools were eligible to participate if they were enrolled in fifth-grade classes as of January 1, 2007. Students in combined fourth-/fifth- or fifth-/sixth-grade classes were excluded, as were those with language barriers or in special education classes, although special education students mainstreamed in regular fifth-grade classes were eligible.
- **Second Year:**
  - **First-cohort students:** In the second year, the 6,349 students from the first year attended 252 schools, 176 of which agreed to permit follow-up testing of students.
  - **Second-cohort students:** 10 districts, 61 schools, 182 teachers, and 4,142 fifth-grade students in the study's second cohort. The same eligibility and exclusion restrictions were used with the first and second cohorts of students.

Research Design:

- **First Year:** Within each district, schools were randomly assigned to an intervention group that would use one of the four curricula or to a control group that did not have access to any of the curricula being tested. Control group teachers could, however, use other supplemental reading programs. The study administered tests to Cohort 1 students near the beginning and end of the 2006-2007 school year, observed classrooms, and collected data from teacher questionnaires, student and school records, and the intervention developers.
- **Second Year:** Schools and students maintained the same treatment (or control) group status in the second year. The study administered tests to Cohort 1 students at the end of the 2007-2008 school year and to Cohort 2 students near the beginning and end of the 2007-2008 school year, observed classrooms, and collected data from teacher questionnaires, student and school records, and the intervention developers. Cohort 2 impact analyses examined the effect of one year of exposure to the interventions after treatment schools and teachers had one year of experience using them. Cohort 1 impact analyses examined the longer-term effects of the implementation of the interventions in the first study year.

Outcomes: Impact estimates in both years focused on student reading comprehension test scores.

Schools participating in the fifth-grade component of the study's second year were in the same treatment or control group in the second year as in the first year. Students in the study's sixth-grade component were classified according to their treatment or control status from the study's first year. See box for a summary of the evaluation design.

There were three key distinctions between the first and second years of the study. First, fewer curricula were included in the fifth-grade component of the study's second year due to the attrition of schools assigned to implement Reading for Knowledge. Project CRISS, ReadAbout, and Read for Real were included in this component in the second year, while Reading for

Knowledge was not.[4] Second, fewer schools participated in the fifth-grade component of the study's second year (61 of the 89 schools that participated in the first year continued participating in Year 2).[5] Third, more schools participated in the study's second year than in the first year due to the study's sixth-grade component, in which follow-up tests were administered to Cohort 1 students at the end of the 2007-2008 school year in a total of 176 schools.

This study provides educators with a sense of the effectiveness of these curricula when used by teachers in "real-world" conditions. Although the study team worked to facilitate study activities such as the collection of data in study schools, the developers provided teacher training and follow-up support to teachers throughout the two study years, and teachers and schools could discontinue use of the curricula during the study period if they believed they were ineffective or too challenging to use. Therefore, the study conditions may be comparable to those many districts might face if they implemented these curricula in their schools.

Collecting Data

Addressing the study questions required information about the curricula and how they were implemented, study participants, and students' performance outcomes. Information about teaching and implementation of the curricula was collected to support an examination of the fidelity of implementation to each curriculum design, the ways the curricula affected more general (non-curriculum-specific) teaching practices related to comprehension and vocabulary instruction, the resources required to implement the curricula, and the way in which the curricula affected teachers' allocation of time during the school day. Data on all three "levels" of study participants—schools, teachers, and students—were collected as a basis for describing their characteristics as they entered the study. Outcomes for the first cohort of students were measured through assessments administered towards the end of the 2006-2007 and 2007-2008 school years. Outcomes for the second cohort of students were measured through assessments administered towards the end of the 2007-2008 school year. More information on the study's key data sources is provided below.

Information About Teaching and Implementation of the Curricula. Five data collection activities focused on teachers, teaching, and implementation of the four reading comprehension curricula. Two of these involved classroom observations, conducted in spring 2007 and spring 2008 for two purposes. To support interpretation of the impact estimates, intervention-specific "fidelity" observations of fifth-grade classes taught by treatment group teachers were conducted to determine the extent to which the teachers adhered to the curriculum content and procedures prescribed by each developer. To describe more general teacher practices related to comprehension and vocabulary instruction (as opposed to practices linked to a specific

[4]Reading for Knowledge was examined as part of the sixth-grade component of the study, because the sixth-grade component focused on examining the longer-term effects of the four curricula implemented in the study's first year with Cohort 1 students (all four study curricula, including Reading for Knowledge, were implemented in the first year).

[5]Of the 28 schools that left the study, 18 were assigned to Reading for Knowledge, 2 were assigned to Project CRISS, 2 were assigned to ReadAbout, 5 were assigned to Read for Real, and 1 was assigned to the control group.

intervention) and determine whether these practices were correlated with intervention impacts, Expository Reading Comprehension (ERC) observations were carried out in both treatment and control group fifth-grade classrooms to record the frequency with which teachers engaged in behaviors that research suggests are effective comprehension and vocabulary teaching practices. The third data collection activity that addressed the implementation of the curricula was a survey of developers on the cost of their curriculum to school districts. The fourth data collection activity related to teaching was a survey of fifth-grade teachers in the study's second year, administered to collect data on the amount of time students spent using informational text in a typical week. The last data collection activity related to teaching was a time allocation form administered to fifth-grade teachers in the second study year to collect data on teachers' allocation of time during the school day.

To help summarize the large amount of ERC observation data collected on general (non-intervention-specific) teaching practices related to comprehension and vocabulary instruction, the following three summary scales were created (for details on these scales, see Chapter II and Appendix F):

- ***Traditional Interaction.*** This scale captures interactive teaching practices, primarily focused on vocabulary instruction and drawing inferences from text, that have been in use for many decades in American schools (Durkin 1978-1979; Brophy and Evertson 1976).

- ***Reading Strategy Guidance.*** This scale captures teachers' use of aspects of strategy instruction (such as using text structure and generating summaries to improve comprehension) to build students' comprehension ability.

- ***Classroom Management and Student Engagement.*** This scale captures teaching practices related to the management of student behavior and students' engagement.

Data on Teacher Characteristics. The fifth-grade Teacher Survey, conducted in early fall 2006, was used to create two scales for examining the relationship between teacher characteristics and impacts (see Appendix F for details):

- ***School Professional Culture.*** The School Professional Culture scale is intended to capture conditions in schools that affect the quality of instruction (Consortium on Chicago School Research 1999; Carlisle 2003). The scale's 35 items—which were included in the Teacher Survey developed for this study—reflect teachers' perceptions of the culture in their school, including relationships with colleagues, access to professional development, experiences with changes being implemented in their school, and leadership support in their school.

- ***Teacher Efficacy.*** The Teacher Efficacy scale is intended to capture teachers' ability to benefit from professional development (Sparks 1988; permission to use scale provided by Hoy and Woolfolk 1993). The scale's 12 items, included in the Teacher Survey developed for this study, ask about teachers' attitudes concerning student engagement, instructional strategies, and classroom management.

Data on Students' Baseline Achievement Levels. Two student assessments administered at the start of the 2006-2007 and 2007-2008 school years allowed the study team to characterize the achievement level of the two cohorts of study students at baseline:

- ***Passage Comprehension subtest of the Group Reading Assessment and Diagnostic Evaluation (GRADE).*** This assessment, published by Pearson Learning Group, measures a student's ability to comprehend text passages (Williams 2001).
- ***Test of Silent Contextual Reading Fluency (TOSCRF).*** This assessment yields a score that reflects skills such as word identification, word meaning, and sentence structure, all of which are important skills for reading comprehension (Hammill et al. 2006).

Data on Student Outcomes. Data on students' post-test outcomes were collected from two sources at the end of the fifth-grade year (spring 2007 for Cohort 1 and spring 2008 for Cohort 2). First, students were retested using the GRADE (Williams 2001). In addition, students were tested for comprehension of social studies and science informational text, using assessments specially developed by the Educational Testing Service (ETS) for the study (Educational Testing Service 2007a and 2007b). To reduce burden, half the students were randomly assigned to take the science test and half to take the social studies test. Data on students' follow-up outcomes were collected from these same assessments at the end of the sixth-grade year (spring 2008) for the first cohort of students.

	Cohort 1 Students	Cohort 2 Students
Study Year 1 (2006-2007 school year)	• Cohort 1 students enter study as fifth graders • Interventions implemented with Cohort 1 treatment students • Administer pre-tests and post-tests	• Not yet included in study
Study Year 2 (2007-2008 school year)	• Cohort 1 students remain in study as sixth graders • Interventions are not implemented with Cohort 1 students • Administer follow-up tests	• Cohort 2 students enter study as fifth graders • Interventions implemented with Cohort 2 treatment students • Administer pre-tests and post-tests

Summary of Findings from the Study's First Year

The key findings from the first year of the study focus on curriculum implementation and impacts on student achievement. The implementation analyses document treatment teachers' training and feelings of preparedness to implement the curricula, adherence to their assigned curriculum, and teaching practices observed among teachers in the treatment and control group classrooms. The impact analyses examine how student outcomes were affected by the curricula and how the impacts relate to conditions and practices in study schools and classrooms. The key findings from the first year of the study were:

- **At the time of the classroom observations in spring 2007, over 80 percent (81 to 91 percent) of treatment teachers reported using their assigned curriculum.** Eighty-one percent of Read for Real teachers, 83 percent of Reading for Knowledge teachers, 87 percent of ReadAbout teachers, and 91 percent of Project CRISS teachers reported using their assigned curriculum.

- **Classroom observation data from the first year of intervention implementation showed that teachers implemented 55 to 78 percent of the behaviors deemed important by the developers for implementing each curriculum.** ReadAbout and Project CRISS teachers implemented, on average, 71 and 78 percent of such behaviors, respectively. Reading for Knowledge teachers implemented 58 and 65 percent of the behaviors deemed important for the two types of instructional days that are part of the curriculum. Finally, Read for Real teachers implemented 55 and 71 percent of the behaviors deemed important for the two types of instructional days that are part of that curriculum.

- **Two of three teacher practice scales were not statistically significantly different between the treatment and control groups.** There were no statistically significant differences in the Reading Strategy Guidance and Classroom Management and Student Engagement scales. Scores on the third scale, Traditional Interaction, were statistically significantly lower for the treatment group than the control group (effect size: -0.52).

- **No statistically significant positive impacts of the curricula on student outcomes were observed in the study's first year.** Reading comprehension test scores were not statistically significantly higher in schools using the selected reading comprehension curricula than in control schools.

- **There was some evidence of statistically significant negative impacts on student test scores in the study's first year.** The treatment group as a whole scored lower than the control group on the GRADE assessment (effect size: -0.08), and the Reading for Knowledge group scored lower than the control group on the ETS science comprehension assessment (effect size: -0.21). On the composite test score, the treatment group as a whole scored lower than the control group and the Reading for Knowledge group scored lower than the control group (effect sizes: -0.08 and -0.14, respectively).

Summary of Implementation Findings from the Study's Second Year

The second year implementation analyses focused on documenting treatment teachers' training, adherence to their assigned curriculum, teaching practices observed among teachers in the treatment and control group classrooms, and understanding teachers' allocation of time during the school day. The key implementation findings from the study's second year are:

- **During summer and early fall 2007, 50 to 91 percent of treatment teachers were trained to use the curricula.** Fifty percent of Read for Real teachers, 89 percent of Project CRISS teachers, and 91 percent of ReadAbout teachers were trained in the use of the curricula.

- **In the spring of the second year of the study, over 80 percent (83 to 96 percent) of treatment teachers reported using their assigned curriculum.** Eighty-three percent of Read for Real teachers, 92 percent of Project CRISS teachers, and 96 percent of ReadAbout teachers reported using their assigned curriculum. The percentage of teachers who reported using each of the three interventions did not differ significantly between the first and second years.

- **Classroom observation data from the second year of intervention implementation showed that teachers implemented 65 to 94 percent of the behaviors deemed important by the developers for implementing each curriculum.** Project CRISS and ReadAbout teachers implemented, on average, 65 and 94 percent of such behaviors, respectively, and Read for Real teachers implemented 75 and 76 percent of the behaviors deemed important for the two types of instructional days that are part of that curriculum. There were no statistically significant differences in average fidelity levels between the first and second study years.

- **Two of three teacher practice scales were not statistically significantly different between the treatment and control groups.** There were no statistically significant differences in the Reading Strategy Guidance and Classroom Management and Student Engagement scales. Scores on the third scale, Traditional Interaction, were statistically significantly lower for the Project CRISS treatment group than the control group (effect size: -0.54).

- **Project CRISS teachers were statistically significantly less likely than control teachers to report engaging in enrichment activities (such as art, music, or physical education), non-curricular activities (such as lunch, recess, or arrival/dismissal activities), and other activities.** Similar patterns were observed for ReadAbout and Read for Real, but those differences were not statistically significant.

What Is the Impact of the Curricula on Students One Year After the End of the Intervention Implementation?

No effects of the curricula on Cohort 1 students were observed in comparisons of outcomes measured one year after the end of the intervention implementation (in the study's second year). For the three intervention groups that had no effect in the first year, effects in the second year remained indistinguishable from zero. For the intervention group that had evidence of a negative effect in Year 1 (Reading for Knowledge), the effect in the second year was indistinguishable from zero. Figures 1 to 4 show impacts of the curricula on Cohort 1 students' follow-up test scores from spring 2008 (impacts on spring 2007 post-test scores are also shown for comparison purposes). Follow-up reading comprehension test scores in spring 2008 were not statistically significantly higher for students who attended treatment schools in the study's first year relative to students who attended control schools in the study's first year.

Figure 1. Effects of Curricula on GRADE Scores, Cohort 1 Students

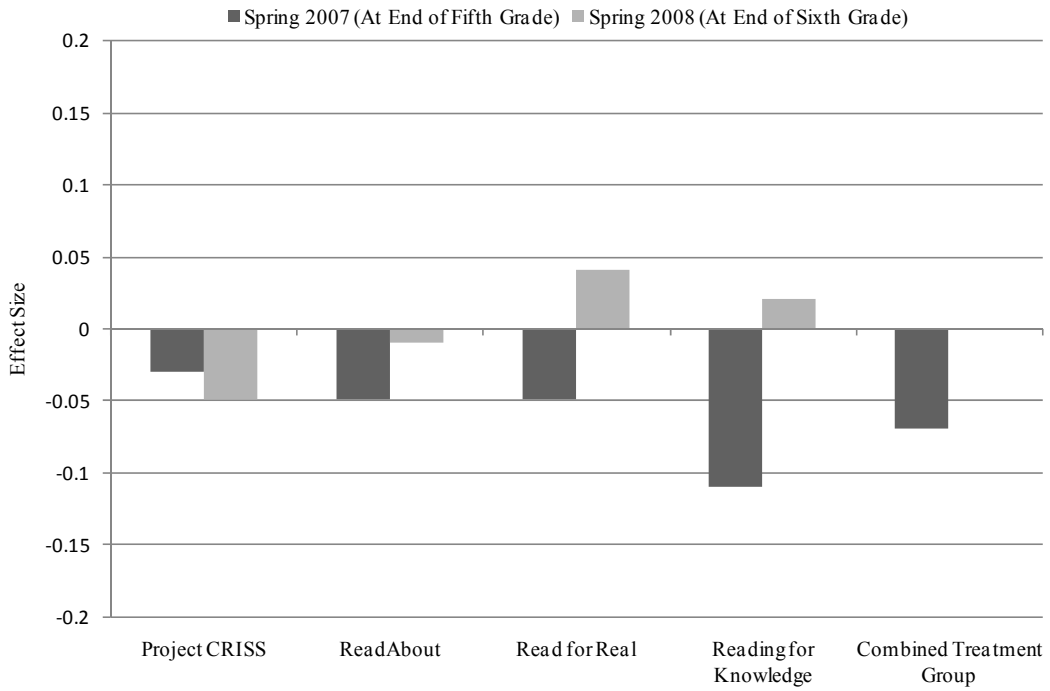


Figure 2. Effects of Curricula on Social Studies Reading Comprehension Scores, Cohort 1 Students

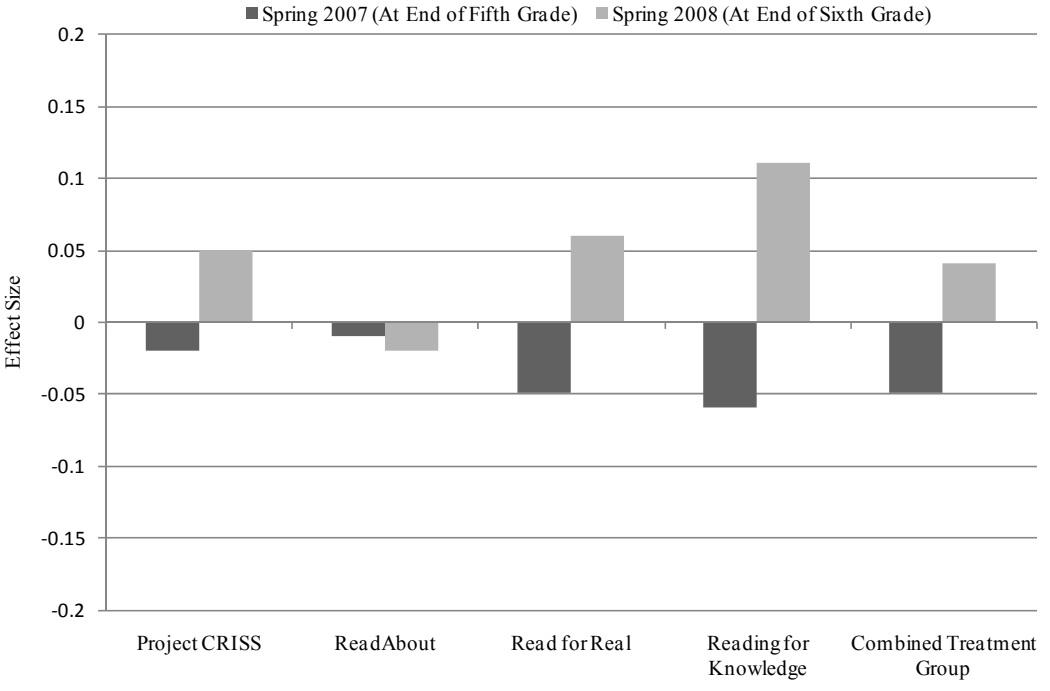
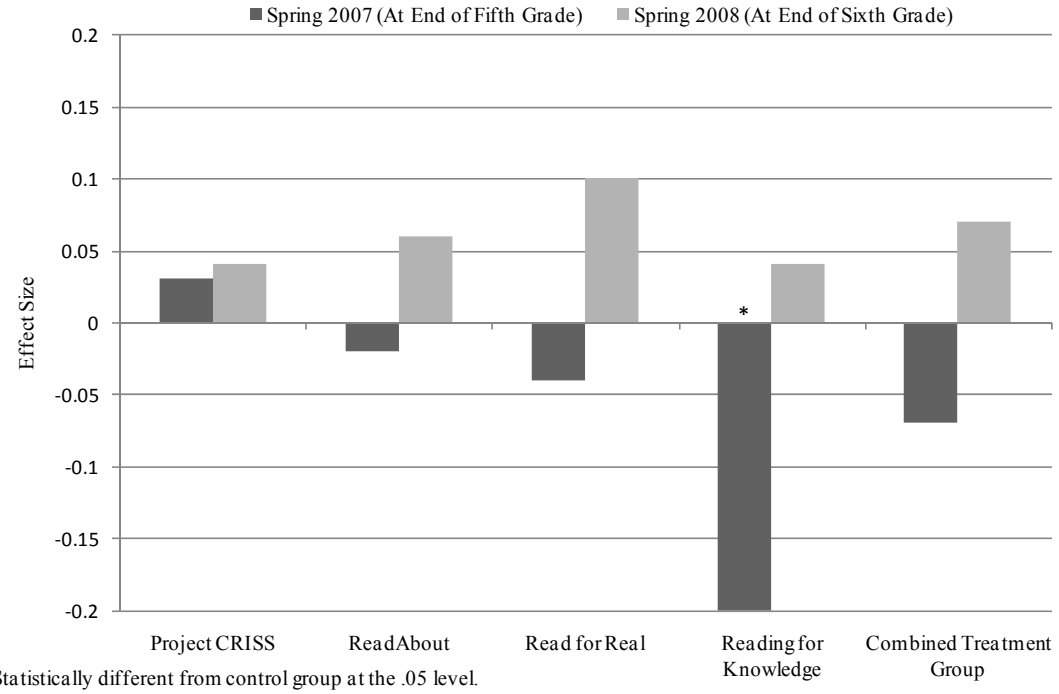
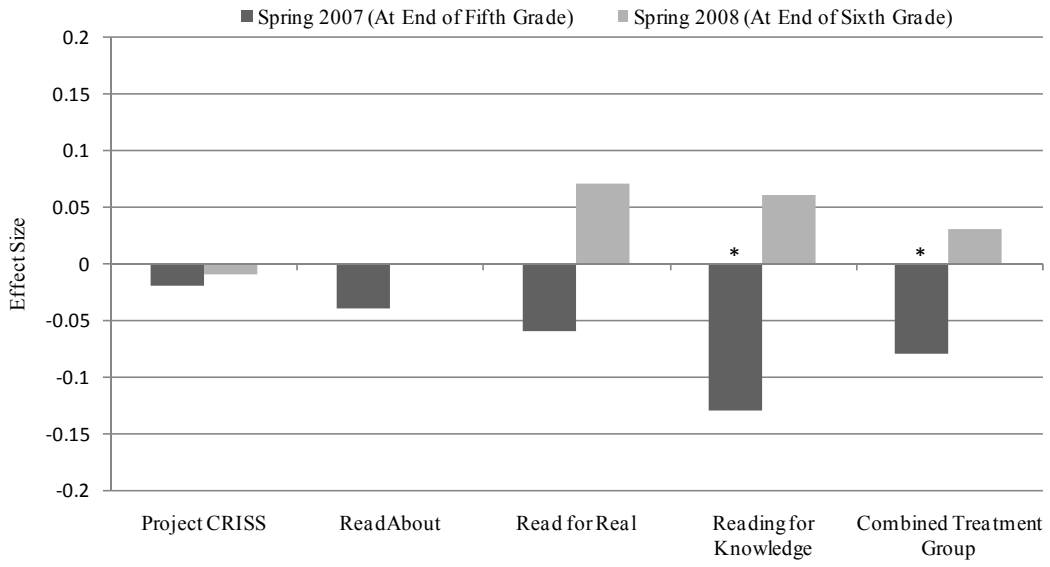


Figure 3. Effects of Curricula on Science Reading Comphrehension Scores, Cohort 1 Students



* Statistically different from control group at the .05 level.

Figure 4. Effects of Curricula on Composite Scores, Cohort 1 Students



NOTE: The composite scores are based on the GRADE scores, social studies reading comprehension scores, and science reading comprehension scores.

* Statistically different from control group at the .05 level.

Were Impacts Larger After Schools and Teachers Had One Year of Experience with the Curricula?

The second key research question examined in the second year of the study was whether impacts of the curricula were larger after schools and teachers had one year of experience using the curricula. As mentioned above, we distinguish between schools and teachers due to the mobility of teachers in and out of study schools. (Focusing on schools that participated in the study in both years, 76 percent of control group teachers and 72 percent of treatment group teachers remained in the study in both years. There were no statistically significant differences in the percentage of teachers remaining in the study across the treatment and control groups.)

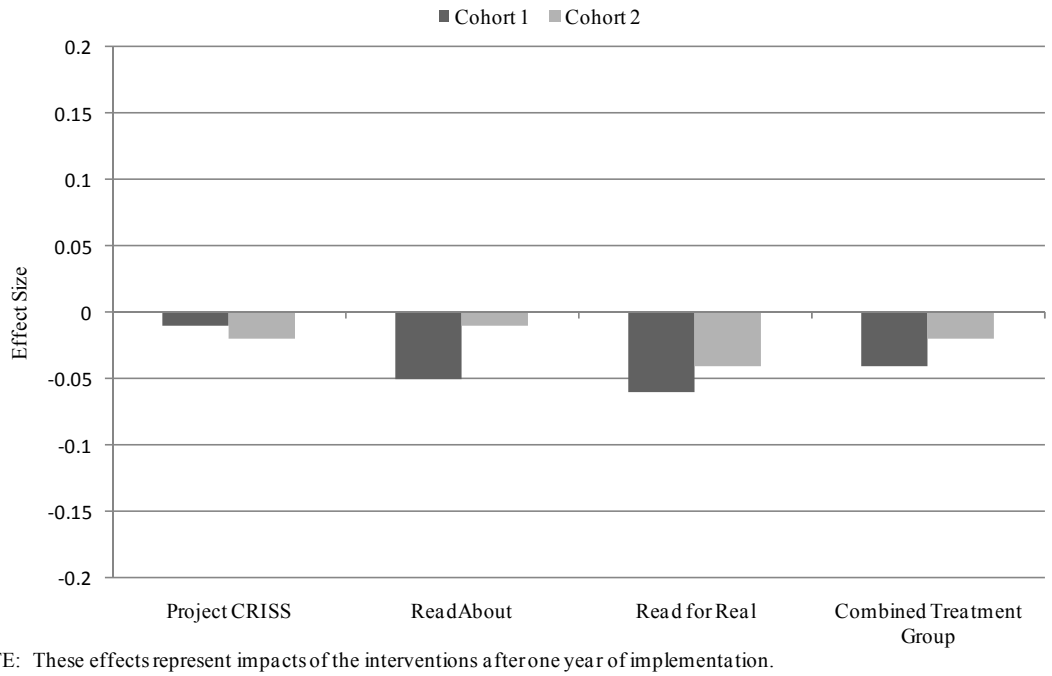
Impacts were not significantly larger after *schools* had one year of experience using the curricula. Overall, we found no statistically significant impacts of the interventions on any of the three student test score outcomes for the second cohort of fifth grade students, and there were no statistically significant differences between the one-year impacts for the first and second cohorts of students (Figures 5 to 8).

To address the research question related to *teacher* experience, the study team focused on post-test data (measured at the end of fifth grade) from first and second cohort students whose teachers were in the study in both the first and second years to assess whether the one-year impacts for the second group of students were larger than the one-year impacts for the first group.

The impact of one of the curricula (ReadAbout) was statistically significantly larger after *teachers* had one year of experience using the curricula (see Figures 9 to 12). When focusing on students of teachers who participated in the study for two years, we found one positive, statistically significant impact among students in the second cohort. In particular, there was a positive, statistically significant impact of ReadAbout on the social studies reading comprehension assessment (effect size: 0.22; Figure 10). To put this in perspective, for a student at the 50th percentile, an effect size of 0.10 represents about 4 percentile points, an effect size of 0.15 represents about 6 percentile points, and an effect size of 0.20 represents about 8 percentile points. To provide additional perspective, a meta-analysis by Rosenshine and Meister (1994) found an average effect size of 0.32 across nine studies examining the impact of multiple reading comprehension strategy instruction on standardized test scores (this meta-analysis focused on reciprocal teaching, which involves the use of guided practice and dialogue between students and teachers to teach students about four comprehension strategies including question generation, summarization, prediction, and clarification). Another meta-analysis by Rosenshine, Meister, and Chapman (1996) found an average effect size of 0.36 across 13 studies examining the impact of question generation on standardized test scores.

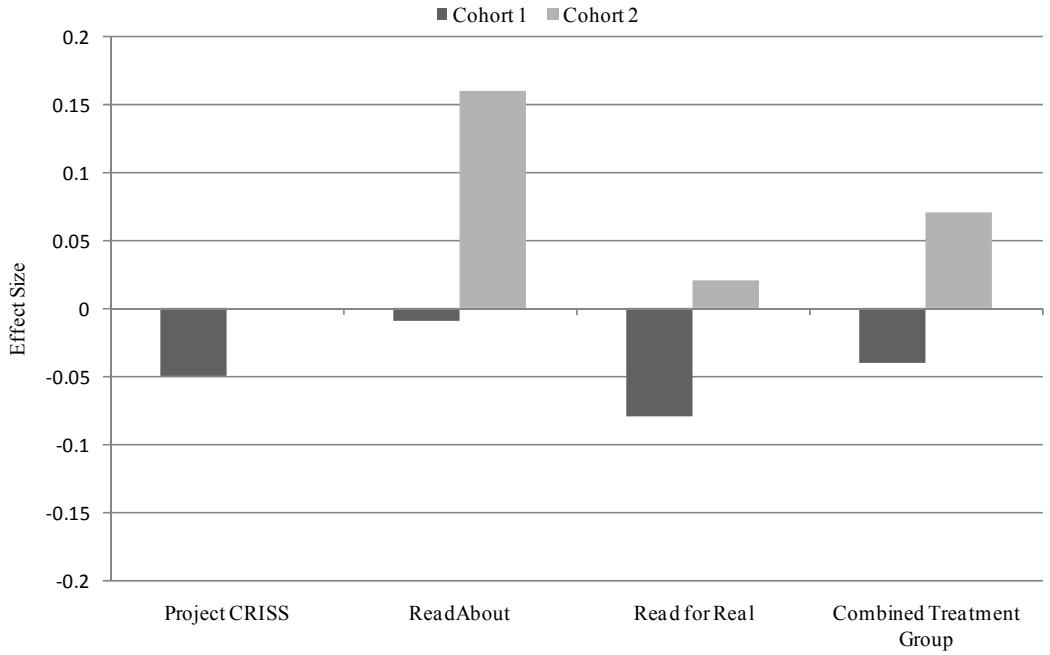
The impact of ReadAbout on the social studies reading comprehension post-test assessment for the second cohort of students was statistically significantly greater than the impact of ReadAbout on this outcome for the first cohort of students taught by the same teachers in the first year of the study (effect size difference: 0.28). ReadAbout's impacts on the other assessments (GRADE and science comprehension) were not statistically significant.

Figure 5. Effects of School Experience with the Curricula on Post-Test GRADE Scores of Fifth-Grade Students



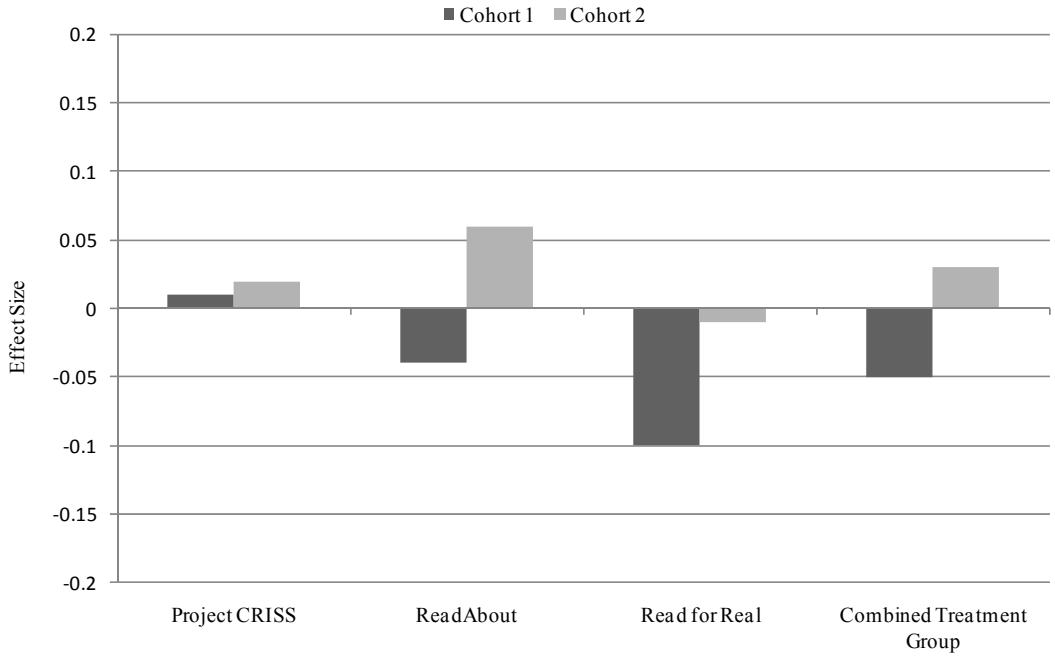
NOTE: These effects represent impacts of the interventions after one year of implementation.

Figure 6. Effects of School Experience with the Curricula on Post-Test Social Studies Reading Comprehension Scores of Fifth-Grade Students



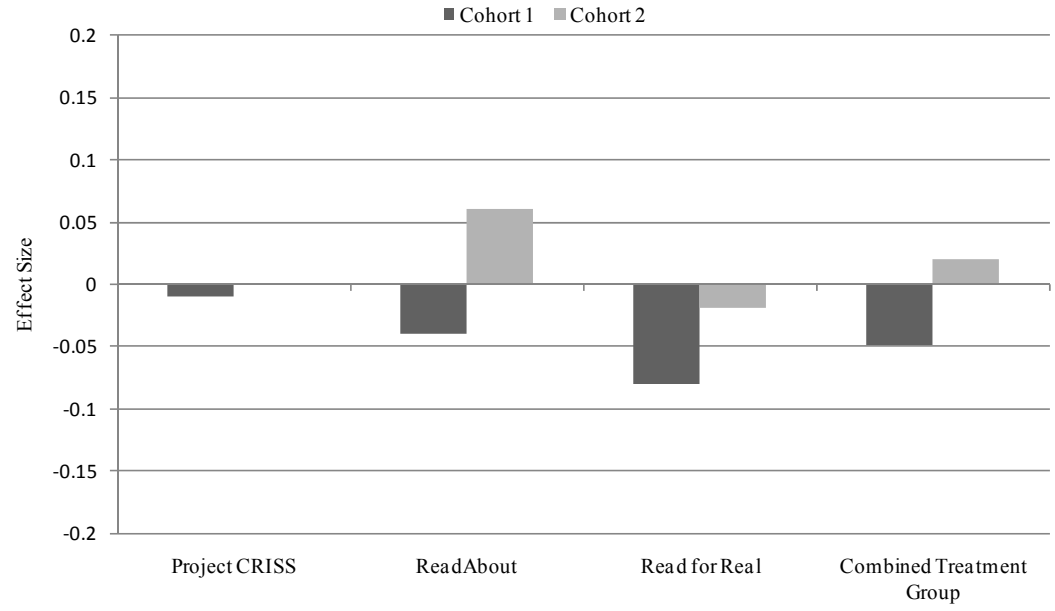
NOTE: These effects represent impacts of the interventions after one year of implementation.

Figure 7. Effects of School Experience with the Curricula on Post-Test Science Reading Comprehension Scores of Fifth-Grade Students



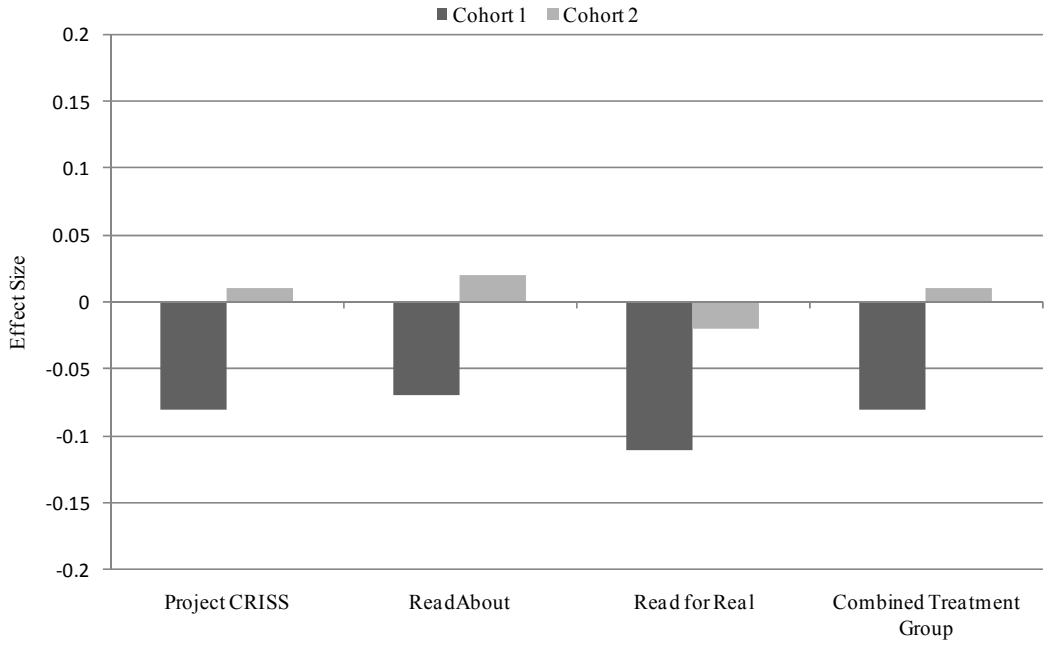
NOTE: These effects represent impacts of the interventions after one year of implementation.

Figure 8. Effects of School Experience with the Curricula on Post-Test Composite Test Scores of Fifth-Grade Students



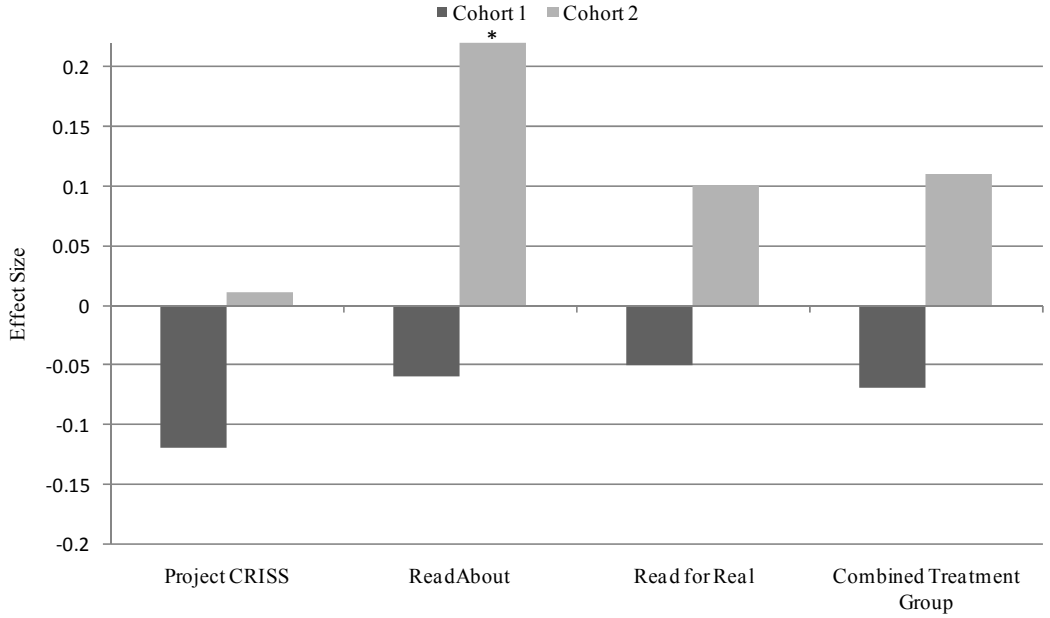
NOTE: These effects represent impacts of the interventions after one year of implementation. The composite scores are based on the GRADE scores, social studies reading comprehension scores, and science reading comprehension scores.

Figure 9. Effects of Teacher Experience with the Curricula on Post-Test GRADE Scores of Fifth-Grade Students



NOTE: These effects represent impacts of the interventions after one year of implementation.

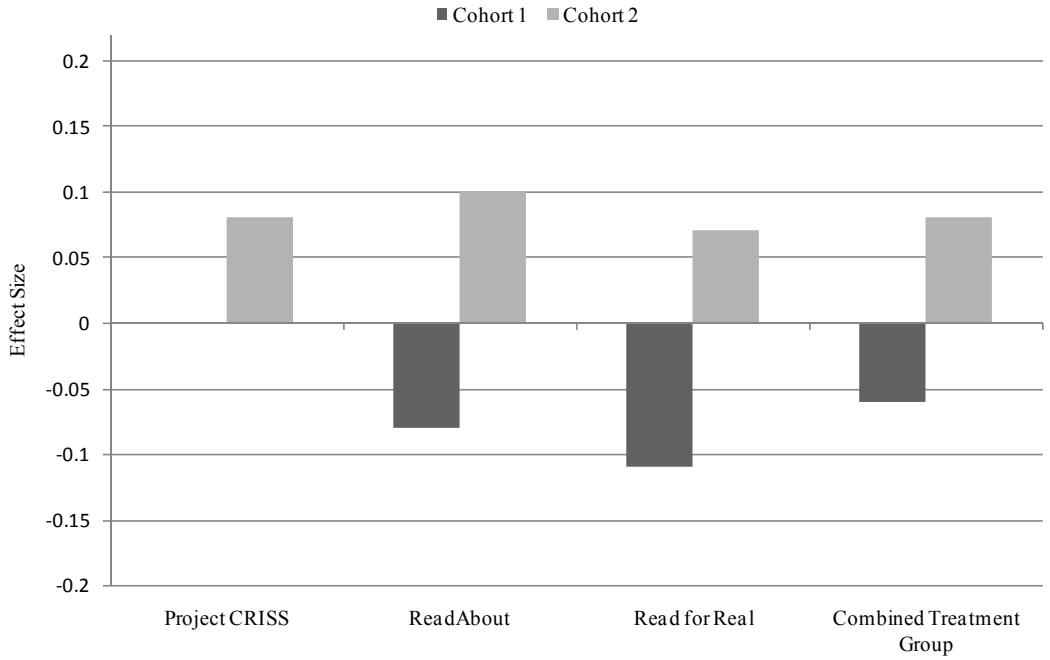
Figure 10. Effects of Teacher Experience with the Curricula on Post-Test Social Studies Reading Comprehension Scores of Fifth-Grade Students



NOTE: These effects represent impacts of the interventions after one year of implementation.

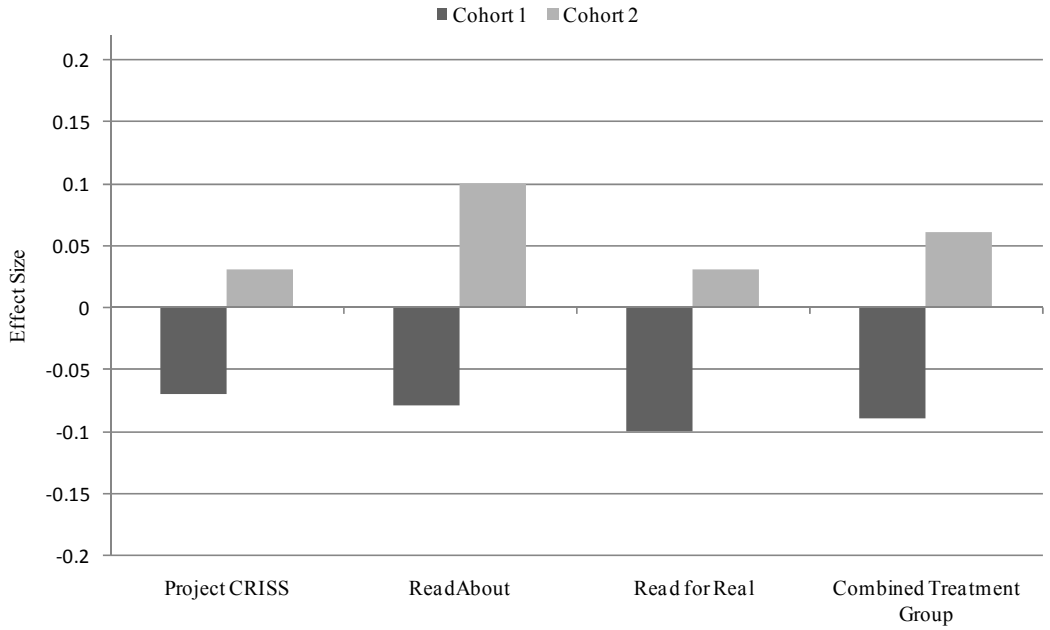
* Statistically different from control group at the .05 level.

Figure 11. Effects of Teacher Experience with the Curricula on Post-Test Science Reading Comprehension Scores of Fifth-Grade Students



NOTE: These effects represent impacts of the interventions after one year of implementation.

Figure 12. Effect of Teacher Experience with the Curricula on Post-Test Composite Test Scores of Fifth-Grade Students



NOTE: These effects represent impacts of the interventions after one year of implementation. The composite scores are based on the GRADE scores, social studies reading comprehension scores, and science reading comprehension scores.

Findings from Nonexperimental Analyses

For this report, the study team conducted a set of nonexperimental analyses to examine the relationship between students' test scores and classroom practices, teacher efficacy in the classroom, teacher professional development, and time students spent using informational text. The study team also examined the correlation of impacts and school characteristics. These findings must be interpreted with caution, as they are correlational in nature and, therefore, do not provide causal evidence of the relationship between the variables examined.

The key findings from these analyses are:

- **Two of the three teacher practice scales were correlated with test scores.** There is evidence of a positive and statistically significant relationship between post-test scores and Classroom Management (14 of 16 correlations were statistically significant) and Reading Strategy Guidance (10 of 16 correlations were statistically significant) scales. The Traditional Interaction scale was not statistically significantly related to post-test scores.

- **Three sets of individual items from the ERC were found to have the largest number of statistically significant positive correlations with test scores (48 of 64).** These items included teaching practices related to (1) explicit comprehension strategy instruction (16 of 24 correlations were positive and statistically significant), (2) teachers' management and responsiveness (18 of 24 correlations were positive and statistically significant), and (3) student engagement (14 of 16 correlations were positive and statistically significant). Among the other individual ERC items, just 15 of 344 correlations were positive and statistically significant.

- **No statistically significant relationships were found between test scores and teacher efficacy, hours of professional development reported by teachers, or time teachers spent with students in reading activities or using informational text.**

I. INTRODUCTION

Improving the ability of disadvantaged children to read and comprehend text is an important element in federal education policy aimed at closing the achievement gap. Title I of the No Child Left Behind Act (NCLB) of 2002 calls on educators to close the gap between low- and high-achieving students, using approaches found effective in scientifically based research. Such research is limited, however, so it is difficult for educators to decide how best to use Title I funds to improve student outcomes. Finding effective interventions to improve reading comprehension is part of this challenge.

There are increasing cognitive demands on student knowledge in middle elementary grades where students become primarily engaged in reading to learn, rather than learning to read (Chall 1983). Children from disadvantaged backgrounds often lack general vocabulary, as well as vocabulary related to academic concepts that enable them to comprehend what they are reading and acquire content knowledge (Hart and Risley 1995). They also often do not know how to use strategies to organize and acquire knowledge from informational text in content areas such as science and social studies (Snow and Biancarosa 2003). Instructional approaches for improving comprehension are not as well developed as those for decoding and fluency (Snow 2002). Although multiple techniques for direct instruction of comprehension in narrative text have been well demonstrated in small studies, there is not as much evidence on the effectiveness of teaching reading comprehension within content areas (National Institute of Child Health and Human Development 2000).

The Institute of Education Sciences (IES) of the U.S. Department of Education (ED) has undertaken a rigorous evaluation of interventions designed to improve reading comprehension as one step toward meeting that research gap. The Impact Evaluation of Reading Comprehension Interventions, begun in 2004, will contribute to the scientific research base available to practitioners. Carefully selected reading comprehension interventions were tested using a rigorous experimental design to determine their effects on reading comprehension among fifth-grade students in selected districts across the country.

Concerns over students' reading achievement[6] helped shape IES's process for defining research on issues related to Title I and the ultimate decision to focus this evaluation on reading comprehension of informational text. IES contracted with Mathematica Policy Research and its subcontractors in October 2002 to help identify issues relevant to Title I evaluation and to propose evaluation design options, and later, in October 2004, to conduct an evaluation.[7] IES

[6]Findings from the 2007 National Assessment of Educational Progress (NAEP) show that one-third of the nation's fourth graders have difficulty reading (U.S. Department of Education 2007). Other estimates suggest as many as 30 percent of elementary, middle, and high school students have reading problems that curtail educational progress and attainment (Moats 1999).

[7]These subcontractors were RMC Research Corporation, RG Research Group, the Vaughn Gross Center for Reading and Language Arts at the University of Texas at Austin, the University of Utah, and Evaluation Research Services.

and Mathematica® drew on input from two expert panels in the design of the study: the Title I Independent Review Panel (IRP) set up by Congress to advise ED on Title I evaluation, and a special Technical Work Group (TWG) of experts on reading comprehension and evaluation design.

With input from these sources, IES decided on an evaluation plan focused on fifth graders, so that the study complemented other IES initiatives to investigate the effectiveness of Reading First for younger students. This focus also reflected the concern that disadvantaged students may continue to struggle with reading as they reach upper elementary grades. The focus was on testing interventions designed to improve comprehension of expository text. Outcomes were defined as the ability to comprehend such text generally and in two specific content areas, science and social studies.

The resulting evaluation addresses a need for reliable information on the effectiveness of commercially available curricula designed to improve students' reading comprehension skills.[8] There is a massive body of research on children's reading and the individual comprehension strategies (or combinations of strategies) that may improve students' reading comprehension, but it offers little guidance on whether (and the extent to which) commercially available curricula improve students' reading comprehension (National Institute of Child Health and Human Development 2000). Moreover, the studies reviewed in the National Reading Panel (NRP) report suffered from a mix of limitations including small sample sizes, a focus on outcome assessments designed by the developers of the interventions being studied, the use of analytic methods that were not aligned with the unit of assignment, and the use of nonexperimental methods.

This study is designed to overcome those limitations. It focuses on curricula designed for commercial distribution. It is based on a rigorous experimental design and a large sample that included 10 districts, 89 schools, 268 teachers, and 6,349 students in the study's first cohort (enrolled in the study in the 2006-2007 school year) and 10 districts, 61 schools, 182 teachers, and 4,142 students in the study's second cohort (enrolled in the study in the 2007-2008 school year). The student assessments used to examine the interventions' impacts on reading comprehension were selected by the study team rather than developers.

The study's first report—based on the first year of data collected in 2006-2007—was released in May 2009 (James-Burdumy et al. 2009) and indicated that, after one school year, there were no statistically significant positive impacts of the interventions, based on comparisons of fifth-grade student test scores in schools that were randomly assigned to use the interventions and schools that were randomly assigned to not use the interventions. There was no clear pattern to the relationship between student, teacher, and school characteristics and the effectiveness of the interventions.

The second year of the study (conducted in the 2007-2008 school year) is the focus of this report, and has two main components. The first component follows students from the study's first year for one more year, using the same outcome measures, to examine whether there is an

[8]Three of the four curricula are currently available commercially. One curriculum—Reading for Knowledge—was designed for commercial use, but is not yet available to the public.

impact of the interventions one year after the interventions were implemented. The second component essentially repeats the first year of the study for three of the four interventions with a new cohort of fifth-grade students to assess whether the interventions are more effective after schools and teachers have had one year of experience using them. In sum, the second year of the study focuses on (1) the impact of the interventions on Cohort 2 fifth graders after one school year of implementation and (2) the impact of the interventions on Cohort 1 sixth graders one year *after* the *end* of the intervention implementation.

This second report presents the background and design of the evaluation, impact results from the 2007-2008 school year (the second year of intervention implementation and data collection), and differences in impacts between the 2006-2007 and 2007-2008 school years. As background for those results, this chapter reviews the existing research on reading comprehension strategies, the study design, identification of study interventions, selection and recruitment of study sites, and the data collected.

The remainder of the report presents findings on the implementation of the reading comprehension interventions and the impacts of those interventions on (1) longer-term, spring 2008 follow-up outcomes[9] for the first cohort of fifth-grade students (enrolled in the study in the 2006-2007 school year) and (2) short-term, spring 2008 post-test outcomes for the second cohort of fifth-grade students (enrolled in the study in the 2007-2008 school year).

Two types of differences in impacts are also presented. First, differences in post-test impacts between the first and second cohorts of students are presented to assess whether the interventions are more effective after teachers and schools have had one year of experience using them (recall that post-test outcomes are measured at the end of fifth grade for both cohorts, after one year of intervention implementation for treatment students). Second, differences in post-test (end of fifth grade) and follow-up (end of sixth grade) impacts for the first cohort of students are presented to assess whether the impacts of the interventions in the second year (when the first cohort of students were in sixth grade and no longer using the interventions) differ from impacts in the first year (when the first cohort of students were in fifth grade and treatment students had just finished one year of intervention implementation).

Finally, the report presents findings from exploratory, nonexperimental analyses that may be of interest to readers, including additional descriptive information on classroom practices and an examination of the relationship between student test scores and classroom practices, teacher efficacy, teachers' professional development, and students' time spent reading; and the relationship between impacts and school characteristics.

---

[9]Short-term, post-test outcomes for the first cohort of students were measured at the end of the year in which the interventions were implemented (the 2006-2007 school year). Longer-term, follow-up outcomes for the first cohort of students were measured at the end of the following school year (the 2007-2008 school year). The interventions were not implemented in the 2007-2008 school year with the first cohort of students. Short-term, post-test outcomes for the second cohort of students were measured at the end of the year in which the interventions were implemented (the 2007-2008 school year).

A. PAST READING RESEARCH HAS FUELED USEFUL RECOMMENDATIONS, BUT LEFT QUESTIONS UNANSWERED

A significant amount of research on specific instructional strategies to enhance reading comprehension is available. Although that research has been used to guide the development of many reading comprehension instructional programs, the effectiveness of those programs has generally not been studied (Liang and Dole 2006). In addition, the research base consists primarily of small-scale studies, many of which suffer from limitations in the rigor of their research design.

The NRP recommendations (National Institute of Child Health and Human Development 2000) and other research syntheses support a variety of techniques and approaches that can be classified into four groups: (1) student comprehension strategies, (2) teaching strategies, (3) instructional delivery, and (4) professional development. These recommendations are summarized below.

Student Comprehension Strategies. The NRP recommendations focus predominantly on teaching students strategies for making meaning out of text. Two recent reviews (National Institute of Child Health and Human Development 2000; Gersten et al. 2001) concluded that research shows the most benefit comes from approaches in which students use multiple strategies flexibly as they read. The NRP (National Institute of Child Health and Human Development 2000) and others (Pearson et al. 1992; Pressley 2002; RAND Reading Study Group 2000) have highlighted two types of strategies as particularly important:

- ***Summarizing.*** Summarizing consists of condensing textual information into essential or main points; it employs multiple strategies, such as determining what is important, categorizing, and organizing information (Brown and Day 1983).
- ***Question generation.*** Question generation involves students, not teachers, asking questions as they read (Martin and Pressley 1991; Wood et al. 1990; Rosenshine et al. 1996). The point of this strategy is for students to actively engage in the text by thinking about questions they want to answer as they read.

Teaching Strategies. A second group of recommendations from the NRP for effective comprehension instruction rests on teaching strategies that appear to influence students' comprehension (National Institute of Child Health and Human Development 2000), including:

- ***Use of engaging text.*** Research has shown that when students read texts that are interesting or that relate to topics of interest to them, they demonstrate improved comprehension compared to when they read other types of text (Renninger et al. 1992). Similarly, other research (Guthrie et al. 1998; Guthrie et al. 2000a; Guthrie et al. 2000b) supports the benefits of using texts containing vivid details that are relevant to the task and easily accessible, with colorful photographs and illustrations (Schraw et al. 1995).
- ***Embedding strategy instruction in texts students use in learning academic disciplines.*** Research suggests that, when strategy instruction (for example, teaching

students about summarizing or question generation) is embedded into the reading of text in academic content areas, students will be more likely to transfer their use of the strategies to texts they read in other content areas and on their own (Pressley 1998; Pressley 2002). Conversely, when strategies are taught in isolation (for example, on reading instruction workbook pages), students do not transfer skills from workbook pages to reading of expository texts (Pearson and Fielding 1991; Pressley 2000).

- ***Cooperative learning.*** Research suggests that cooperative learning—having students work together in groups, interacting with their peers while discussing text—can encourage students to think about and internalize comprehension strategies (National Institute of Child Health and Human Development 2000). Practicing a strategy in a small group has been found to contribute to the success of at least some researcher-developed instructional activities (National Institute of Child Health and Human Development 2000; Gersten et al. 2001).

Instructional Delivery. A third set of NRP recommendations focuses on instructional delivery—how best to implement instruction on student comprehension strategies (National Institute of Child Health and Human Development 2000). These recommendations encourage using direct, or explicit, instruction and explanation, two methods supported by research:

- ***Direct, or explicit, instruction.*** Teachers model how the comprehension strategy or skill is used (often called a "think aloud"), give feedback to students as they begin to use the strategy, and provide opportunities for students to practice using the strategy or skill independently (Rosenshine and Stevens 1986; Adams et al. 1982; Darch and Gersten 1986; Darch and Kame'enui 1987; Lloyd et al. 1980; Patching et al. 1983).

- ***Direct explanation of strategies.*** Teachers first *name and explain* a strategy, describe *when* and *how* it might best be used, and tell *why* it is important for improving reading. They next engage in a significant amount of explanation and cognitive modeling to show how to use the strategy. Students practice the strategy in teacher-mediated activities until they are able to use the strategy independently (Duffy et al. 1987; Duke and Pearson 2002; National Institute of Child Health and Human Development 2000; RAND Reading Study Group 2000).

Professional Development. A fourth focus of NRP recommendations—professional development in the teaching of reading comprehension strategies—has been found to be important in promoting effective teaching of reading comprehension (National Institute of Child Health and Human Development 2000). With sufficient professional development, teaching of comprehension strategies improves (Brown et al. 1996). Ongoing professional development consisting of one-on-one coaching, collaborative sharing, and lesson observation and feedback has been shown to help teachers learn to teach comprehension strategies (Duffy et al. 1987). This body of research suggests that building skills in teaching reading comprehension requires a good deal of professional development and that thorough use of comprehension strategy instruction is difficult for many teachers.

The NRP's research review and other research summaries referenced above suggest that interventions to improve reading comprehension can have positive effects on student outcomes, but many of the individual studies on comprehension instruction have limitations that highlight the importance of this study. First, many studies have been based on instruction delivered to students by well-trained graduate students or teachers personally trained by the researchers, which leaves open the question of how useful the interventions would be in "real-world" classrooms with teachers not exposed to such training (Klingner et al. 1998; Shany and Biemiller 1995). Another limitation is that reading materials that researchers used were sometimes different from those students typically encountered in classrooms (Anderson and Roit 1993; Baumann and Bergeron 1993). Although individual and even multiple strategies have been researched, no large-scale, rigorous studies of supplemental comprehension curricula designed for commercial distribution have been conducted. Developers of most current commercial programs indicate that their programs are "research-based," but they generally mean that instructional activities in the programs have been the focus of research studies. However, the *complete* program usually has not been rigorously researched (Liang and Dole 2006). Finally, many studies used outcome measures that were closely aligned to the specific goal of the intervention (see, for example, Baumann 1984; Hare and Borchardt 1984; Raphael and Pearson 1985; Taylor and Beach 1984). Positive effects are more likely with closely aligned outcome measures, but policy interest generally focuses on broader measures of reading comprehension.

B. STUDY DESIGN: FOCUS ON RIGOR AND UNDERSTANDING INTERVENTIONS

To address the limitations of earlier research noted in the prior section, the plan for this evaluation is based on a rigorous experimental design coupled with an emphasis on understanding the thoroughness of teachers' implementation of interventions under regular school conditions. The experimental design ensures a strong basis for answering the study's key research questions:

1. What is the impact of the reading comprehension curricula as a whole on reading comprehension, and how do the impacts of the individual curricula compare to one another?
2. How are student, teacher, and school characteristics related to impacts of the curricula?
3. Which instructional practices are related to impacts of the curricula?
4. What is the impact of the curricula on students one year after the end of the intervention implementation?
5. Are impacts larger after schools and teachers have had one year of experience with the curricula?

The first research question provides primary answers about intervention effectiveness. It addresses the question faced by school districts interested in investing in a curriculum to improve students' reading comprehension. The second and third questions help to understand what lies behind the results and might suggest directions for future research. In addition, answers to those

questions provide school districts with more detailed information on the conditions in which the interventions might be effective.

The second year of data collection permits the study team to address the fourth and fifth questions about the longer-term effects of the curricula. In particular, the fourth question addresses whether the interventions have an impact on students one year after the intervention implementation ended. The fifth question addresses whether intervention impacts are larger after schools and teachers have had one year of experience using the curricula.

1. First-Year Study Design

The study's second year (2007-2008) design builds on the study's first year design (2006-2007), the main features of which are summarized here. The study was based on a rigorous random assignment design; a competitive process for identifying interventions for the study; voluntary participation of districts, schools, and teachers; and a comprehensive data collection to facilitate answering the study's key research questions. The study design laid out below is also described in James-Burdumy et al. (2006).[10]

a. Random Assignment

Prior to the 2006-2007 school year, schools in districts that agreed to participate were randomly assigned to one of the five study arms (four intervention groups and one control group). For example, in a district with 10 schools, 2 schools were assigned to each treatment group and 2 schools were assigned to the control group. Curriculum developers provided training for teachers in schools assigned to their intervention. Teachers and schools assigned to a treatment or intervention group developed their own strategies for incorporating the assigned reading comprehension curriculum into their daily schedules and their core reading instruction. (As described in more detail in the next section, the curricula being evaluated in this study were designed to supplement—not replace—the core reading curriculum being used by each teacher.) Teachers in control group schools continued to teach reading using the methods they had been using before the study. Due to the experimental design, differences in outcomes of students in the treatment and control groups are attributable to the interventions being tested.

This study tests whether interventions are effective when districts and schools volunteer to participate. Eligible districts that were invited to participate in the study were under no obligation to participate, and only some of them (10 of 71) agreed to do so. When districts agreed to participate, they did so after holding discussions with leaders of schools that they felt best met the selection priorities for the study.

[10]Early study design proposals are laid out in Glazerman and Myers (2004).

SUMMARY OF FIRST- AND SECOND-YEAR EVALUATION DESIGN

Intervention:

- **First Year:** Four reading comprehension curricula (Project CRISS, ReadAbout, Read for Real, and Reading for Knowledge) were implemented with first-cohort students.
- **Second Year:**
  - **First-cohort students:** Interventions were *not* implemented with first-cohort students.
  - **Second-cohort students:** Due to attrition of schools assigned to the Reading for Knowledge group, only three curricula (Project CRISS, ReadAbout, and Read for Real) were implemented with second-cohort students.

Participants:

- **First Year:** 10 districts, 89 schools, 268 teachers, and 6,349 fifth-grade students in the study's first cohort. Districts were recruited from among those with at least 12 Title I schools, and schools were recruited only if they did not already use any of the four selected curricula. Students in those schools were eligible to participate if they were enrolled in fifth-grade classes as of January 1, 2007. Students in combined fourth-/fifth- or fifth-/sixth-grade classes were excluded, as were those with language barriers or in special education classes, although special education students mainstreamed in regular fifth-grade classes were eligible.
- **Second Year:**
  - **First-cohort students:** In the second year, the 6,349 students from the first year attended 252 schools, 176 of which agreed to permit follow-up testing of students.
  - **Second-cohort students:** 10 districts, 61 schools, 182 teachers, and 4,142 fifth-grade students in the study's second cohort. The same eligibility and exclusion restrictions were used with the first and second cohorts of students.

Research Design:

- **First Year:** Within each district, schools were randomly assigned to an intervention group that would use one of the four curricula or to a control group that did not have access to any of the curricula being tested. Control group teachers could, however, use other supplemental reading programs. The study administered tests to Cohort 1 students near the beginning and end of the 2006-2007 school year, observed classrooms, and collected data from teacher questionnaires, student and school records, and the intervention developers.
- **Second Year:** Schools and students maintained the same treatment (or control) group status in the second year. The study administered tests to Cohort 1 students at the end of the 2007-2008 school year and to Cohort 2 students near the beginning and end of the 2007-2008 school year, observed classrooms, and collected data from teacher questionnaires, student and school records, and the intervention developers. Cohort 2 impact analyses examined the effect of one year of exposure to the interventions after treatment schools and teachers had one year of experience using them. Cohort 1 impact analyses examined the longer-term effects of the implementation of the interventions in the first study year.

Outcomes: Impact estimates in both years focused on student reading comprehension test scores.

The integrity of the study design was maintained throughout the study's first year. Two treatment schools did not end up using their assigned intervention in the first year of the study, but student testing (at both baseline and post-test) was conducted in both of these schools to ensure that the integrity of the study's treatment and control groups was maintained.[11] See Appendix B for diagrams showing the flow of schools and students through the study.

[11]One school stopped implementing the intervention early in the school year when the only teacher who attended training discontinued using the program. The other school (in another district) never implemented the

b. Selection of Interventions

An open, competitive process was used to solicit proposals from curriculum developers. The invitation to submit proposals described the type of interventions to be included in the study. The reading comprehension interventions needed to supplement—not displace—the core reading, science, and/or social studies instruction in fifth-grade classrooms. They needed to take an average of 30 to 45 minutes per day to implement and they needed to encompass an entire school year.

A total of 13 proposals were submitted. Complete proposals were reviewed by an expert panel to assess the extent to which the proposals met substantive criteria for inclusion in a pilot implementation year. These criteria related to the intervention's theoretical and empirical underpinnings, evidence of the intervention's efficacy or effectiveness (based on previous research conducted by the developer or other researchers), the intervention's design and support proposed for teachers, institutional capability, and the appropriateness of the intervention for the study's target population.

Five programs were selected to participate in the 2005-2006 pilot year.[12] After the pilot year, four of the five interventions were selected for the full implementation of the study. To make this decision, the expert panel reviewed curriculum materials, initial proposals, and data collected during the pilot year. After discussing the interventions with IES and the study team, the panel recommended the four curricula they concluded best met the study's selection criteria, which included ease of use, intensity of the professional development provided, the extent to which curriculum activities were clearly specified, theoretical and empirical support for the program's content, and the developer's capacity to support a large-scale implementation. Based on these recommendations, IES then selected the following interventions (see Table II.1 for a summary of these interventions):

- **Project CRISS (**developed by CRISS) (Santa et al. 2004): Project CRISS focuses on five keys to learning—background knowledge, purpose setting, author's craft (which involves identifying and using the structure of text to help improve comprehension), active learning, and metacognition. The program is designed to be used during language arts, science, or social studies periods.
- **ReadAbout** (developed by Scholastic) (Scholastic 2005): Students are taught reading comprehension skills such as author's purpose, main idea, cause and effect, compare and contrast, summarizing, and inferences, primarily through a computer program.

(continued)

program after teachers were trained; the school indicated that its schedule could not accommodate the required 45 minutes of instructional time.

[12]During the pilot year, each developer recruited three Title I schools, trained an average of three teachers per school, and provided support to teachers during the year. The study team observed training and instruction, reviewed training and instructional materials, and provided formative feedback to the developers so they could refine their interventions. To eliminate any potential conflict of interest, the subcontractor who interacted with developers during the pilot year to refine the interventions was not involved in the impact study.

Students apply what they have learned during this time to a selection of science and social studies trade books.

- **Read for Real** (developed by Chapman University and Zaner-Bloser) (Crawford et al. 2005): In Read for Real, teachers work with a six-volume set of books to teach reading strategies appropriate for use before, during, and after reading (such as previewing, activating prior knowledge, setting a purpose, main idea, graphic organizers, and text structures). Each of these units includes vocabulary, fluency, and writing activities.

- **Reading for Knowledge** (developed by the Success for All Foundation) (Madden and Crenson 2006): Reading for Knowledge makes extensive use of cooperative learning strategies and a process called SQRRRL (Survey, Question, Read, Restate, Review, Learn).

c. District and School Recruiting

The study team began recruiting school districts for the study in January 2006. The team focused on districts that served low-income students and had enough schools to support the random assignment of schools in each participating district to the five arms of the study. Interested districts worked with the study team to identify schools that served low-income students and did not already use any of the four curricula identified for the study (or other similar comprehension curricula).

d. Study Participants

By August 2006, participating districts and schools had been identified and participation agreements with districts obtained. A total of 10 districts and 89 schools agreed to participate in the study's first year. Table I.1 shows the Year 1 sample sizes by intervention/control group.

As expected—given the types of districts and schools being recruited—the participating districts and schools were statistically significantly different from schools and districts nationwide in several respects. The districts included in the study were statistically significantly more disadvantaged, larger, and more urban than the average U.S. district (Table I.2). In particular, study districts had a higher percentage of students eligible for free or reduced-price lunch than the average district in the United States (57 percent vs. 39 percent). Study districts included more schools (65 vs. 6) and students (38,490 vs. 3,069) than the average U.S. district, and were more likely to be in urban areas (70 percent vs. 13 percent) than the average district.

Similar statistically significant patterns were found for the schools participating in the study (Table I.3). For example, study schools were more likely to be eligible for Title I funds (96 percent vs. 74 percent) and more likely to be operating schoolwide Title I programs, as compared to the average U.S. school (93 percent vs. 50 percent).[13] Study schools also included a

[13]Schools in which poor children make up at least 40 percent of enrollment are eligible to use Title I funds for schoolwide programs that serve all children in the school.

TABLE I.1

NUMBER OF STUDY DISTRICTS, SCHOOLS, TEACHERS, AND STUDENTS IN STUDY SAMPLE IN YEAR 1

Intervention	Number of Districts	Number of Schools	Number of Teachers	Number of Students[a]
Cohort 1 Post-Test				
Project CRISS	10	17	52	1,319
ReadAbout	10	17	50	1,245
Read for Real	9[b]	16	54	1,228
Reading for Knowledge	10	18	53	1,195
Control Group	10	21	59	1,362
Total	**10**	**89**	**268**	**6,349**

[a]This number includes all consenting students in the analysis sample. In Year 1, across all treatment groups, 87-88 percent of cohort 1 students in the analysis sample were tested at post-test (spring 2007).

[b]One district did not have enough participating schools to include all four intervention groups. The interventions that were assigned in that district were selected randomly.

TABLE I.2

CHARACTERISTICS OF DISTRICTS IN THE STUDY

Characteristics	U.S. Districts[a]	Districts in Study	Difference	*p-value*
Number of Schools per District	5.8	65.1	-59.3*	0.00
Percentage of Schools in Each District That Are:				
Title I Eligible	3.6	48.9	-45.3*	0.00
Schoolwide Title I	2.3	45.7	-43.4*	0.00
District Location (Percentage)				
Urban	12.8	70.0	-57.2*	0.00
Urban fringe	—[b]	—[b]	—[b]	—[b]
Town	16.8	0.0	16.8	0.16
Rural area	—[b]	—[b]	—[b]	—[b]
Number of Full-Time Teachers per District	120	573	-453*	0.00
Number of Students per District	3,069	38,490	-35,421*	0.00
Percentage of Students Eligible for Free or Reduced-Price Lunch[c]	38.6	57.3	-18.7*	0.02
Number of Districts	**16,019**	**10**		

SOURCE: 2005–2006 Common Core of Data (CCD).

[a]Data include districts with one or more regular schools. Regular schools are defined as public schools that do not focus primarily on vocational, special, or alternative education.

[b]Value suppressed to protect district confidentiality.

[c]Data are missing for 3 percent of districts with at least one regular school nationwide.

*Statistically different at the .05 level.

TABLE I.3

CHARACTERISTICS OF SCHOOLS IN THE FIRST YEAR OF THE STUDY

Characteristics	U.S. Schools[a]	Schools in Study	Difference	*p-value*
Schools Receiving Title I (Percentage)				
Title I Eligible School[b]	74.3	95.5	-21.2*	0.00
Schoolwide Title I[b]	49.5	93.3	-43.8*	0.00
School Location (Percentage)				
Urban[c]	28.8	68.5	-39.7*	0.00
Urban fringe	30.0	16.9	13.2*	0.01
Town and rural area[d]	41.1	14.6	26.5*	0.00
Students per Teacher (Average)	14.9	16.3	-1.4	0.33
Number of Students per School (Average)	449.8	560.3	-110.5*	0.00
Students Eligible for Free or Reduced-Price Lunch (Percentage)[e]	49.3	72.3	-23.0*	0.00
Student Race/Ethnicity (Percentage)[f]				
White	56.2	26.7	29.6*	0.00
Black	16.3	36.9	-20.6*	0.00
Hispanic	20.0	31.5	-11.4*	0.00
Asian	4.1	1.9	2.1*	0.03
Native American	2.0	1.0	1.0	0.32
GRADE Score (Average)	100.0	100.0	0.0	1.00
Number of Schools	**50,905**	**89**		

SOURCE: 2005–2006 Common Core of Data (CCD). Data from the last row of the table are from two sources: (1) the study team's baseline GRADE test administration, and (2) national GRADE norm information provided by the GRADE test's developer.

[a]Data include regular primary and middle schools that reported having fifth-grade classrooms. Regular primary and middle schools are defined as public elementary/secondary schools that do not focus primarily on vocational, special, or alternative education.

[b]Data are missing for 2 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[c]Data are missing for 0.7 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[d]The town and rural area categories have been combined to protect school confidentiality.

[e]Data are missing for 4 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[f]Data are missing for 0.8 percent of regular primary and middle schools that reported having fifth-grade classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation.

*Statistically different at the .05 level.

higher percentage of black (37 percent vs. 16 percent) and Hispanic (32 percent vs. 20 percent) students than the average school, reflecting the more urban nature of the study districts and schools.

Because over 90 percent of the schools participating in the study were schoolwide Title I schools, we also compared study schools to schoolwide Title I schools in the U.S. to assess how similar our study schools were to other Title I schools in the U.S. (Table I.4). Those comparisons showed that study schools were more likely than U.S. schoolwide Title I schools to be urban (69 percent vs. 39 percent) and less likely to be in a town or rural area (15 percent vs. 39 percent), and included a smaller percentage of white students (27 percent vs. 38 percent) and a higher percentage of black students (37 percent vs. 24 percent). Study schools also included more students than the average schoolwide Title I school (560 vs. 457).

e. The Sample Design Ensured an 80 Percent Probability of Detecting Impacts of at Least 0.17 Standard Deviations in the Study's First Year

The study design called for a sample that enabled us to detect impacts of individual interventions whose effect size was 0.25 standard deviations, with 80 percent probability. This calculation was based on assumptions regarding the intraclass correlation, school- and student-level R^2 (described below), and an adjustment for multiple comparisons. To attain this target effect size with 80 percent probability, the design called for recruiting 100 schools in 10 districts with 7,800 participating students. After recruitment was completed and 89 schools agreed to participate in the study, we were able, with 80 percent probability, to detect impacts of individual interventions on post-test student test scores in the study's first year of at least 0.17 standard deviations. The increase in statistical power was due to a greater benefit from covariate adjustment than anticipated. We originally assumed that there would be an intraclass correlation of 0.10, and school- and student-level R^2 of 0.50. The major factor contributing to the increased power was that the school-level R^2 turned out to be 0.89.

To put this in perspective, the average gain in GRADE scores among students in the control group between baseline and follow up was 0.44 standard deviations over a period of 245 calendar days. The full school year is about 270 calendar days. Assuming a constant rate of achievement gain over time, a 0.17 standard deviation gain would take about one-third of a school year (0.17/(0.44*270/245) = 0.35). The study's ability to detect impacts as low as 0.17 standard deviations can also be compared with the findings of a meta-analysis by Rosenshine and Meister (1994), which found an average effect size of 0.32 across nine studies of the impact of multiple reading comprehension strategy instruction on standardized test scores. (This meta-analysis focused on reciprocal teaching, which involves the use of guided practice and dialogue between students and teachers to teach students about four comprehension strategies including question generation, summarization, prediction, and clarification.) Another meta-analysis by Rosenshine, Meister, and Chapman (1996) found an average effect size of 0.36 across 13 studies examining the impact of question generation on standardized test scores.

With respect to teacher practices, which are of interest for the descriptive, implementation analysis, the study had less power due to smaller sample sizes of teachers and larger intraclass correlations (in the range of 0.20 to 0.30). For example, the smallest difference on the Traditional

TABLE I.4

CHARACTERISTICS OF SCHOOLS IN THE FIRST YEAR OF THE STUDY, COMPARED TO SCHOOLWIDE TITLE I SCHOOLS IN THE UNITED STATES

Characteristics	U.S. Schoolwide Title I Schools[a]	Schools in Study	Difference	*p-value*
Schools Receiving Title I (Percentage)				
Title I Eligible School	100.0	95.5	4.5*	0.00
Schoolwide Title I	100.0	93.3	6.7*	0.00
School Location (Percentage)[b]				
Urban	39.2	68.5	-29.4*	0.00
Urban fringe	22.0	16.9	5.2	0.24
Town and rural area[c]	38.8	14.6	24.2*	0.00
Students per Teacher (Average)	14.9	16.3	-1.4	0.33
Number of Students per School (Average)	456.7	560.3	-103.5*	0.00
Students Eligible for Free or Reduced-Price Lunch (Percentage)[d]	69.3	72.3	-3.0	0.17
Student Race/Ethnicity (Percentage)				
White	38.2	26.6	11.6*	0.00
Black	24.1	36.9	-12.8*	0.00
Hispanic	30.7	31.5	-0.8	0.82
Asian	3.3	1.9	1.3	0.17
Native American	2.5	1.0	1.5	0.20
Number of Schools	**24,754**	**89**		

SOURCE: 2005–2006 Common Core of Data (CCD).

[a]Data include regular primary and middle schools that reported having fifth-grade classrooms and that are schoolwide Title I schools. Regular primary and middle schools are defined as public elementary/secondary schools that do not focus primarily on vocational, special, or alternative education.

[b]Data are missing for 0.6 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[c]The town and rural area categories have been combined to protect school confidentiality.

[d]Data are missing for 1.5 percent of regular primary and middle schools that reported having fifth-grade classrooms.

*Statistically different at the .05 level.

Interaction scale of a single intervention that the study could detect with 80 percent probability was 0.75[14] in the study's first year.

f. Data Collection

Addressing the reading comprehension evaluation questions required collecting information about the interventions and how they were implemented, the study participants, and students' performance outcomes. We used information about implementation of the interventions to examine the fidelity of implementation to curriculum designs, to describe teaching practices related to comprehension and vocabulary instruction, and to examine the resources required to implement the interventions. Data were collected on all three "levels" of participants—schools, teachers, and students—as a basis for describing their characteristics as they entered the study and the preparation teachers had for using the new interventions (Table I.5). We measured subsequent student outcomes through reading comprehension test scores. The text in this section describes the data collection conducted by the study team during the first year of the implementation of the interventions—the 2006-2007 school year. The data collection conducted during the second study year (the 2007-2008 school year) is described in Section 2 of this chapter below.

(i) Information on Teaching and Intervention Implementation

In the study's first year, three data collection activities focused on teachers, teaching, and implementation of the four reading comprehension interventions. Two of these activities involved classroom observation. The first of these activities, "fidelity observations" of classes taught by treatment group teachers, were conducted to determine the extent to which teachers adhered to the curriculum content and procedures prescribed by each developer. The second data collection activity, "Expository Reading Comprehension" (ERC) observations, were carried out in both treatment and control group teachers' classrooms to record the frequency with which teachers engaged in behaviors that experts consider to be best practices for vocabulary and comprehension instruction. The third data collection activity pertaining to implementation of the interventions was a survey of developers on the cost of their curricula.

Fidelity Observations Were Used to Assess Adherence to Each Intervention. To support interpretation of the impact estimates, fidelity observations were conducted to provide a picture of how thoroughly the reading comprehension interventions were delivered. A separate fidelity observation form was developed for each intervention to capture whether treatment group teachers demonstrated behaviors or performed specific instructional activities inherent to the intervention. To create the forms, the evaluation team drew from each intervention's curriculum content and materials and then had the developer review the form to confirm that it accurately reflected the teaching practices and behaviors the developer expected as part of the curriculum's

[14]The minimum detectable effects reported in this paragraph are the effects that the study could detect with 80 percent probability (the standard level of power for reporting minimum detectable effects). The study could detect smaller effects with lower probability, which is why some of the reported statistically significant impacts are smaller than the effect sizes stated here.

TABLE I.5

SCHEDULE OF DATA COLLECTION ACTIVITIES

Data Collection Activity	Month and Year
Year 1	
Cohort 1	
Student Reading Tests—Baseline	August-October 2006
Teacher Survey	August-November 2006
Classroom Observations	January-April 2007
Student Reading Tests—Post-Test	April-June 2007
School Information Form	April-June 2007
Developer Survey	April-May 2007
Student Records	May-October 2007
Year 2	
Cohort 1	
Student Reading Tests—Follow Up	April-October 2008
Sixth-Grade Teacher Survey	April-October 2008
Cohort 2	
Student Reading Tests—Baseline	July-November 2007
Classroom Observations	January-May 2008
Teacher Survey—Students' Use of Informational Text	January-May 2008
Student Reading Tests—Post-Test	April-July 2008
Teacher Time Allocation Form	April-June 2008
Developer Survey	April-May 2008
Student Records	May-October 2008

implementation. Trained observers used the forms to record, primarily in yes/no format, the occurrence of 7 to 28 teaching practices, depending on the intervention.[15, 16]

The fidelity observation (one per teacher) was conducted only for teachers who reported using the curriculum. Treatment teachers were asked to schedule an observation in the spring at a time when they would be using the reading comprehension intervention. If teachers reported they had never or were no longer using the curriculum, the fidelity observation was not conducted (see Chapter II, Section C for information on the relatively minor extent to which this occurred). However, to create a full picture of the extent to which treatment teachers implemented the interventions, our analysis of implementation fidelity (presented in Chapter II) includes all teachers who were expected to implement each intervention (the analysis treats non-implementing teachers as not having engaged in the fidelity form behaviors). In particular, ones and zeros, respectively, were used in the data file to indicate whether a teacher engaged in or did not engage in a behavior listed on each curriculum's fidelity form. For teachers who reported they had never or were no longer using the curriculum, zeros were entered in the data file for all fidelity form behavior variables.

Observations of Instructional Practices Not Linked to Specific Curricula Provided a Basis for Assessing Differences in Teacher Practice. Structured observations across both treatment and control classrooms were conducted to provide descriptive information on the teaching practices in use in study classrooms. Unlike the fidelity observations described above, these observations focused on behaviors that reading experts posit as contributing to reading comprehension, rather than on the specific procedures developed by each curriculum developer. This approach—which provides a snapshot of the reading instruction fifth-grade students received from teachers using expository texts—measures how much teachers used specific vocabulary and comprehension-related teaching practices.

The ERC Classroom Observation Instrument, designed by a team of experts in reading instruction and classroom observation, was structured so that study team observers could record tallies of the number of times teachers displayed the instructional behaviors.[17] This approach was favored over the alternative approach of requiring observers to make more global judgments of the extent to which each behavior was observed, because the former approach was believed to be more likely to yield an unbiased measure of performed behaviors.

---

[15]For one intervention, these yes/no items were supplemented by questions about the focus of comprehension, vocabulary, and writing instruction, the length of instructional rotations and the number of students in the rotation, and the type of program materials used.

[16]The fidelity forms provide data on whether or not teachers engaged in a behavior; they do not provide data on the number of times the teachers engaged in each behavior or the quality of the behaviors.

[17]Tallies (or counts) of the number of times teachers engaged in these teaching practices were used to create scales summarizing teachers' practices. The process of creating these scales involved three main steps: (1) coding the tallies into ordinal categories, (2) conducting an exploratory factor analysis to determine conceptual groupings of items, and (3) estimating an item response theory (IRT) model using the categorical variables formed in the first step. These steps are explained in detail in Chapter II, Section D and Appendix F. Appendix I presents key descriptive statistics (such as means and standard deviations) for the full set of fidelity and ERC observation items.

The team of reading experts determined the critical behaviors to be recorded. Based on a review of prominent reading research (including Palincsar and Brown 1984 and Rosenshine et al. 1996), they identified the key behaviors associated with improved reading achievement, developed measures of those behaviors, and then refined the measures using trial observations of classroom teachers.

The behaviors identified for the ERC form (and the teacher practice scales based on those behaviors) were indeed related to student test score outcomes observed in this evaluation. Two of the three scales created (the Reading Strategy Guidance and Classroom Management scales) were statistically significantly related to follow-up student test scores (see Appendix F for more information on how criterion validity was assessed).

The behaviors recorded on the ERC form comprised practices related to comprehension and vocabulary. Observers documented occurrences of eight comprehension-related behaviors, such as activating prior knowledge, providing explicit instruction on how to use comprehension strategies, and asking students to justify their responses. For each behavior, observers recorded the number of times the practice occurred in the form of (1) teacher modeling; (2) teacher explaining, reviewing, providing examples, or elaborating; or (3) student practice. Six behaviors related to vocabulary were tallied. Observers noted, for example, the number of times teachers provided an explanation or definition or the number of times teachers provided examples, contrasting examples, multiple meanings, or elaborations on student responses.

Analysis of teacher behavior data was based on observations conducted on one day—when informational texts were used—for each treatment and control teacher. Observations were conducted in January through April 2007, so teachers had time over the first part of the school year to become familiar and practiced with the new curriculum. Study staff observed any class period in which teachers were using informational text, including reading/language arts, science, social studies, and test preparation.[18] Observers tallied the targeted behaviors in 10-minute intervals (recording up to 11 tallies within each interval) and observed as many intervals in which informational text was used as occurred (up to 10 intervals within each class period), to capture all instruction involving informational text. We conducted observations of 98 percent of the teachers in the first study year.[19] On average, classrooms were observed 1.8 times during the day of observations (this ranged from a minimum of 1 time to a maximum of 3 times).[20] Classrooms were observed for 49 minutes during the day of observations, on average (this ranged from a minimum of 15 minutes to a maximum of 123 minutes).

---

[18]In departmentalized schools, all teachers who taught a given classroom of students for reading/language arts, science, or social studies were considered a teaching unit, and all were observed.

[19]Response rates for each arm of the study (four treatment groups and one control group) are provided in Appendix E.

[20]Although classrooms, on average, were observed multiple times during the day, they were only observed for a single day, which may reduce the reliability of the teacher practice scales based on the ERC data (relative to observations conducted over multiple days). The teacher practice scales based on a single day of observations still allow us to calculate valid estimates of treatment/control differences on the scales (presented in Chapter II), but the correlations based on these scales (presented in the correlational analyses in Chapter V) may be attenuated.

Observers participated in four days of training, and inter-rater reliability of at least 80 percent was achieved during the training. The training included detailed explanations of behavior items and practice observing videotaped classes. Each observer who achieved at least 80 percent reliability with a master trainer (defined as within one tally for each item in the time interval) was certified to conduct classroom observations for the study.

Assessments of inter-rater reliability continued during data collection to ensure that no erosion of consistency had occurred. Pairings of a master trainer with each observer at least once during the first two weeks of observation, coupled with randomly assigned pairings of regular observers throughout the field period, provided inter-rater reliability data on 25 percent of the teachers and classrooms observed.[21] A variety of measures were used to assess inter-rater reliability, including simple sums of tallies and mean tallies for each teacher across the 10-minute intervals. Later, we computed scales from the tallies (see Chapter II, Section D and Appendix F), and the inter-rater reliability for the three scales ranged from 0.94 to 0.98.

Developer Survey Provided Data on Costs of Implementing the Programs. Since treatment schools did not have to pay to receive the reading program assigned to them for the study, we asked developers about the costs that non-study schools would incur to implement their program in the 2006-2007 school year. Using an ingredients approach (Levin and McEwan 2001), we identified all the items schools would need to purchase to implement and obtain support for the interventions. We then asked developers to specify the unit charge for each item, and we calculated total costs per reading comprehension program based on the quantities needed of each unit. This approach allowed us to compare (1) the implementation and support services that developers provided to study districts, schools, and teachers with what they typically provided to others outside the study purchasing their services in the 2006-2007 school year, and (2) program costs and implementation and support services provided across developers.

(ii) Data to Describe Teachers, Schools, and Students

An essential part of documenting study results is describing the participants and assessing the similarity of the treatment and control groups. Data collection therefore included a Teacher Survey, School Information Form, student assessments, and Student Records Form.

Teacher Survey Obtained Data on Teacher Characteristics and Attitudes. The teacher survey data collected allowed the study team to describe the teachers participating in the study, assess the similarity of treatment and control group teacher characteristics, and examine the relationship between teacher characteristics and intervention impacts. The Teacher Survey—conducted in treatment and control schools in August through November 2006 (as teachers began the first study year)—included items about the teacher's background and experience, grade levels taught, educational credentials, gender, age, and race/ethnicity. The survey also included items from School Professional Culture and Teacher Efficacy scales (see below for details on these scales). For treatment teachers only, the survey contained questions about the

[21]When a behavior was not observed during an interval, observers recorded a tally of zero. Reliability was computed both with and without these zeroes (the latter was done to guard against inflation of inter-rater reliability).

training they received on the study curriculum. Treatment teachers were asked to rate the training on various dimensions and to indicate how well prepared to use the curriculum they felt as a result of the training.

In nondepartmentalized schools, the questionnaire was given to all fifth-grade teachers. In departmentalized schools, the survey was usually administered to reading/language arts teachers (in a few treatment schools it was given to science or social studies teachers instead because they had received the intervention training and the reading/language arts teachers had not). A response rate of 93 percent was achieved. Item responses were used to create the following scales (see Appendix F for details):

- ***Teacher Efficacy.*** This scale was included on the Teacher Survey because it is correlated with teachers' ability to benefit from professional development (Sparks 1988).[22] It is based on 12 items from the Teacher Survey developed for this study (items used with permission from Hoy and Woolfolk 1993). These items ask about teachers' attitudes about student engagement, instructional strategies, and classroom management. The reliability of this scale was .90.

- ***School Professional Culture.*** This scale was designed to capture conditions in schools that affect quality of instruction (Consortium on Chicago School Research 1999; Carlisle 2003). It is based on 35 items from the Teacher Survey developed for this study and reflects teachers' perceptions of the culture in their school, including relationships with colleagues, access to professional development, experiences with changes being implemented in their school, and leadership support in their school. The reliability of this scale was .87.

School Information Forms Captured Data on School Characteristics. At the end of the first study year (between May and October 2007), schools provided information that could help describe the study context, contribute school-level variables to the impact analysis, and permit the study team to examine the relationship between impacts and conditions in schools. Items on the form included school enrollment, the percentage of students eligible for free or reduced-price lunches, the percentage classified as ELL, and the textbooks, basal reading series, and special programs or supplementary curricula the schools were using for reading instruction just before the study began. Data were collected from 94 percent of the schools.

Baseline Data on Students Were Collected from Tests and Records. Data on student achievement levels were used to characterize the student sample at baseline. Starting in the third week of school (after enrollment had settled and parental consent had been obtained), the study team administered two standardized tests to fifth graders. Table I.6 describes the norming samples and presents reliability and validity statistics for these two assessments (and a third administered at follow up). Descriptions of the two baseline tests are as follows:

[22]The items included on the Teacher Survey are an abbreviated version of a teacher efficacy scale (Hoy and Woolfolk 1993; Gibson and Dembo 1984).

TABLE I.6

FEATURES OF TESTS USED IN THE STUDY

Characteristic	Group Reading Assessment and Diagnostic Evaluation (GRADE), Passage Comprehension Subtest	Test of Silent Contextual Reading Fluency (TOSCRF)	Educational Testing Service (ETS) Social Studies/Science Reading Comprehension Assessments
General Information	Commercially available norm-referenced, group-administered reading assessment. The Passage Comprehension subtest measures students' ability to comprehend extended text as a whole. Students read a passage and then answer multiple-choice questions about the passage. Level 5, Form A was used for grade 5 students, and Level 6, Form A was used for grade 6 students. (Two alternative forms at each test level are available.)	Commercially available norm-referenced, group-administered assessment of silent reading fluency. The test measures the speed with which students can recognize the individual words in a series of printed passages that are printed in uppercase without punctuation or spaces between words.	Two pairs of tests developed specifically for the Evaluation of Reading Comprehension Interventions – one pair for students in grade 5 and one pair for students in grade 6. The tests measure students' ability to comprehend expository text; each pair includes one test emphasizing the reading of science-based passages and one emphasizing the reading of social studies-based passages. Students read a passage and then answer multiple-choice questions about the passage.
Norm Sample	National norms for the full test are based on samples of students in 46 states—16,408 in spring 2000 and 17,024 in fall 2000. Norms for the Passage Comprehension subtest are as follows: fifth-grade norms are based on 473 students in spring and 570 students in fall; sixth-grade norms are based on 539 students in spring and 513 in fall. The average student in the norm sample has a standard score of 100, and the standard deviation of standard scores is 15.	National norms are based on a sample of 1,898 students in 23 states tested in spring and fall of 2004. The average student in the norm sample has a standard score of 100, and the standard deviation of standard scores is 15.	Not nationally normed.
Reliability	For the Level 5 Passage Comprehension subtest, split-half reliability coefficient is .94. Alternate form reliability is .89. Test-retest reliability is .77 (corrected for the effects of restriction of range). For the Level 6 Passage Comprehension subtest, split-half reliability coefficient is .94. Alternate form reliability is .88. Test-retest reliability is .94 (corrected for the effects of restriction of range).	Alternate form reliabilities range from .83 to .87. Test-retest reliabilities range from .85 to .88 (corrected for the effects of restriction of range).	Internal consistency reliabilities (Cronbach's Alpha) for the four tests are: .85 for the grade 5 science test .84 for the grade 5 social studies test .82 for the grade 6 science test .80 for the grade 6 social studies test
Validity	Evidence of content, criterion-related, and construct validity.	Evidence of content, criterion-related, and construct validity.	Not provided.
Grade Range	PK – 12	2 – 12	5 and 6
Age Range	Not provided.	7.0 – 18.11	Not provided.
Number of Test Items	Six passages, each with six questions.	Twelve printed passages that become progressively more difficult in their content, vocabulary, and grammar.	Five passages, each with six questions.
Average Passage Length	Level 5, Form A – 158 words Level 6, Form A – 195 words	NA	Grade 5: science test – 391 words; social studies test – 454 words Grade 6: science test – 559 words; social studies test – 563 words

Table I.6 *(continued)*

Characteristic	Group Reading Assessment and Diagnostic Evaluation (GRADE), Passage Comprehension Subtest	Test of Silent Contextual Reading Fluency (TOSCRF)	Educational Testing Service (ETS) Social Studies/Science Reading Comprehension Assessments
Readability Scores	Level 5, Form A: Flesch-Kincaid grade levels range from 3.9 to 8.5. Mean=6.1. Lexile measures range from 510 to 1130. Mean=803. Level 6, Form A: Flesch-Kincaid grade levels range from 4.5 to 7.5. Mean=6.4. Lexile measures range from 630 to 1040. Mean=903.	NA	Grade 5 science passages: Flesch-Kincaid grade levels range from 3.7 to 6.2. Mean=5.5. Lexile measures range from 590 to 930. Mean=850. Grade 5 social studies passages: Flesch-Kincaid grade levels range from 4.6 to 5.6. Mean=5.2. Lexile measures range from 680 to 790. Mean=748. Grade 6 science passages: Flesch-Kincaid grade levels range from 4.0 to 9.9. Mean=7.1. Lexile measures range from 920 to 1050. Mean=1002. Grade 6 social studies passages: Flesch-Kincaid grade levels range from 4.2 to 11.6. Mean=8.1. Lexile measures range from 750 to 1330. Mean=1042.
Test Time	The subtest is untimed, but the estimated time for completion is 25 minutes.	3 minutes	The tests are untimed, but the estimated time for completion is 30 minutes.

SOURCES: Hammill et al., Test of Silent Contextual Reading Fluency (TOSCRF), Examiner's Manual, Austin, TX: Pro Ed, 2006; Williams, K. T., Group Reading Assessment and Diagnostic Evaluation (GRADE) Technical Manual, Circle Pines, MN: American Guidance Service, Inc., 2001. Information about the science and social studies tests was provided by ETS in a technical report.

NA = not available.

- ***The Passage Comprehension subtest of the Group Reading Assessment and Diagnostic Evaluation (GRADE).*** The GRADE (published by Pearson Learning Group) is a multiple-choice, paper-and-pencil, group-administered, untimed test that measures baseline skills and student improvement in critical reading areas (Williams 2001). The Passage Comprehension subtest measures the ability to comprehend extended text as a whole, using short passages in different genres and questions that "incorporate the metacognitive strategies of questioning, predicting, clarifying, and summarizing, as well as inclusion of a variety of sentence structures" (http://www.pearsonlearning.com). A response rate of 95 percent was achieved.

- ***Test of Silent Contextual Reading Fluency (TOSCRF).*** This paper-and-pencil, group-administered, timed test measures skills such as word identification, word meaning, and sentence structure, all of which are important for reading comprehension. Commonly known as the "slasher test," this assessment presents words using uppercase letters without any spaces or punctuation and requires students to insert slashes between letters to distinguish words (http://www.proedinc.com). Since the test allows students only three minutes for completion, it was conducted on the same day as the baseline GRADE test. Ninety-four percent of students completed the TOSCRF test.

The study team also asked schools to provide data on each student. Although these data were collected at the end of fifth grade, some stable items that serve as baseline student characteristics were obtained. The data included date of birth, gender, race/ethnicity, ELL and disability status, and eligibility for free or reduced-price lunch. Districts abstracted most or all of these data from their databases, with some data gathered manually by school staff or local study team staff. Overall, we obtained records for 96 percent of students.

(iii) Data Used to Measure Student Outcomes

Data on students' post-test outcomes were collected from two sources at the end of the fifth-grade year (between April and June 2007). First, students were retested using the GRADE (Williams 2001) and an 88 percent completion rate was achieved. Second, students were tested for comprehension of social studies and science text, using assessments developed specifically for the study.

The Educational Testing Service (ETS) developed tests to assess comprehension of informational text, drawing from its item bank and creating some new items (Educational Testing Service 2007a and 2007b). The multiple-choice, paper-and-pencil, group-administered, untimed assessments included either social studies or science passages. The questions asked about the passages' main idea, significant details, vocabulary, and author's purpose, and asked students to draw inferences. To reduce burden, half the students were randomly assigned to take the science test and half to take the social studies test. Generally, the tests were conducted within the same week (but not on the same day) in which the GRADE was administered. Eighty-seven percent of students completed the science or social studies test.

	Cohort 1 Students	Cohort 2 Students
Study Year 1 (2006-2007 school year)	• Cohort 1 students enter study as fifth graders • Interventions implemented with Cohort 1 treatment students • Administer pre-tests and post-tests	• Not yet included in study
Study Year 2 (2007-2008 school year)	• Cohort 1 students remain in study as sixth graders • Interventions are not implemented with Cohort 1 students • Administer follow-up tests	• Cohort 2 students enter study as fifth graders • Interventions implemented with Cohort 2 treatment students • Administer pre-tests and post-tests

2. Second-Year Study Design

The second year of the study was based largely upon the structure of the study's first year, but with three key distinctions:

1. ***Fewer curricula included in the fifth-grade component of the study's second year.*** Nine of the 18 schools randomly assigned to implement Reading for Knowledge elected not to continue implementing it in their fifth-grade classrooms in the second year of the study. Due to this attrition, Reading for Knowledge was not included in the fifth-grade component of the study's second year (in which fifth-grade teachers in treatment schools implemented the study interventions and their Cohort 2 students' outcomes were compared to outcomes of Cohort 2 fifth-grade students in control schools; see text box for a summary of the two second year study components).

2. ***Fewer schools participating in the fifth-grade component of the study's second year.*** Because 18 Reading for Knowledge schools were not included and because 10 other schools decided not to continue participating in the study's second year,[23] there were fewer schools participating in the fifth-grade component of the study's second year. In total, 61 schools (of the 89 that participated in Year 1) participated in the fifth-grade component of the second study year.

[23]Two Project CRISS schools (out of 17), two ReadAbout schools (out of 17), five Read for Real schools (out of 16), and one control school (out of 21) decided not to continue participating in the study's second year.

3. ***More schools participating due to the study's sixth-grade component (in which follow-up tests were administered to Cohort 1 students at the end of the 2007-2008 school year).*** Because many Cohort 1 students were attending different schools in sixth grade, a large number of schools were added to the study to facilitate the administration of follow-up tests to the first cohort of students. Cohort 1 students attended a total of 252 schools in the study's second year, 176 of which permitted the study team to conduct follow-up student testing for this study component.[24]

Second Year Study Components at a Glance

- **Fifth-grade component** – In this component, a second cohort of fifth-grade students from a subset of the study's original schools was added to the study, maintaining the original treatment assignments. Fifth-grade teachers in treatment schools implemented their assigned interventions and fifth-grade teachers in control schools continued teaching reading using methods they would have used in the absence of the study. Pre-tests and post-tests administered to students were used to assess the impact of the interventions on the second cohort of students. The rationale for including this component in the study is that impacts may be larger after schools and teachers have had one year of experience using the curricula.

- **Sixth-grade component** – In this component, the first cohort of students (all but 64 of whom were in sixth grade in the study's second year) was tracked for one additional year and follow-up tests were administered at the end of the school year to assess whether the interventions had statistically significant impacts one year after the end of their implementation. Fourteen sixth-grade students (0.2 percent) had the same teacher in sixth grade as in fifth grade, but the study interventions were *not* implemented in the second year when first-cohort students were in sixth grade. There are two main rationales for including this component in the study: (1) it is possible that impacts of the interventions could emerge in the second year even after the intervention implementation has ended and (2) to examine whether the negative effects of Reading for Knowledge observed in the first year continued into the second year.

a. Interventions

As noted above, the fifth-grade component of the second year included three of the four interventions that had been included in the first year of the study. Project CRISS, ReadAbout, and Read for Real were included in the fifth-grade component of the second year of the study, but, as noted above, Reading for Knowledge was not because 9 of the 18 Reading for Knowledge schools elected not to continue implementing the intervention in the second year.

The design of the study did not call for the interventions to be implemented in the sixth-grade component of the study, and, indeed, the interventions were not implemented in that component. Rather, the design called for following first-cohort students for one additional year

[24]While we cannot rule out the possibility that the nonparticipation of 76 schools in the follow-up testing of sixth-grade students affected the findings from the study's sixth-grade component, there were no statistically significant differences in the percentage of study students attending these nonparticipating schools between the four treatment groups and the control group. This suggests that the study's impact findings should not be biased by these schools' nonparticipation in follow-up testing.

after the *end* of the implementation of the interventions in the study's first year (through the end of the 2007-2008 school year), to assess whether implementation in the study's first year had longer-term effects on students' outcomes (measured at the end of the study's second year). Because this component is focused on assessing impacts of the interventions implemented in the study's first year, impacts of all four interventions that were implemented in the first study year (including Reading for Knowledge) were estimated in the sixth-grade component of the second year of the study.

b. District and School Recruiting

The study team began recruiting school districts for the second year of the study in March 2007. For the fifth-grade component of the second year of the study, the team focused on the 10 districts and 89 schools that participated in the study in the first year, with the goal of recruiting all of them to participate in the second year. Ultimately, all 10 of the districts and 61 of the 89 schools participated in the second year. See Appendix B and Section d below for information on the number of schools participating in Year 2 by treatment group.

For the sixth-grade component of the study's second year, the team focused on recruiting all schools that the first cohort of students attended during the 2007-2008 school year. As noted above, this was a much larger number than the 89 schools that participated in the first year of the study, as many Cohort 1 students were attending different schools in the study's second year, due to either moving on to middle school to attend sixth grade or moving to a neighborhood served by a different school in the district. Ultimately, we were able to administer follow-up tests to Cohort 1 students in 176 of the 252 schools that Cohort 1 students attended in the study's second year. The 76 schools in which we were unable to administer follow-up tests were schools that included few study students, with an average of 7 study students per school (compared to an average of 33 study students per school in the 176 schools in which we were able to conduct follow-up testing).

c. Treatment and Control Groups

Schools participating in the fifth-grade component of the study's second year were in the same treatment or control group in the second year as in the first year. Students in the study's sixth-grade component were classified according to their treatment status from the study's first year. For example, students who attended Read for Real schools in the study's first year are in the Read for Real group in the analyses for the study's sixth-grade component, regardless of the school they attended in the study's second year. Likewise, students who attended control schools in the study's first year are in the control group for the analyses of the study's sixth-grade component. This enabled the study team to assess the longer-term effectiveness of the single year of curricula implementation provided to students in the first year of the study. Because of the way in which multiple elementary schools fed into a single middle school serving sixth-grade students, first-cohort students from the treatment group could attend school in sixth grade with first-cohort students from the control group. For example, a student who attended Read for Real school "A" in *fifth* grade and a student who attended control school "B" in *fifth* grade could have both attended middle school "C" in *sixth* grade. It therefore follows that treatment students might be in the same classrooms as control students in *sixth* grade. Following the example above, these

two students might have been taught by teachers "D" and "E" respectively in *fifth* grade, and—in *sixth* grade when they were both attending school "C"—might have both been in a classroom taught by teacher "F." Thirty percent of sixth-grade students attended the same school in sixth-grade as they did in fifth-grade (because their school's grade structure included sixth grade). Very few sixth-grade students (0.2 percent) had the same teacher in sixth grade as in fifth grade. As noted above, none of the sixth-grade students received instruction in the study interventions in sixth grade.

d. Study Participants

The sixth-grade component of the second year of the study included 6,349 students who were attending 252 schools. The fifth-grade component of the second year of the study included 4,142 students from 61 schools. Table I.7 shows sample sizes for each intervention group and the control group in the second year of the study (2007-2008 school year).

Consistent with the first year—given the types of districts and schools being recruited—the schools participating in the second study year were statistically significantly different from schools nationwide in several respects. As in the first year, the schools included in the fifth-grade component of the second year of the study were statistically significantly more disadvantaged, larger, and more urban than the average U.S. school, and included higher percentages of black and Hispanic students (Table I.8).

Because over 90 percent of the schools participating in the fifth-grade component of the second year of the study were schoolwide Title I schools, we also compared fifth-grade component study schools to schoolwide Title I schools in the U.S. to assess how similar they were to other Title I schools in the U.S. (Table I.9). Findings from those comparisons mirrored the findings presented above for schools participating in the first year of the study, showing that study schools were more likely than U.S. schoolwide Title I schools to be urban, to include a higher percentage of black students, and to include more students.

The study team conducted similar comparisons for schools participating in the sixth-grade component of the second year of the study (Tables I.10 and I.11). The pattern of findings was the same as that described above for the schools participating in the fifth-grade component. In particular, the schools included in the sixth-grade component of the second year of the study were statistically significantly more disadvantaged, larger, and more urban than the average U.S. school, and included higher percentages of black and Hispanic students (Table I.10). A similar pattern was observed when the sixth-grade component schools were compared to schoolwide Title I schools in the U.S. (Table I.11).

Some turnover of fifth-grade teachers was observed between the first and second study years. Table I.12 shows the number of teachers participating in the study in Year 1, the number participating in Year 2, and the number of Year 2 teachers that were either new to the study in the second year or were returning to the study for a second year after having participated in the study's first year. The percentage of Year 1 teachers that remained in the study for a second year ranged from 41 percent for Read for Real to 71 percent for the control group. The Read for Real percentage reflects the fact that 11 of the 16 Read for Real schools from the first year of the study continued participating in the second year. A higher percentage of schools continued

TABLE I.7

NUMBER OF STUDY DISTRICTS, SCHOOLS, TEACHERS, AND STUDENTS IN STUDY SAMPLE IN YEAR 2

Intervention	Number of Districts	Number of Schools	Number of Teachers	Number of Students[a]
Sixth-Grade Component (Cohort 1 Follow Up)				
Project CRISS	10	133[c]	439[d]	1,319
ReadAbout	10	114[c]	432[d]	1,245
Read for Real	9[b]	124[c]	412[d]	1,228
Reading for Knowledge	10	104[c]	420[d]	1,195
Control Group	10	142[c]	365[d]	1,362
Total	**10**	**252[e]**	**907[e]**	**6,349**
Fifth-Grade Component (Cohort 2 Post-Test)				
Project CRISS	10	15	49	1,201
ReadAbout	10	15	46	1,108
Read for Real	9[b]	11	31	639
Control Group	10	20	56	1,194
Total	**10**	**61**	**182**	**4,142**

[a]This number includes all consenting students in the analysis sample. In spring 2008 (the end of the second year of the study), across all treatment groups, 75-76 percent of Cohort 1 students were tested at follow up, and 88 percent of Cohort 2 students in the analysis sample were tested at post-test.

[b]One district did not have enough participating schools to include all four intervention groups. The interventions that were assigned in that district were selected randomly.

[c]This refers to the number of schools that Cohort 1 students attended in the second study year. While some Cohort 1 students remained in the same school in the second year, other students moved to a new school due to student mobility (for example, resulting from family relocation or matriculation to sixth grade). This resulted in a larger number of schools attended by Cohort 1 students in Year 2 than in Year 1. For example, Cohort 1 students in the ReadAbout intervention group attended 114 schools in the second year, compared to 17 in the first year.

[d]This refers to the number of science, social studies, and English/Language Arts teachers of Cohort 1 students in the second study year. For example, Cohort 1 students in the ReadAbout intervention group had 432 science, social studies, and English/Language Arts teachers in the second year, while Cohort 1 students in the Read for Real intervention group had 412 science, social studies, and English/Language Arts teachers in the second year.

[e]This total refers to the number of unique schools and teachers in Year 2 that are linked to sixth graders from Cohort 1. Because some Cohort 1 students from different treatment groups in Year 1 were enrolled in school with and had the same teachers as Cohort 1 students from other treatment groups in Year 2, the total number of schools and teachers in this row does not correspond to the sum of schools or teachers across the treatment and control groups in the five rows above this number. Fourteen Cohort 1 students (0.2 percent) had the same teacher in sixth grade as in fifth grade. Across all treatment and control groups, 1,912 Cohort 1 students (30 percent) attended the same school in fifth and sixth grade (because some study schools included sixth grade). Note that the study interventions were not implemented in any sixth-grade classrooms.

TABLE I.8

CHARACTERISTICS OF SCHOOLS IN THE FIFTH-GRADE COMPONENT OF THE SECOND YEAR OF THE STUDY

Characteristics	U.S. Schools[a]	Schools in Study	Difference	*p-value*
Schools Receiving Title I (Percentage)				
Title I Eligible School[b]	74.3	98.4	-24.1*	0.00
Schoolwide Title I[b]	49.5	95.1	-45.6*	0.00
School Location (Percentage)[c]				
Urban	28.9	68.9	-40.0*	0.00
Urban fringe	30.0	18.0	12.0*	0.04
Town and rural area[d]	41.1	13.1	28.0*	0.00
Students per Teacher (Average)	14.9	16.1	-1.2	0.51
Number of Students per School (Average)	449.8	574.0	-124.2*	0.00
Students Eligible for Free or Reduced-Price Lunch (Percentage)[e]	49.3	73.6	-24.3*	0.00
Student Race/Ethnicity (Percentage)[f]				
White	56.2	24.2	32.0*	0.00
Black	16.3	39.2	-22.9*	0.00
Hispanic	20.0	31.9	-11.9*	0.00
Asian	4.1	1.9	2.2	0.07
Native American	2.0	0.9	1.0	0.39
GRADE Score (Average)	100.0	100.6	-0.6	1.00
Number of Schools	**50,933**	**61**		

SOURCE: 2005–2006 Common Core of Data (CCD). Data from the last row of the table are from two sources: (1) the study team's baseline GRADE test administration, and (2) national GRADE norm information provided by the GRADE test's developer.

[a]Data include regular primary and middle schools that reported having fifth-grade classrooms. Regular primary and middle schools are defined as public elementary/secondary schools that do not focus primarily on vocational, special, or alternative education.

[b]Data are missing for 1.7 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[c]Data are missing for 0.7 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[d]The town and rural area categories have been combined to protect school confidentiality.

[e]Data are missing for 3.6 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[f]Data are missing for 0.8 percent of regular primary and middle schools that reported having fifth-grade classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation.

*Statistically different at the .05 level.

TABLE I.9

CHARACTERISTICS OF SCHOOLS IN THE FIFTH-GRADE COMPONENT OF THE SECOND YEAR OF THE STUDY, COMPARED TO SCHOOLWIDE TITLE I SCHOOLS IN THE UNITED STATES

Characteristics	U.S. Schoolwide Title I Schools[a]	Schools in Study	Difference	*p-value*
Schools Receiving Title I (Percentage)				
Title I Eligible School	100.0	98.4	1.6*	0.00
Schoolwide Title I	100.0	95.1	4.9*	0.00
School Location (Percentage)[b]				
Urban	39.2	68.9	-29.7*	0.00
Urban fringe	22.0	18.0	4.0	0.45
Town and rural area[c]	38.8	13.1	25.7*	0.00
Students per Teacher (Average)	14.9	16.1	-1.2	0.51
Number of Students per School (Average)	456.8	574.0	-117.2*	0.00
Students Eligible for Free or Reduced-Price Lunch (Percentage)[d]	69.3	73.6	-4.3	0.11
Student Race/Ethnicity (Percentage)[b]				
White	38.2	24.2	14.0*	0.00
Black	24.1	39.2	-15.1*	0.00
Hispanic	30.7	31.9	-1.3	0.77
Asian	3.3	1.9	1.3	0.24
Native American	2.5	0.9	1.5	0.27
Number of Schools	**24,779**	**61**		

SOURCE: 2005–2006 Common Core of Data (CCD).

[a]Data include regular primary and middle schools that reported having fifth-grade classrooms and that are schoolwide Title I eligible schools. Regular primary and middle schools are defined as public elementary/secondary schools that do not focus primarily on vocational, special, or alternative education.

[b]Data are missing for 0.6 percent of regular primary and middle schools that reported having fifth-grade classrooms.

[c]The town and rural area categories have been combined to protect school confidentiality.

[d]Data are missing for 1.4 percent of regular primary and middle schools that reported having fifth-grade classrooms.

*Statistically different at the .05 level.

TABLE I.10

CHARACTERISTICS OF SCHOOLS IN THE SIXTH-GRADE COMPONENT
OF THE SECOND YEAR OF THE STUDY

Characteristics	U.S. Schools[a]	Schools in Study	Difference	*p-value*
Schools Receiving Title I (Percentage)				
Title I Eligible School	70.2	81.9	-11.8*	0.00
Schoolwide Title I	65.7	77.5	-28.9*	0.00
School Location (Percentage)[b]				
Urban	25.8	67.8	-42.0*	0.00
Urban fringe	26.8	17.0	9.8*	0.01
Town	12.0	2.6	9.4*	0.00
Rural area	35.5	12.6	22.9*	0.00
Students per Teacher (Average)	15.2	16.6	-1.4	0.17
Number of Students per School (Average)	480.7	649.3	-168.6*	0.00
Students Eligible for Free or Reduced-Price Lunch (Percentage)[c]	49.2	63.2	-14.1*	0.00
Student Race/Ethnicity (Percentage)[b]				
White	57.8	29.5	28.3*	0.00
Black	15.6	33.3	-17.6*	0.00
Hispanic	19.2	30.9	-11.7*	0.00
Asian	3.7	3.7	0.0	0.92
Native American	2.5	0.8	1.7*	0.02
GRADE Score (Average)	100.0	100.0	0.0	1.00
Number of Schools	**35,687**	**230**		

SOURCE: 2005–2006 Common Core of Data (CCD). Data from the last row of the table are from two sources: (1) the study team's baseline GRADE test administration, and (2) national GRADE norm information provided by the GRADE test's developer.

[a]Data include regular primary and middle schools that reported having sixth-grade classrooms. Regular primary and middle schools are defined as public elementary and secondary schools that do not focus primarily on vocational, special, or alternative education.

[b]Data are missing for 0.9 percent of regular primary and middle schools that reported having sixth-grade classrooms.

[c]Data are missing for 2.2 percent of regular primary and middle schools that reported having sixth-grade classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation.

*Statistically different at the .05 level.

TABLE I.11

CHARACTERISTICS OF SCHOOLS IN THE SIXTH-GRADE COMPONENT OF THE SECOND YEAR OF THE STUDY, COMPARED TO SCHOOLWIDE TITLE I SCHOOLS IN THE UNITED STATES

Characteristics	U.S. Schoolwide Title I Schools[a]	Schools in Study	Difference	*p-value*
Schools Receiving Title I (Percentage)				
Title I Eligible School	100.0	81.9	18.1*	0.00
Schoolwide Title I	100.0	77.5	22.5*	0.00
School Location (Percentage)[b]				
Urban	36.1	67.8	-31.2*	0.00
Urban fringe	20.3	17.0	3.3	0.21
Town	11.5	2.6	8.9*	0.00
Rural area	32.1	12.6	19.5*	0.00
Students per Teacher (Average)	15.5	16.6	-1.0	0.28
Number of Students per School (Average)	474.3	649.3	-175.0*	0.00
Students Eligible for Free or Reduced-Price Lunch (Percentage)[c]	69.3	63.2	6.0*	0.00
Student Race/Ethnicity (Percentage)[b]				
White	38.6	29.5	9.1*	0.00
Black	23.2	33.3	-10.0*	0.00
Hispanic	30.7	30.9	-0.2	0.92
Asian	3.2	3.7	-0.4	0.49
Native American	3.2	0.8	2.4*	0.01
Number of Schools	**16,121**	**230**		

SOURCE: 2005–2006 Common Core of Data (CCD).

[a]Data include regular primary and middle schools that reported having sixth-grade classrooms and that are schoolwide Title I eligible schools. Regular primary and middle schools are defined as public elementary and secondary schools that do not focus primarily on vocational, special, or alternative education.

[b]Data are missing for 0.9 percent of regular primary and middle schools that reported having sixth-grade classrooms.

[c]Data are missing for 2.2 percent of regular primary and middle schools that reported having sixth-grade classrooms.

*Statistically different at the .05 level.

TABLE I.12

FIFTH-GRADE TEACHERS IN STUDY SAMPLE IN YEARS 1 AND 2, BY EXPERIMENTAL CONDITION

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge
Year 1					
Total Number of Teachers Participating in the Study	59	52	50	54	53
Year 2					
Number of Year 1 Teachers Remaining in the Study	42	35	34	22	n.a.
Number of Teachers Who are New to the Study	12	18	12	9	n.a.
Total Number of Teachers Participating in the Study	54	53	46	31	n.a.

n.a. = not applicable.

participating in the second year in the control group (20 of 21 schools), the CRISS group (15 of 17 schools), and the ReadAbout group (15 of 17 schools).

We conducted statistical tests of the differences in the percentages of teachers remaining in the study across groups to address the potential concern that the interventions had an impact on the percentage of teachers that remained in the study. Findings from two sets of analyses suggest that the interventions did not affect teacher attrition. In the first analysis—in which we compared the percentages of teachers remaining in the study between the treatment and control groups—there was one statistically significant difference (fewer Read for Real teachers than control group teachers remained in the study). In the second analysis—in which we repeated the first analysis while restricting the sample to schools that participated in the study in both years—there were no statistically significant differences in the percentages of teachers remaining in the study across the groups. This second comparison was important because a larger number of schools in the Read for Real group elected not to participate in the second year of the study compared to the other groups. Taken together, these analyses suggest that the interventions had no impact on teacher attrition, and that that the one significant difference observed between the Read for Real group and the control group was due to schools (not teachers) leaving the study.

e. The Sample Design Ensured an 80 Percent Probability of Detecting Impacts in the Study's Second Year of at Least 0.14 Standard Deviations for Fifth Graders and 0.25 Standard Deviations for Sixth Graders

We were able to detect impacts of individual interventions on post-test scores of the second cohort of students of at least 0.14 standard deviations with 80 percent probability. This minimum

detectable effect is based on an intraclass correlation, school- and student-level R^2 values that were calculated using regression adjustment, and an adjustment for multiple comparisons.[25] With respect to teacher practices in the fifth-grade component of the study's second year, the study had less power due to smaller sample sizes of teachers and larger intraclass correlations (in the range of 0.46 to 0.50). For example, the smallest difference on the Reading Strategy Guidance scale of a single intervention that the study could detect with 80 percent probability was 0.91.[26] For the regression-adjusted impacts on follow-up scores of the first cohort of students, we were able to detect effects of at least 0.25 standard deviations with 80 percent probability.[27]

f. Data Collection

The data collected in the second year of the study differed from the data collected in the first study year depending on the component of the study. The sixth grade component of the study's second year included two data collection activities. First, we administered follow-up tests to students at the end of the 2007-2008 school year (when first-cohort students were in sixth grade), which was approximately one year after the end of the intervention implementation. This permitted the study team to examine whether there were impacts of the interventions in the second year, after Cohort 1 students were no longer using the interventions.

The tests administered at follow up included:

- **GRADE Passage Comprehension Subtest.** A response rate of 76 percent was obtained.
- **ETS assessments of reading comprehension in science and social studies for sixth-grade students.** Students were assigned to take the test in the same content area in which they took the test in Year 1. Seventy-five percent of students completed either the science or social studies test.[28]

Second, we administered a teacher survey to all sixth-grade teachers who taught English/language arts, science, or social studies. This survey gathered information on teachers'

---

[25]We obtained an intraclass correlation of 0.12, a school-level R^2 of 0.94, and a student-level R^2 of 0.51 for post-test scores of the second cohort of students.

[26]The minimum detectable effects reported for impacts on teacher practices are the effects that the study could detect with 80 percent probability (the standard level of power for reporting minimum detectable effects). The study could detect smaller effects with lower probability. Therefore, some of the reported statistically significant impacts are smaller than the effect sizes stated here. In addition, the minimum detectable effects reported in this paragraph are effects for the *average* intervention. Some of the interventions might have larger or smaller minimum detectable effects depending on the sample sizes by intervention.

[27]We obtained an intraclass correlation of 0.18, a school-level R^2 of 0.77, and a student-level R^2 of 0.36 for follow-up test scores of the first cohort of students.

[28]See Appendix E for information on response rates by treatment group for both second year study components.

education background, teaching experience, and certification. A response rate of 54 percent was obtained.

In the fifth-grade component of the study's second year, the study team essentially repeated the first year of the study with a new cohort of fifth-grade students.[29] The same data were collected, with four exceptions. In the second year:

1. The study team did not collect school-level data using the school information form. Instead, information on schools was collected from the *Common Core of Data* (National Center for Education Statistics 2008).

2. We administered a survey to fifth-grade teachers to obtain information on the amount of time students in their class spent using informational text in a typical week. Treatment group teachers were also asked to indicate how much time their students spent using the study curricula in a typical week. This form allowed the study team to examine (1) the extent to which treatment teachers were using the study curricula and (2) whether the interventions affected the amount of time students spent working with informational text. Eighty-five percent of teachers completed this form.

3. We asked teachers to fill out a form indicating how they allocated their time during a given school day. The time log was designed to show how much time teachers spent on various activities, including time spent on reading activities and time spent on the study curricula (for treatment teachers). This form also allowed treatment teachers to report whether any activities needed to be eliminated or reduced to make room for the implementation of the study curricula. This form allowed the study team to (1) assess whether the interventions affected the type of activities in which teachers were engaged (and the amount of time spent on those activities) and (2) determine whether or not teachers reduced the amount of time devoted to a particular activity, or eliminated an activity entirely, so that instruction in the study curricula could be provided (including how much time on average teachers reported reducing those activities). Eighty-nine percent of teachers completed this form.

4. The teacher survey administered in the first year of the study was not administered to all fifth-grade teachers. In the second year, it was only administered to teachers who participated in the study in the first year but who had not completed the survey during that year.

In the fall of 2007, 97 percent of Cohort 2 students took each of the pre-test assessments—the GRADE Passage Comprehension Subtest and the TOSCRF. Response rates on the GRADE and ETS comprehension assessments administered in spring 2008 of the second year to Cohort 2 students were 88 percent. Response rates on the classroom observations were 92 percent on the ERC and ranged from 81 to 93 percent on the fidelity forms, with an overall response rate of 88 percent across all fidelity forms.

[29]See section 1 above for information on the data collection conducted in the first year of the study.

This page is intentionally left blank.

II. IMPLEMENTATION FINDINGS

In impact studies, understanding the extent and quality of implementation can help researchers interpret statistically significant impact results (or the absence of impacts), form hypotheses about whether and how subsequent implementation experiences might yield different impact results, and understand whether schools are able to implement the interventions in a way that is consistent with developers' recommendations.

In this study, implementation is measured from two perspectives, as recommended by Gersten et al. (2005). The first, and most common, perspective focuses on assessing the extent to which teachers demonstrate adherence to procedures or practices deemed critical for implementing a particular intervention.[30] On this study, the developers specified the set of practices deemed essential to implementation, from which the study team developed fidelity forms that could be used by observers in the classroom to capture whether teachers were engaging in these practices. This approach is appealing because it corresponds to the common understanding of faithful "program implementation," and the forms can be easy for observers to complete.

However, this method also has several drawbacks (Gersten et al. 2000, 2005; Desimone 2002). Developers often find it difficult to identify the critical elements of their intervention with precision. Some developers' materials are detailed and exacting, while others allow teachers great latitude. These differences correspond to variation in the level of detail that observers can be asked to look for in the classroom. As a result, 80 percent implementation of Intervention A may not be equivalent to, or as difficult to achieve as, 80 percent implementation of Intervention B. In addition, some programs provide teachers with menus of options to choose from for part of the lesson (e.g., choosing either a vocabulary development or writing activity—or both—for small group follow-up instruction, depending on time allocations for that day). Differences in quality of implementation may also go unnoted with procedural checklists. Two teachers may achieve identical scores, one following procedures in a rote fashion and the other in a dynamic and engaging fashion (Gersten et al. 2005).

The alternative perspective involves a common observational system to assess teaching practices, regardless of the details of the curricula observed. For example, the Project Follow Through implementation study of seven instructional models (Stallings 1975) used a common observational procedure to describe reading and mathematics instruction in classrooms operating under the seven intervention models as well as control group classrooms.

Researchers have used this approach to examine the instructional practices associated with enhanced academic outcomes, using the same definition of practices, regardless of the intervention (for example, Cooley and Leinhardt 1980; Rosenshine and Stevens 1986; Dynarski et al. 2007; Glazerman et al. 2008). In a multi-treatment impact study, consistent definitions of

[30] O'Donnell (2008) defines fidelity of implementation as "the determination of how well an intervention is implemented in comparison with the original program design" (pp. 33-34).

instructional practices make it possible to use observational measures of implementation to describe how the various treatments differ from each other and from the control condition, and to use them as mediating variables in the impact analysis.

Both approaches were used in this evaluation. We developed and used a procedural fidelity form for each of the four interventions to gauge whether teachers actually followed the procedures specified by the developers. This form did not rate or rank the quality of implementation of a procedure. Instead, it measured the absence or presence of the procedures specified by the developers. We also developed a common observational system for use in all intervention and control classrooms when students and teachers were working with informational text, to record the frequency of teaching practices that earlier small-scale experimental research suggested were associated with enhanced comprehension outcomes.

In Sections A and B below, we summarize the features of the four interventions and the extent of preparation and training the teachers in the intervention classrooms received. Section C presents results from the intervention-specific fidelity analysis, focusing on two aspects of fidelity: (1) fidelity in the second year of intervention implementation for the study and (2) comparisons of fidelity between the first and second years of implementation of the study interventions. Section D presents descriptive information on teacher practices in the second year of implementing the interventions, including comparisons of educational practices across treatment and control groups using three scales derived from the observational data. This section also includes comparisons of instructional practices between the first and second years of the study. Section E presents information on teachers' allocation of time in the second year of the study, including the amount of time students typically spent using informational text in a typical week and the amount of time treatment group teachers spent using their assigned intervention in a given day.

A. INTERVENTION FEATURES

All four study interventions share a set of common comprehension strategies, instructional strategies, and student activities, but there are some differences in emphasis (Table II.1) and cost. All of the interventions focus on teaching students four core reading comprehension strategies (although they are not always labeled in the same way):

- ***Elements of text structure.*** This strategy involves an awareness of the structure and organizational elements of text and how they can be used to enhance comprehension of text. Elements of text structures[31] in informational text include headings, subheadings, visuals, and graphics, and organizational elements include cause and effect, compare and contrast, problem and solution, and sequencing. Project CRISS calls this strategy "author's craft." ReadAbout refers to "reading skills," while Read

[31]There is variation in the terminology used to describe these strategies by the developers on this study. For example, headings and subheadings are categorized as text *structures* by CRISS but as text *features* by Reading for Knowledge.

TABLE II.1

SUMMARY OF READING COMPREHENSION PROGRAMS

Program/ Developer	Program Focus	Teacher Training	Instructional Components[a]	Student Materials
Project CRISS/ CRISS	Focuses on five metacognitive Keys to Learning to help students become strategic learners: (1) background knowledge; (2) purpose setting; (3) author's craft (text structure); (4) active involvement (writing, discussion); and (5) organization (transforming information using writing and graphic organizers).	18 hours of initial training, 6 hours of follow-up training. Monthly trainer visits to each school to observe teachers and provide feedback. CRISS Cornerstones manual and DVD provide follow-up lessons for teacher learning community teams. Includes administrator and parent training components. <u>Year 2</u> • New and returning teachers participated in the initial and follow-up training listed above. • Building facilitators (CRISS leaders who assist other teachers with implementation) received 3 additional hours of training.	• Teacher's edition of *Learning How to Learn* provides detailed lesson plans for each chapter. Recommended use: 30-45 minutes per day. • Strategies are learned and practiced using *Tough Terminators,* a science trade book. • Uses variety of graphic organizers and note-taking, discussion, vocabulary, and writing strategies. • Students apply strategies to regular science and social studies texts.	Student book, *Learning How to Learn,* includes 19 chapters in a four-step format: (1) prepare, (2) be involved, (3) organize, and (4) apply. Each chapter focuses on two to four learning strategies.
ReadAbout/ Scholastic	Students are taught 10 comprehension skills: identifying author's purpose, identifying cause and effect, comparing and contrasting, drawing conclusions, distinguishing fact and opinion, locating main idea and details, making inferences, identifying problem and solution, sequencing events, and summarizing. Students also learn seven reading strategies: visualizing, setting a purpose, monitoring, rereading, summarizing, questioning, and repairing.	6 hours of initial training (plus access to the online course, *Improving Reading Comprehension)*, 6 hours of follow-up training in the fall, 6 hours of follow-up training in the spring. <u>Year 2</u> • New teachers participated in the initial and follow-up training listed above. • Returning teachers received 6 hours of refresher training in Year 2.	• Adaptive computer software used three times per week for 20 minutes. Software teaches comprehension skills, vocabulary, and content knowledge. • Students use offline materials once per week for 20 minutes. Offline materials include whole-class or small-group lessons on comprehension skills, vocabulary strategies, text types, or writing skills. Students rotate among computer, teacher-led, and independent reading groups. • Teacher materials include suggestions for English language learners and differentiated instruction.	Three core components are: (1) a software program, (2) SmartFile topic cards (supplemental print articles), and (3) a content library of science and social studies trade books. Reading passages are classified by three topics (science, social studies, and life), and five reading bands with Lexile ranges. Includes an assessment and writing topic at the end of each reading topic.

Table II.1 *(continued)*

Program/ Developer	Program Focus	Teacher Training	Instructional Components[a]	Student Materials
Read for Real/ Zaner-Bloser	Each unit focuses on (1) a Before Reading strategy (previewing, activating prior knowledge, or setting a purpose); (2) a During Reading strategy (making connections, interacting with text, or clarifying understanding); and (3) an After Reading strategy (recalling, evaluating, or responding).	12 hours of initial training, which includes an overview of research-based reading strategies, as well as training on using the curriculum. Follow up includes six hours of on-site training, plus telephone support and an online teacher support forum. <u>Year 2</u> • New teachers participated in the initial training listed above but did not receive follow-up training. • Returning teachers did not receive additional training in Year 2.	• Each unit has three reading selections for students to learn, practice, and apply a comprehension strategy. • Lessons take 30-45 minutes per day. • Teacher Guide includes a script for guiding reading and discussion of each story, activities for English language learners, writing activities, and comprehension tests.	*Read for Real* literacy series has six leveled books for grades three through eight. Each book has six units, and each unit has three reading selections. New vocabulary words are defined in sidebars, and a student "reading partner" in the text models thinking about each strategy. Vocabulary, writing, and fluency activities follow each reading selection. Includes unit tests and answer keys.
Reading for Knowledge/ Success for All (SFA)	Program focuses on four key comprehension strategies: (1) clarifying, (2) predicting, (3) summarizing, and (4) questioning. Includes vocabulary-building strategies in each lesson.	12 hours of initial training, 6 hours of follow-up training, and quarterly teacher meetings with SFA trainer. Four professional development videos guide teacher learning community meetings.	• Detailed daily lesson plans for 17 units (eight days each) covering 136 lessons. Lessons take 45 minutes per day. • Lessons follow same process: Set the stage; Active instruction; Teamwork (paired reading, team talk); and Reflection (teams share with class). • The four key strategies are introduced to students using video-based lessons. • Major cooperative learning component in the program.	Reading comprehension strategies are taught using a Student Edition for each strategy, a Video Viewing Guide, a set of science and social studies trade books, Strategy Practice sheets, and Strategy Cue cards to encourage transfer of skills to other content reading. Includes unit tests and answer keys.

[a]The amount of time reported for lessons is based on programs' recommended usage, not on actual usage by teachers in the study.

for Real calls this practice "interacting with text" and Reading for Knowledge considers these elements to be part of the "predicting strategy."

- ***Self-questioning.*** This strategy involves asking oneself questions about the text before, during, and after reading as a way to improve comprehension. Project CRISS and Read for Real call this "setting a purpose," while ReadAbout and Reading for Knowledge call this "questioning."

- ***Clarifying understanding.*** This strategy involves methods for clarifying the meaning of words, sentences, or passages that a student does not understand. These practices are called "fix-up strategies" by Project CRISS, "monitoring, rereading, or repairing" by ReadAbout, "clarifying understanding" by Read for Real, and simply "clarifying" by Reading for Knowledge.

- ***Summarizing.*** The summarizing strategy involves identifying the main ideas and important details in a passage and providing succinct summaries either verbally or in writing. Project CRISS, ReadAbout, and Reading for Knowledge call this summarizing, and Read for Real labels it "recalling."

Two of the curricula go beyond these four core strategies and provide students with additional comprehension tools (see box below for a summary of the intervention features discussed in this section). Project CRISS and Read for Real also teach students to think about what they already know concerning the topic before they start reading or while they are reading. They call this strategy variously "background knowledge," "activating prior knowledge," or "making connections."

All of these interventions also have certain instructional methods or student activities in common. For example, all of the curricula include teacher-directed instruction; such instruction can include explaining, modeling, and guided practice. Delivering the four interventions also involves student practice activities, such as having students read aloud or complete worksheets or graphic organizers.

Other instructional methods figure in three of the four curricula. Three of the programs (Project CRISS, ReadAbout, and Reading for Knowledge) have students practice their reading comprehension skills and strategies as they read selected science and social studies trade books.[32] All of the programs except Project CRISS provide assessments at the end of each unit. Two programs use technology as a teaching tool and for student practice—ReadAbout includes adaptive computer software so that students practice the comprehension strategies using text at the appropriate reading level for them and Reading for Knowledge includes four videotapes that introduce and model the program's four reading strategies. Reading for Knowledge also includes a cooperative learning component in which teachers track individual and team participation "points" to provide incentives for both individual and group effort.

---

[32]Trade books are books published for a general readership rather than specifically for the classroom and are distributed to the general public through booksellers.

SUMMARY OF INTERVENTION FEATURES

	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge
Comprehension Strategies				
Identification of text structure	√	√	√	√
Self-questioning	√	√	√	√
Clarifying understanding	√	√	√	√
Summarizing	√	√	√	√
Activating prior knowledge	√		√	
Instructional Methods and Student Activities				
Teacher-directed instruction	√	√	√	√
Student practice	√	√	√	√
End-of-unit assessments		√	√	√
Practice skills using content-area trade book(s)	√	√		√
Technology used as teaching tool		√		√
Cooperative learning component				√

Although the four curricula tested in the evaluation have much in common in terms of comprehension strategies, instructional methods, and student activities, they are offered to educators under different pricing structures (Table II.2). One developer includes all curriculum components in one price, while the others list separate prices for various curriculum components. For example, to implement Read for Real, districts would pay one price for all program materials (based on the number of participating classrooms), with teacher training and support included in that amount. To implement ReadAbout, districts would pay a per-classroom price that would encompass licenses, classroom kits, and initial training. For Project CRISS, on the other hand, districts would pay separate prices for training and for optional materials. The Reading for Knowledge developer was unable to provide a purchase price because the program was adapted from Success for All for the study and its pricing structure had not yet been determined.

Despite these differences in pricing arrangements, it is possible to discern how prices vary across curricula and for districts of different sizes. Costs for the intervention programs range from roughly $3,000 up to $187,000 per district, depending on the size of the school district and certain standardizing assumptions (Table II.3). Costs that would have been incurred by non-study districts to purchase these programs in the 2006-2007 or 2007-2008 school years range from about $3,000 to almost $14,000 for a sample small district to about $34,000 to $187,000 for a sample large district, after various discounts for districts with many schools have been considered. The costs for all the programs would drop after the first year, when materials have been purchased, software has been installed, and experienced teachers within the district may be able to provide some or all of the training. Costs would fall most dramatically for ReadAbout, since its licenses (the most expensive component of the program) are valid in perpetuity.

TABLE II.2

PROGRAM COSTS

Costs	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge
Base Cost	Program components are purchased separately	Licenses:[a] $6,000 for 60 students and two classroom kits; $9,500 for 100 students and three kits; $19,500 for 360 students and 12 kits. Kits include teacher materials (Topic Planners [cards that preview the ReadAbout software passages and vocabulary]; Know About ReadAbout Guide; assessments, reports, and the Differentiated Instruction Guide; the ReadAbout Software Manual; the SAM software manual; and the SAM Reference Guide) and classroom materials for students (SmartFiles [cards that extend the software], a poster, and Bonus Card Stickers). Each school receives Professional Papers, ReadAbout Installation Kits, and an SRI Installation Kit. Licenses also cover initial teacher training.	Program is purchased by buying the program materials: $475.75/ classroom (25 students); $18.99/extra copy	Program cost not yet known (the curriculum was adapted from Success for All for the study during the pilot year)
Costs Not Included in Base Cost				
Initial Training	$45/person if district provides trainer; $55/person with national trainer, plus $800/day per trainer (for two to four days of training), plus travel expenses[b]	No additional cost for the one day of initial training	No additional cost if entire district is participating; otherwise, $1,000/day (two days) per trainer, plus travel expenses	No additional cost for the two days of initial training
Follow-Up Training	$800/day trainer honorarium (for one to two days of training), plus the trainer's travel expenses[b]	$2,500/one-day training for up to about 20 teachers ($2,000/half-day seminar x two seminars = $4,000 x 37 percent discount for multiple seminars = $2,500; for two or more trainings, there is a 44 percent discount).	No follow-up training	No additional cost for the one day of follow-up training
Additional Services and Support	Parent workshop: Cost per booklet, $4 for 1-50 parents, $3 for 51-200 parents, and $2 for 201+ parents Email and telephone consultation were added for schools in the study.	$2,500/school for technology installation[c] $2,800/school for premium technical support (web, telephone, emails)[d]	A website with an electronic bulletin board, a helpdesk, and email or telephone consultation were added for schools in the study.	No additional cost for quarterly visits in which a trainer observes and then meets with each teacher to discuss goal setting, planning, and other feedback; or email and group teleconferencing
Materials	Optional: Classroom set, $550: one teacher's manual with Critterman DVD, 31 Tough Terminators (student book), and 30 student workbooks; extra student book, $10; extra student workbook, $8; video, $445 each; posters, $125/set of 30 posters; administrator materials, $55/each; Cornerstones (follow-up booklet and CD-ROM for teachers' independent use), $35/each	No additional materials	No additional materials	No additional materials

SOURCE: Developer Interviews: Reading Program Costs and Services.

[a]Licenses are valid in perpetuity.

[b]Districts typically use their own trainers after the first year. If insufficient capacity was built during the first year, however, districts can continue to pay for national trainers.

[c]Installation costs are a one-time fee.

[d]There is a premium technical support discount of 15 percent for 11 to 20 schools, 25 percent for 21 to 30 schools, 30 percent for 31 to 50 schools, 35 percent for 51 to 80 schools, and 40 percent for 81 or more schools.

TABLE II.3

ESTIMATED PROGRAM COSTS FOR TYPICAL SMALL, MEDIUM, AND LARGE DISTRICTS

District Size	Project CRISS (in Dollars)[a]		ReadAbout (in Dollars)[b]		Read for Real (in Dollars)[c]		Reading for Knowledge
Small (districts with < 2,500 students); assumptions: • One elementary school • Two fifth-grade teachers • 50 students and parents	0	Base cost	6,000	Base cost	952	Base cost	Program cost not yet known (the curriculum was adapted from Success for All for the study during the pilot year).
	2,510	Initial training	0	Initial training	2,000	Initial training	
	800	Follow-up training	2,500	Follow-up training	0	Follow-up training	
	200	Additional support	5,300	Additional support	0	Additional support	
	1,100	Materials	0	Materials	0	Materials	
	4,610	**Total**	**13,800**	**Total**	**2,952**	**Total**	
Medium (districts with 2,500-9,999 students); assumptions: • Four elementary schools • 12 fifth-grade teachers • 300 students and parents	0	Base cost	19,500	Base cost	5,709	Base cost	Program cost not yet known (the curriculum was adapted from Success for All for the study during the pilot year).
	3,060	Initial training	0	Initial training	2,000	Initial training	
	800	Follow-up training	2,500	Follow-up training	0	Follow-up training	
	600	Additional support	21,200	Additional support	0	Additional support	
	6,600	Materials	0	Materials	0	Materials	
	11,060	**Total**	**43,200**	**Total**	**7,709**	**Total**	
Large (districts with ≥10,000 students); assumptions: • 17 elementary schools • 68 fifth-grade teachers • 1,700 students and parents	0	Base cost	97,500	Base cost	32,351	Base cost	Program cost not yet known (the curriculum was adapted from Success for All for the study during the pilot year).
	8,540	Initial training	0	Initial training	2,000	Initial training	
	1,600	Follow-up training	6,720	Follow-up training	0	Follow-up training	
	3,400	Additional support	82,960	Additional support	0	Additional support	
	37,400	Materials	0	Materials	0	Materials	
	50,940	**Total**	**187,180**	**Total**	**34,351**	**Total**	

[a]Assumptions: A national trainer is provided for three days of initial training and one day of follow-up training; one trainer would be used for the small and medium district; two trainers would be used for the large district; the trainers' travel expenses would be in addition to the amounts shown. The optional classroom set is purchased.

[b]Assumptions: Licenses come in packets at $6,000 for 60 students, $9,500 for 100 students, and $19,500 for 360 students. The small district requires a set of 60 licenses, the medium district a set of 360 licenses, and the large district five sets of 360 licenses. The small and medium districts receive a 37 percent discount on the follow-up training, and the large district (which requires three follow-up trainings to train the 68 teachers) receives a 44 percent discount. The large district also qualifies for a 15 percent discount on premium technical support, since it has 17 schools.

[c]Assumptions: One trainer would be used for the small and medium district; two trainers would be used for the large district; the trainers' travel expenses would be in addition to the amounts shown.

B. TEACHER TRAINING AND SUPPORT

The training that prepares teachers to implement a new curriculum can be an important determinant of how well they deliver it, and thus whether and how it affects student outcomes. In this evaluation, developers trained teachers in the treatment group schools. Understanding this training and the extent to which teachers participated in it can inform our interpretation of the interventions' estimated impacts on student outcomes. This information also can contribute to our understanding of the observed differences in teacher practices between the treatment and control groups, since differences in practice could be expected to emerge only if a large percentage of teachers participated in the training (see Section D of this chapter for information on the comparison of treatment and control group teaching practices).

Initial Teacher Training. Implementing the interventions involved a considerable amount of support and training for teachers (Table II.4). This training and support varied across the two years of the study and across interventions. In the first year of the study, across the four interventions, the developers' training plans called for providing an average of 12 hours of initial training to prepare treatment group teachers to use the interventions. The initial training prescribed for the interventions ranged from 6 hours for ReadAbout to 18 hours for Project CRISS.[33]

In the second year of the study, all developers' training plans called for providing initial training to any teachers new to the study. The length of the training in the second year was the same as in the first year. In the second study year, two of the three developers' training plans also called for providing refresher training to returning teachers (Table II.4). Scholastic's training plans called for one day of refresher training for returning ReadAbout teachers. Project CRISS's training plans called for returning teachers to participate in the same initial training provided to teachers new to the study. A new three-hour training—for a subset of returning teachers identified to assist other teachers with their CRISS implementation—was also called for in the second year.

Follow-up Teacher Training During the School Year. In Year 1, all of the developers' training plans called for follow-up training to help support teachers during the school year (Table II.4). Across the four interventions, an average of 7.5 hours of follow-up training were prescribed by the developers of the interventions to help support and further build upon teachers' skills in the use of the interventions. Three programs (Project CRISS, Read for Real, and Reading for Knowledge) provided 6 hours of follow-up training, and ReadAbout provided 12 hours of follow-up training.

In Year 2, two of the three developers' training plans called for follow-up training (Table II.4). Six hours of follow-up training were prescribed for new and returning Project CRISS teachers. Twelve hours of follow-up training were prescribed for new ReadAbout teachers (no follow-up training was prescribed for returning ReadAbout teachers). Read for Real did not provide follow-up training in the second year.

---

[33]Two-thirds of initial training sessions were held before the school year started. The timeline for the initial training in both study years is shown in Appendix D.

TABLE II.4

SUMMARY OF TEACHER TRAINING

	Initial Training	Follow-Up Training and Ongoing Support
Project CRISS	18 hours of initial training, which includes 12 hours on using the strategies in the teacher's guide and 6 hours on using the student text and workbook. Teachers receive a training manual, teacher's guide, student text, and a wraparound edition of the student workbook. In Year 2, 18 hours of initial training were prescribed for new and returning teachers.	6 hours of follow-up training. Monthly trainer visits to each school to observe teachers and provide feedback. Developer encourages teachers to use biweekly study teams in which teachers review and discuss their use of CRISS strategies. In Year 2, 6 hours of follow-up training were prescribed for new and returning teachers. Year 2 training also included 3 hours of training for building facilitators (CRISS leaders who assist other teachers with implementation).
ReadAbout	6 hours of initial training covering program components (computer software, SmartFiles, Topic Planners), reading strategies, and test data interpretation. In Year 2, 6 hours of initial training were prescribed for new teachers. One day of refresher training was prescribed for returning teachers in Year 2.	12 hours of follow-up training (6 hours in the fall and 6 hours in the spring) to provide more in-depth understanding of program components and strategies and to provide instruction in using student data to make instructional decisions. In Year 2, 12 hours of follow-up training were prescribed for new teachers (no follow-up training was prescribed for returning teachers).
Read for Real	12 hours of initial training on connecting to prior knowledge, active reading strategies, vocabulary, text analysis, graphic organizers, Know-Want to Know-Learned (KWL), and using writing to assess comprehension. In Year 2, 12 hours of initial training were prescribed for new teachers. Refresher training was not provided to returning teachers in Year 2.	6 hours of follow-up training. Telephone support and online teacher support forum. Follow-up training was not provided in Year 2.
Reading for Knowledge[a]	12 hours of initial training, which includes an overview of the four critical comprehension strategies, as well as instruction in cooperative learning and monitoring strategy use.	6 hours of follow-up training. Developer encourages teachers to meet once per month to discuss program implementation. Each quarter, Success for All trainer attends teacher meetings, provides support and feedback (on-site and by phone), and observes reading and content area classes.

[a]Reading for Knowledge was not included in the fifth-grade component of the second year of the study.

Participation in Study Intervention Training. Over 90 percent of teachers participated in the initial training sessions provided by developers in Year 1 (Table II.5). A statistically significantly smaller percentage of teachers participated in the training sessions in Year 2 (ranging from 50 percent for Read for Real up to 91 percent for ReadAbout).

Compared to the other programs, a smaller percentage of Read for Real teachers participated in training in both study years. A statistically significantly lower percentage of Read for Real teachers (50 percent) participated in training relative to ReadAbout (91 percent) and Project CRISS (89 percent) (the *p*-values from tests comparing the percentage of teachers trained across treatment groups are not shown in the table). One potential explanation for this finding is that Read for Real developers did not provide make-up training for teachers who missed training sessions.

TABLE II.5

TEACHER TRAINING PARTICIPATION

	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge
Year 1				
Percentage of Teachers Trained[a]	100.0	100.0	91.2	96.8
Year 2				
Percentage of Teachers Trained[a]	88.8	91.4	49.5	n.a.
Difference Between Year 2 and Year 1				
Percentage of Teachers Trained	-11.2* (0.02)	-8.6* (0.02)	-41.7* (0.03)	n.a.
Number of Teachers, Year 1[b]	**52**	**50**	**54**	**53**
Number of Teachers, Year 2[b]	**49**	**46**	**8**	**n.a.**

SOURCE: Teacher training stipend claim forms.

NOTE: The *p-values* from tests of differences in Year 1 and Year 2 means are presented in parentheses. These tests account for clustering of teachers within schools.

[a]Three developers (Project CRISS, ReadAbout, and Reading for Knowledge) provided nonstandard training for teachers who missed the original training sessions. The nonstandard training involved working with teachers individually to cover content they missed.

[b]The number of teachers shown in this row is the number of teachers participating in the study, except for Read for Real in Year 2. In Year 2, the only Read for Real Teachers who were to receive training were those who were new to the study in the second year, so that is the number reported. For the other interventions, training was to be provided to all teachers in Year 2 (including teachers new to the study and those returning to the study for a second year).

n.a. = not applicable.

*Statistically different at the .05 level.

C. OBSERVED FIDELITY OF IMPLEMENTATION

Knowing the extent to which the interventions were implemented as intended is useful for interpreting impacts. Fidelity observations were conducted in spring of the 2006-2007 and 2007-2008 school years to assess whether treatment group teachers were implementing the procedures of the intervention assigned to their school (see Chapter I for more information). Fidelity observations were conducted in all treatment classrooms in which the teachers reported using the interventions.

We did not observe the handful of teachers in each intervention condition (two to four teachers per intervention in the second year of the study) who reported not using the interventions.[34] Fidelity observations were not conducted for these teachers because the goal of the fidelity analysis was to measure teachers' adherence to the specific set of procedures deemed important by developers for implementing each intervention model. Therefore, teachers who reported not implementing the interventions would not be adhering to the curriculum model if they happened to implement practices suggested by the curriculum model. (Data are not available to assess whether these teachers unintentionally implemented practices suggested by the curricula models.)

When analyzing the fidelity observation data, we assumed that these teachers did not implement any of the procedures listed on their assigned treatment group's fidelity form. This procedure was followed to ensure that the fidelity data reflect the full sample of teachers assigned to each intervention.

In the text that follows, we discuss the following key implementation findings:

- **In the spring of the second year of the study, over 80 percent (83 to 96 percent) of treatment teachers reported using their assigned curriculum.** Eighty-three percent of Read for Real teachers, 92 percent of Project CRISS teachers, and 96 percent of ReadAbout teachers reported using their assigned curriculum. The percentage of teachers who reported using each of the three interventions did not differ significantly between the first and second years of the study.

- **Classroom observation data from the second year of intervention implementation showed that teachers implemented 65 to 94 percent of the practices deemed important by the developers for implementing each curriculum.** On average, Project CRISS teachers implemented 65 percent of such practices and ReadAbout teachers implemented 94 percent of such practices. Read for Real teachers implemented 75 and 76 percent of such practices for the two types of instructional formats that comprise the program. There were no statistically

[34]The more general observations of teaching practices relating to vocabulary and comprehension instruction were conducted for these teachers.

significant differences in average fidelity levels between the first and second study years.[35]

- **Teachers who participated in both study years were more likely to report using their assigned curriculum than teachers new to the study in the second year.** These differences were statistically significant for ReadAbout (100 percent vs. 82 percent, *p*-value = .013) and Project CRISS (100 percent vs. 76 percent, *p*-value = .005), but not for Read for Real (90 percent vs. 80 percent).

- **Project CRISS teachers who participated in both study years were observed implementing statistically significantly more practices than Project CRISS teachers new to the study in the second year.** Project CRISS teachers who participated in both years were observed implementing 72 percent of the practices deemed important by developers, while teachers new to the study in Year 2 were observed implementing 51 percent of the practices (*p*-value = 0.008).

Below, we present information on the extent to which treatment group teachers were observed implementing the procedures of the study intervention on the day they were observed. We present this information separately for each intervention because each intervention had a set of intervention-specific practices that the developer deemed important for implementation. For each intervention, we present fidelity rates from the second year of intervention implementation. We then describe whether those rates differed significantly from those observed in the first year. In addition, we examine whether fidelity rates are higher for teachers who were in the study both years relative to teachers new to the study in the second year.

We report fidelity rates for each intervention in two ways: (1) the percentage of fidelity form items observed—restricted to items that fell within a section for which teachers were observed and (2) the percentage of all fidelity form items observed. We report findings both ways because it is not possible to determine the reason why items in a particular section were not observed. For example, if a behavior was not observed, we do not know if this was because the teacher should not have been implementing that behavior on that day or because the teacher forgot, or intentionally decided not, to implement it. However, with two of the interventions (Read for Real and ReadAbout), the calculations that are restricted to fidelity form items that fell within a section for which teachers were observed may be particularly relevant. According to the developer, Read for Real lessons follow a progression in which teachers must complete a given lesson before moving on to the next lesson. In some cases, teachers may not have had time to finish a lesson before the school day ended, therefore some sections would not have been observed by the study team's classroom observers. This does not necessarily mean the teacher implemented the curriculum poorly, as the program is designed to have teachers begin their lessons on the next day based on where they finished on the prior day. By presenting the percentage of items observed—restricted to items that fell within a section for which teachers

[35]The fidelity levels reported in this bullet for ReadAbout and Read for Real are based on fidelity form practices that fell within a window observed by the study's classroom observers. The fidelity levels reported for Project CRISS are based on all practices on the Project CRISS fidelity form because developers expected teachers to implement all fidelity form practices during each lesson.

were observed—the fidelity ratings may be more likely to be based on the sections of the lesson the teacher was completing that day. For example, if a teacher was working on the part of the lesson that involved discussing text after students read it, we would not expect to observe teachers engaging in before-reading activities such as activating prior knowledge.

The fidelity calculations based on the restricted set of items that fall within observed sections of the ReadAbout fidelity protocol are also relevant, but for a different reason. According to the developer, every ReadAbout lesson does not need to include all teaching practices shown on the fidelity form. In fact, some lessons might include two practices, while other lessons might include more. Therefore, presenting the percentage of items observed—restricted to items that fell within the sections for which teachers were observed—aims to focus the fidelity analysis on only those teaching practices being implemented that day.

The fidelity calculations based on the restricted set of items that fall within observed sections of the fidelity protocol are not necessarily relevant for Project CRISS. According to the developer, teachers using Project CRISS should complete all teaching practices during every lesson. Therefore, the most relevant analysis of the Project CRISS fidelity data is based on the percentage of all fidelity form items observed. If an item was not observed, it is likely that the teacher failed to exhibit that item when he or she should have done so.

We also report fidelity rates for each intervention for teachers who participated in both years of the study and for teachers who were new to the study in the second year. We conducted tests comparing fidelity rates for teachers who participated in both years of the study and teachers who were new to the study in the second year to examine how differences in experience implementing the study curricula are related to the quality of implementation.

1. Project CRISS

As noted above, according to the developer, each Project CRISS lesson should include all items that appear on the fidelity protocol. Therefore, in the text that follows, we focus on fidelity rates based on all fidelity form items (column 3 in Table II.6 and column 2 in Table II.7).

Project CRISS teachers were observed engaging in 65 percent of the key Project CRISS teaching practices in Year 2 (Table II.7). This percentage did not differ significantly from the 63 percent of teaching practices observed in Year 1 (Table II.8). In Year 2, Project CRISS teachers engaged most frequently in asking students to read a written text (92 percent), leading students in transforming information activities (86 percent), including informal or formal writing in transforming information activities (80 percent), and using transforming activities to teach the content of the lesson (76 percent) (Table II.6).

There were no statistically significant differences in the extent to which Project CRISS teachers engaged in individual key teaching practices between the first and second years of implementation (Table II.8).

TABLE II.6

FIDELITY OF IMPLEMENTATION OF INDIVIDUAL TEACHING PRACTICES FOR THE PROJECT CRISS CURRICULUM IN YEAR 2

	(1) Percentage of Teachers Observed Implementing Section[a]	(2) Among Teachers Implementing Section, Percentage of Teachers Observed Implementing Behavior	(3) Among All Teachers, Percentage of Teachers Observed Implementing Behavior[a]
Section I. Preparing for Understanding			
Provide instruction or lead activities to generate background knowledge about a topic or concept before students read about it	71.34	94.29	67.35
Help students set goals and determine a purpose before beginning to read		90.91	61.22
Section II. Engaging Students with Content and Transforming Information			
Have students read a written text		100.00	91.84
Lead students during and/or after reading in transforming information activities (for example, graphic organizer, guided discussion)		93.33	85.71
Include informal or formal writing in the transforming activities (including note taking)	91.84	92.86	79.59
Use the transforming activities to teach the content of the lesson		94.87	75.51
Discuss or reflect on students' metacognitive processes during the transforming activities		56.76	42.86
Section III. Reflecting on Content and Learning Processes			
Lead the whole class in a reflection discussion at the end of the lesson using questions such as: (A) Metacognition: How did you evaluate your comprehension? (B) Background knowledge: Did I assist you in thinking about what you already knew? (C) Purpose setting: Did you have clear purposes? (D) Active involvement: How were you actively engaged? (E) Discussion: How did discussion clarify your thinking? (F) Writing: How did you use writing to help you learn? (G) Transformation: What were the different ways you transformed information? How did this help you? (H) Teacher modeling: Did I do enough modeling?	16.33	100.00	16.33
Sample Size[b]		**53**	

Table II.6 *(continued)*

SOURCE: Classroom observations.

NOTE: There are two possible explanations for why teachers were not observed implementing a section of behaviors from the fidelity form: (1) a section might not have been appropriate given the stage of the lesson; or (2) the behaviors might have been appropriate, but the teacher failed to exhibit them. Because it is not possible to determine which explanation is the reason the behaviors in that section were not observed, we report fidelity rates under both assumptions. The findings in the second column can be viewed as an "upper bound" on fidelity (corresponding to the first explanation), while the findings in the third column can be viewed as a "lower bound" (corresponding to the second explanation).

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]The number of teachers presented in this row is the number participating in the study.

TABLE II.7

OVERALL FIDELITY OF IMPLEMENTATION FOR THE PROJECT CRISS CURRICULUM IN YEAR 2

	(1) Restricted to Behaviors in Observed Sections	(2) All Behaviors[a]
Percentage of Teachers Who Were Observed Implementing:		
80 to 100 percent of the CRISS fidelity form behaviors	71.11	30.61
40 to 79 percent of the CRISS fidelity form behaviors	28.89	61.22
0 to 39 percent of the CRISS fidelity form behaviors	0.00	8.16
Mean Percentage of the CRISS Fidelity Form Behaviors That Teachers Were Observed Implementing	83.71	65.05
Sample Size[b]	**53**	

SOURCE: Classroom observations.

NOTE: There are two possible explanations for why teachers were not observed implementing a section of behaviors from the fidelity form: (1) a section might not have been appropriate given the stage of the lesson; or (2) the behaviors might have been appropriate, but the teacher failed to exhibit them. Because it is not possible to determine which explanation is the reason the behaviors in that section were not observed, we report overall fidelity rates under both assumptions. The findings in the first column can be viewed as an "upper bound" on overall fidelity (corresponding to the first explanation), while the findings in the second column can be viewed as a "lower bound" (corresponding to the second explanation).

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]The number of teachers presented in this row is the number participating in the study.

TABLE II.8

FIDELITY OF IMPLEMENTATION FOR THE PROJECT CRISS CURRICULUM IN YEARS 1 AND 2

	Year 1	Year 2	Difference	*p-value*
Percentage of Teachers Who Reported Using Project CRISS	94.23	94.09	0.14	.979
Section I. Preparing for Understanding				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provide instruction or lead activities to generate background knowledge about a topic or concept before students read about it	67.31	68.00	0.69	.941
Help students set goals and determine a purpose before beginning to read	63.46	60.72	-2.74	.779
Section II. Engaging Students with Content and Transforming Information				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Have students read a written text	84.62	95.00	10.38	.139
Lead students during and/or after reading in transforming information activities (for example, graphic organizer, guided discussion)	82.69	87.53	4.84	.541
Include informal or formal writing in the transforming activities (including notetaking)	76.92	82.29	5.37	.536
Use the transforming activities to teach the content of the lesson	76.92	76.17	-0.75	.931
Discuss or reflect on students' metacognitive processes during the transforming activities	46.15	44.60	-1.55	.883
Section III. Reflecting on Content and Learning Processes				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Lead the whole class in a reflection discussion at the end of the lesson using questions such as:	—[b]	—[b]	—[b]	—[b]
(A) Metacognition: How did you evaluate your comprehension?				
(B) Background knowledge: Did I assist you in thinking about what you already knew?				
(C) Purpose setting: Did you have clear purposes?				
(D) Active involvement: How were you actively engaged?				
(E) Discussion: How did discussion clarify your thinking?				
(F) Writing: How did you use writing to help you learn?				
(G) Transformation: What were the different ways you transformed information? How did this help you?				
(H) Teacher modeling: Did I do enough modeling?				
Overall Fidelity				
Percentage of Teachers Who Were Observed Implementing:[a]				
80 to 100 percent of the fidelity form behaviors listed above	23.08	29.04	5.96	.398
40 to 79 percent of the fidelity form behaviors listed above	57.69	53.62	-4.07	.506
0 to 39 percent of the fidelity form behaviors listed above	19.23	17.34	-1.89	.913
Mean Percentage of the Fidelity Form Behaviors Listed Above That Teachers Were Observed Implementing	62.50	64.74	2.24	.617
Sample Size[c]	**54**	**53**		

SOURCE: Classroom observations.

Table II.8 (*continued)*

NOTE: The reported differences are regression adjusted to account for school effects. The Year 1 mean is the raw mean, the Year 2 mean is the Year 1 mean plus the regression-adjusted difference between years. Note that this explains the difference from the raw Year 2 means reported in Table II.7.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the behaviors listed in this table. In addition, the calculations presented in this table are based on all behaviors in all sections of the fidelity form, not only on the behaviors in the observed sections.

[b]Value suppressed to protect teacher confidentiality.

[c]The number of teachers presented in this row is the number participating in the study.

Project CRISS teachers who participated in both years of the study were statistically significantly more likely to report using Project CRISS than teachers who were new to the study in the second year (Table II.9, 100 percent vs. 76 percent, *p*-value = .005). Teachers who participated in both years of the study were observed implementing a statistically significantly higher mean percentage of practices advocated by the developer than teachers who were new to the study in the second year (72 percent vs. 51 percent, *p*-value = .008). Four statistically significant differences between teachers new to the study in the second year and teachers who participated in both years of the study were observed on individual teaching practices. Statistically significantly more teachers who participated in both years of the study than teachers new to the study in the second year were observed implementing the following practices: (1) having students read a written text (100 percent vs. 76 percent, *p*-value = .005), (2) leading students during or after reading in a transforming information activity (97 percent vs. 65 percent, *p*-value = .002), (3) using the transforming activities to teach the content of the lesson (87 percent vs. 59 percent, *p*-value = .030), and (4) discussing or reflecting on students' metacognitive processes during the transforming activities (53 percent vs. 24 percent, *p*-value = .048). There were no statistically significant differences among teachers who participated in both years of the study and teachers who were new to the study in the second year on the remaining items (which addressed teaching practices related to before- and after-reading activities).

2. Read for Real

As mentioned above, each Read for Real lesson follows a specified progression that must be completed before a new lesson can begin, and lessons may span two days. Therefore, in this section we focus primarily on the percentage of items observed restricting to items that fall within sections for which teachers were observed (columns 2 and 5 of Table II.10 and column 1 in Table 11).

The Read for Real intervention involved two types of instructional days, both of which were observed for the study. On Read for Real "Learn" days (days on which teachers modeled the comprehension strategies for students), Read for Real teachers were assessed based on 25 items. On Read for Real "Practice" days (days on which the teachers worked with students as they practiced the comprehension strategies), Read for Real teachers were assessed based on a similar protocol with 17 items (See Tables II.10, II.12, and II.13 for a list of the items included in the "Learn" and "Practice" day protocols).

Learn Days. On the "Learn" days in the second year of the study, restricting to items in sections for which teachers were observed, Read for Real teachers were observed engaging in 75 percent of the teaching practices deemed important by developers for implementing Read for Real (Table II.11). Restricting to items in sections for which teachers were observed, 100 percent of teachers were observed implementing three practices: (1) discussing the comprehension strategy with students, (2) reading or asking students to read the explanation of the "During Reading" strategy, and (3) reading or asking students to read about organizing information (Table II.10).

TABLE II.9

FIDELITY OF IMPLEMENTATION FOR THE PROJECT CRISS CURRICULUM, BY TEACHER EXPERIENCE WITH THE CURRICULUM

	Teachers in Year 2 Only	Teachers in Both Years	Difference	*p-value*
Percentage of Teachers Who Reported Using Project CRISS	76.47	100.00	23.53	.005*
Section I. Preparing for Understanding				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provide instruction or lead activities to generate background knowledge about a topic or concept before students read about it	52.94	73.33	20.39	.163
Help students set goals and determine a purpose before beginning to read	52.94	63.33	10.39	.496
Section II. Engaging Students with Content and Transforming Information				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Have students read a written text	76.47	100.00	23.53	.005*
Lead students during and/or after reading in transforming information activities (for example, graphic organizer, guided discussion)	64.71	96.67	31.96	.002*
Include informal or formal writing in the transforming activities (including notetaking)	64.71	86.67	21.96	.080
Use the transforming activities to teach the content of the lesson	58.82	86.67	27.85	.030*
Discuss or reflect on students' metacognitive processes during the transforming activities	23.53	53.33	29.80	.048*
Section III. Reflecting on Content and Learning Processes				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Lead the whole class in a reflection discussion at the end of the lesson using questions such as:	—[b]	—[b]	—[b]	—[b]
(A) Metacognition: How did you evaluate your comprehension?				
(B) Background knowledge: Did I assist you in thinking about what you already knew?				
(C) Purpose setting: Did you have clear purposes?				
(D) Active involvement: How were you actively engaged?				
(E) Discussion: How did discussion clarify your thinking?				
(F) Writing: How did you use writing to help you learn?				
(G) Transformation: What were the different ways you transformed information? How did this help you?				
(H) Teacher modeling: Did I do enough modeling?				

Table II.9 *(continued)*

	Teachers in Year 2 Only	Teachers in Both Years	Difference	*p-value*
Overall Fidelity				
Percentage of Teachers Who Were Observed Implementing:[a]				
80 to 100 percent of the fidelity form behaviors listed above	23.53	33.33	9.80	.491
40 to 79 percent of the fidelity form behaviors listed above	41.18	56.67	15.49	.318
0 to 39 percent of the fidelity form behaviors listed above	35.29	10.00	25.29	.035*
Mean Percentage of the Fidelity Form Behaviors Listed Above That Teachers Were Observed Implementing	50.74	72.08	21.34	.008*
Sample Size	**18**	**35**		

SOURCE: Classroom observations.

NOTE: These differences are not regression adjusted.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the behaviors listed in this table. In addition, the calculations presented in this table are based on all behaviors in all sections of the fidelity form, not only on the behaviors in the observed sections.

[b]Value suppressed to protect teacher confidentiality.

*Statistically different at the .05 level.

TABLE II.10

FIDELITY OF IMPLEMENTATION OF INDIVIDUAL TEACHING PRACTICES FOR THE READ FOR REAL CURRICULUM IN YEAR 2

	Learn Observation Days			Practice Observation Days		
	(1) Percentage of Teachers for Whom Behavior is Included in Observation Window[a]	(2) Among Teachers for Whom Behavior Falls in Observation Window, Percentage of Teachers Observed Implementing Behavior	(3) Among All Teachers, Percentage of Teachers Observed Implementing Behavior[a]	(4) Percentage of Teachers for Whom Behavior Is Included in Observation Window[a]	(5) Among Teachers for Whom Behavior Falls in Observation Window, Percentage of Teachers Observed Implementing Behavior	(6) Among All Teachers, Percentage of Teachers Observed Implementing Behavior[a]
Before Reading						
Reads or asks a student to read the explanation of the Before Reading focus strategy	75.00	91.67	61.11	42.86	83.33	35.71
Discusses the strategy with students	75.00	100.00	66.67	42.86	83.33	35.71
Reads or asks a student to read the information in the My Thinking box	75.00	66.67	44.44	n.a.	n.a.	n.a.
Asks students to apply the strategy	68.75	70.00	38.89	50.00	85.71	42.86
Discusses students' comments	n.a.	n.a.	n.a.	57.14	75.00	42.86
During Reading						
Reads or asks a student to read the explanation of the During Reading focus strategy	81.25	100.00	72.22	57.14	87.50	50.00
Discusses the strategy with the students	81.25	76.92	55.56	n.a.	n.a.	n.a.
Reads or asks a student to read the information in the My Thinking box (notes from the reading partner)	81.25	84.62	61.11	57.14	75.00	42.86
Asks students to share their thinking about the strategy	81.25	61.54	44.44	n.a.	n.a.	n.a.
Reminds students to write notes about the strategy	n.a.	n.a.	n.a.	64.29	88.89	57.14
Stops and addresses the My Thinking notes at the "red strategy buttons"	81.25	76.92	55.56	64.29	77.78	50.00
Reads and/or asks students to read the selection	56.25	81.82	68.75	64.29	100.00	64.29
After Reading						
Reads or asks a student to read the After Reading focus strategy	37.50	83.33	27.78	50.00	57.14	28.57
Discusses or asks questions about the strategy	37.50	83.33	27.78	42.86	50.00	21.43
Reads or asks a student to read the information in the My Thinking box	37.50	83.33	27.78	n.a.	n.a.	n.a.
Gives a written assignment highlighting the After Reading focus strategy	n.a.	n.a.	n.a.	42.86	50.00	21.43
Calls on students to implement the After Reading focus strategy	37.50	83.33	27.78	n.a.	n.a.	n.a.

Table II.10 *(continued)*

	Learn Observation Days			Practice Observation Days		
	(1) Percentage of Teachers for Whom Behavior is Included in Observation Window[a]	(2) Among Teachers for Whom Behavior Falls in Observation Window, Percentage of Teachers Observed Implementing Behavior	(3) Among All Teachers, Percentage of Teachers Observed Implementing Behavior[a]	(4) Percentage of Teachers for Whom Behavior Is Included in Observation Window[a]	(5) Among Teachers for Whom Behavior Falls in Observation Window, Percentage of Teachers Observed Implementing Behavior	(6) Among All Teachers, Percentage of Teachers Observed Implementing Behavior[a]
Comprehension						
Administers the open book comprehension test	18.75	—[b]	—[b]	35.71	—[b]	—[b]
Corrects tests with the class	18.75	—[b]	—[b]	28.57	0.00	0.00
Discusses responses	18.75	—[b]	—[b]	28.57	0.00	0.00
Organizing Information						
Reads or asks a student to read the information from the reading partner	18.75	100.00	16.67	n.a.	n.a.	n.a.
Discusses the graphic organizer	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Asks students to complete the graphic organizer	n.a.	n.a.	n.a.	35.71	80.00	28.57
Writing for Comprehension						
Reads or asks a student to read the information from the reading partner	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Reads or asks a student to read the summary	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Asks students to write a summary based on their completed graphic organizer	n.a.	n.a.	n.a.	—[b]	—[b]	—[b]
Identifies how the paragraphs and sentences in the summary correspond to the information on the graphic organizer	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Discusses the three parts of a summary	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Introduction	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Body	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Conclusion	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.
Sample Size[c]				**31**		

SOURCE: Classroom observations.

NOTE: Teachers could have started the lesson at any item/behavior ("start" item) in any section of the fidelity form and ended the lesson at any item/behavior ("end" item) in a subsequent section of the fidelity form. The window of observation consists of all the items in the interval from the "start" to the "end" item. There are two possible explanations for why teachers were not observed implementing behaviors outside their observation window: (1) behaviors outside of that window might not have been appropriate given the stage of the lesson; or (2) the behaviors might have been appropriate, but the teacher failed to exhibit them. Because it is not possible to determine which explanation is the reason behaviors outside the window were not observed, we report fidelity rates under both assumptions. The findings in the second and fifth columns can be viewed as an "upper bound" on fidelity (corresponding to the first explanation), while the findings in the third and sixth columns can be viewed as a "lower bound" (corresponding to the second explanation).

Table II.10 *(continued)*

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]Value suppressed to protect teacher confidentiality.

[c]The number of teachers presented in this row is the number participating in the study. Roughly half the teachers were observed on "Learn" days, and roughly half were observed on "Practice" days.

n.a. = not applicable.

TABLE II.11

OVERALL FIDELITY OF IMPLEMENTATION FOR THE READ FOR REAL CURRICULUM IN YEAR 2

	(1) Restricted to Behaviors in Observation Windows	(2) All Behaviors[a]
Learn Observation Days		
Percentage of Teachers Who Were Observed Implementing:[b]		
80 to 100 percent of the Read for Real fidelity form behaviors	50.00	0.00
0 to 79 percent of the Read for Real fidelity form behaviors[c]	50.00	100.00
Mean Percentage of the Read for Real Fidelity Form Behaviors That Teachers Were Observed Implementing	74.78	36.68
Practice Observation Days		
Percentage of Teachers Who Were Observed Implementing:[b]		
40 to 100 percent of the Read for Real fidelity form behaviors[d]	81.52	42.86
0 to 39 percent of the Read for Real fidelity form behaviors	18.18	57.14
Mean Percentage of the Read for Real Fidelity Form Behaviors That Teachers Were Observed Implementing	76.02	33.61
Sample Size[e]	**31**	

SOURCE: Classroom observations.

NOTE: Teachers could have started the lesson at any item/behavior ("start" item) in any section of the fidelity form and ended the lesson at any item/behavior ("end" item) in a subsequent section of the fidelity form. The window of observation consists of all the items in the interval from the "start" to the "end" item. There are two possible explanations for why teachers were not observed implementing behaviors outside their observation window: (1) behaviors outside of that window might not have been appropriate given the stage of the lesson; or (2) the behaviors might have been appropriate, but the teacher failed to exhibit them. Because it is not possible to determine which explanation is the reason behaviors outside the window were not observed, we report overall fidelity rates under both assumptions. The findings in the first column can be viewed as an "upper bound" on fidelity (corresponding to the first explanation), while the findings in the second column can be viewed as a "lower bound" (corresponding to the second explanation).

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]The vocabulary and fluency items have been left out of the table because developers noted they were not essential for implementation of the Read for Real intervention.

[c]The 0 to 39 percent and 40 to 79 percent categories were combined to protect teacher confidentiality.

[d]The 40 to 79 percent and 80 to 100 percent categories were combined to protect teacher confidentiality.

[e]The number of teachers presented in this row is the number participating in the study. Roughly half the teachers were observed on "Learn" days, and roughly half were observed on "Practice" days.

TABLE II.12

FIDELITY OF IMPLEMENTATION FOR THE READ FOR REAL CURRICULUM IN YEARS 1 AND 2

	Learn Observation Days				Practice Observation Days			
	Year 1	Year 2	Difference	*p-value*	Year 1	Year 2	Difference	*p-value*
Percentage of Teachers Who Reported Using Read for Real	86.79	83.33	-3.46	.703	86.79	83.33	-3.46	.703
Before Reading								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the explanation of the Before Reading focus strategy	55.00	68.75	13.75	.415	54.55	35.71	-18.84	.247
Discusses the strategy with students	45.00	75.00	30.00	.073	54.55	35.71	-18.84	.247
Reads or asks a student to read the information in the My Thinking box	55.00	50.00	-5.00	.773	n.a.	n.a.	n.a.	n.a.
Asks students to apply the strategy	45.00	43.75	-1.25	.942	57.58	42.86	-14.72	.366
Discusses students' comments	n.a.	n.a.	n.a.	n.a.	48.48	42.86	-5.62	.731
During Reading								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the explanation of the During Reading focus strategy	60.00	81.25	21.25	.179	48.48	50.00	1.52	.926
Discusses the strategy with the students	65.00	62.50	-2.50	.881	n.a.	n.a.	n.a.	n.a.
Reads or asks a student to read the information in the My Thinking box (notes from the reading partner)	60.00	68.75	8.75	.600	42.42	42.86	0.44	.979
Asks students to share their thinking about the strategy	60.00	50.00	-10.00	.562	n.a.	n.a.	n.a.	n.a.
Reminds students to write notes about the strategy	n.a.	n.a.	n.a.	n.a.	36.36	57.14	20.78	.196
Stops and addresses the My Thinking notes at the "red strategy buttons"	65.00	62.50	-2.50	.881	69.70	50.00	-19.70	.207
Reads and/or asks students to read the selection	70.00	56.25	-13.75	.408	69.70	64.29	-5.41	.723
After Reading								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the After Reading focus strategy	35.00	31.25	-3.75	.819	24.24	28.57	4.33	.762
Discusses or asks questions about the strategy	25.00	31.25	6.25	.688	21.21	21.43	0.22	.987
Reads or asks a student to read the information in the My Thinking box	20.00	31.25	11.25	.453	n.a.	n.a.	n.a.	n.a.
Gives a written assignment highlighting the After Reading focus strategy	n.a.	n.a.	n.a.	n.a.	15.15	21.43	6.28	.610
Calls on students to implement the After Reading focus strategy	15.00	31.25	16.25	.256	n.a.	n.a.	n.a.	n.a.
Comprehension								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Administers the open book comprehension test	—[b]	—[b]	—[b]	—[b]	9.09	7.14	-1.95	.831
Corrects tests with the class	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]
Discusses responses	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]
Organizing Information								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the information from the reading partner	20.00	18.75	-1.25	.928	n.a.	n.a.	n.a.	n.a.
Discusses the graphic organizer	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Asks students to complete the graphic organizer	n.a.	n.a.	n.a.	n.a.	12.12	28.57	16.45	.177

Table II.12 *(continued)*

	Learn Observation Days				Practice Observation Days			
	Year 1	Year 2	Difference	*p-value*	Year 1	Year 2	Difference	*p-value*
Writing for Comprehension								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the information from the reading partner	__[b]	__[b]	__[b]	__[b]	n.a.	n.a.	n.a.	n.a.
Reads or asks a student to read the summary	__[b]	__[b]	__[b]	__[b]	n.a.	n.a.	n.a.	n.a.
Asks students to write a summary based on their completed graphic organizer	n.a.	n.a.	n.a.	n.a.	__[b]	__[b]	__[b]	__[b]
Identifies how the paragraphs and sentences in the summary correspond to the information on the graphic organizer	__[b]	__[b]	__[b]	__[b]	n.a.	n.a.	n.a.	n.a.
Discusses the three parts of a summary								
Introduction	__[b]	__[b]	__[b]	__[b]	n.a.	n.a.	n.a.	n.a.
Body	__[b]	__[b]	__[b]	__[b]	n.a.	n.a.	n.a.	n.a.
Conclusion	__[b]	__[b]	__[b]	__[b]	n.a.	n.a.	n.a.	n.a.
Overall Fidelity								
Percentage of Teachers Who Were Observed Implementing:[c]								
40 to 100 percent of the fidelity form behaviors listed above[d]	50.00	37.50	–12.50	.468	48.48	42.86	-5.62	.731
0 to 39 percent of the fidelity form behaviors listed above	50.00	62.50	12.50	.468	51.52	57.14	5.62	.731
Mean Percentage of the Fidelity Form Behaviors Listed Above that Teachers Were Observed Implementing	34.60	34.25	-0.35	.962	34.05	31.51	-2.54	.748
Sample Size[e]	**57 (Year 1), 31 (Year 2)**							

SOURCE: Classroom observations.

NOTE: The reported differences are regression adjusted to account for school effects. The Year 1 mean is the raw mean, the Year 2 mean is the Year 1 mean plus the regression-adjusted difference between years. Note that this explains the difference from the raw Year 2 means reported in Table II.11.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the behaviors listed in this table. In addition, the calculations presented in this table are based on all behaviors on the fidelity form, not only on the behaviors in the observed windows.

[b]Value suppressed to protect teacher confidentiality.

[c]The vocabulary and fluency items have been left out of the table because developers noted they were not essential for implementation of the Read for Real intervention.

[d]The 40 to 79 percent and the 80 to 100 percent categories have been combined to protect teacher confidentiality.

[e]The number of teachers presented in this row is the number participating in the study. Roughly half the teachers were observed on "Learn" days, and roughly half were observed on "Practice" days.

n.a. = not applicable.

TABLE II.13

FIDELITY OF IMPLEMENTATION FOR THE READ FOR REAL CURRICULUM, BY TEACHER EXPERIENCE WITH THE CURRICULUM

	Learn Observation Days				Practice Observation Days			
	Teachers in Year 2 Only	Teachers in Both Years	Difference	*p-value*	Teachers in Year 2 Only	Teachers in Both Years	Difference	*p-value*
Percentage of Teachers Who Reported Using Read for Real	80.00	90.00	10.00	.505	80.00	90.00	10.00	.505
Before Reading								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the explanation of the Before Reading focus strategy	77.78	57.14	-20.64	.411	—[b]	—[b]	—[b]	—[b]
Discusses the strategy with students	77.78	71.43	-6.35	.789	—[b]	—[b]	—[b]	—[b]
Reads or asks a student to read the information in the My Thinking box	66.67	28.57	-38.10	.149	n.a.	n.a.	n.a.	n.a.
Asks students to apply the strategy	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]
Discusses students' comments	n.a.	n.a.	n.a.	n.a.	—[b]	—[b]	—[b]	—[b]
During Reading								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the explanation of the During Reading focus strategy	77.78	85.71	7.93	.710	—[b]	—[b]	—[b]	—[b]
Discusses the strategy with the students	66.67	57.14	-9.53	.719	n.a.	n.a.	n.a.	n.a.
Reads or asks a student to read the information in the My Thinking box (notes from the reading partner)	77.78	57.14	-20.64	.411	—[b]	—[b]	—[b]	—[b]
Asks students to share their thinking about the strategy	55.56	42.86	-12.70	.642	n.a.	n.a.	n.a.	n.a.
Reminds students to write notes about the strategy	n.a.	n.a.	n.a.	n.a.	—[b]	—[b]	—[b]	—[b]
Stops and addresses the My Thinking notes at the "red strategy buttons"	66.67	57.14	-9.53	.719	—[b]	—[b]	—[b]	—[b]
Reads and/or asks students to read the selection	66.67	42.86	-23.81	.171	—[b]	—[b]	—[b]	—[b]
After Reading								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the After Reading focus strategy	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]
Discusses or asks questions about the strategy	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]
Reads or asks a student to read the information in the My Thinking box	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Gives a written assignment highlighting the After Reading focus strategy	n.a.	n.a.	n.a.	n.a.	—[b]	—[b]	—[b]	—[b]
Calls on students to implement the After Reading focus strategy	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.

Table II.13 *(continued)*

	Learn Observation Days				Practice Observation Days			
	Teachers in Year 2 Only	Teachers in Both Years	Difference	*p-value*	Teachers in Year 2 Only	Teachers in Both Years	Difference	*p-value*
Comprehension								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Administers the open book comprehension test	0.00	28.57	28.57	.098	—[b]	—[b]	—[b]	—[b]
Corrects tests with the class	0.00	28.57	28.57	.098	0.00	0.00	n.a.	n.a.
Discusses responses	0.00	28.57	28.57	.098	0.00	0.00	n.a.	n.a.
Organizing Information								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the information from the reading partner	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Discusses the graphic organizer	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Asks students to complete the graphic organizer	n.a.	n.a.	n.a.	n.a.	—[b]	—[b]	—[b]	—[b]
Writing for Comprehension								
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Reads or asks a student to read the information from the reading partner	0.00	14.29	14.29	.271	n.a.	n.a.	n.a.	n.a.
Reads or asks a student to read the summary	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Asks students to write a summary based on their completed graphic organizer	n.a.	n.a.	n.a.	n.a.	—[b]	—[b]	—[b]	—[b]
Identifies how the paragraphs and sentences in the summary correspond to the information on the graphic organizer	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Discusses the three parts of a summary								
Introduction	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Body	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Conclusion	—[b]	—[b]	—[b]	—[b]	n.a.	n.a.	n.a.	n.a.
Overall Fidelity								
Percentage of Teachers Who Were Observed Implementing:[c]								
80 to 100 percent of the fidelity form behaviors listed above	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]
40 to 79 percent of the fidelity form behaviors listed above	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]	—[b]
0 to 39 percent of the fidelity form behaviors listed above	55.56	71.43	15.87	.547	63.64	33.33	-30.31	.386
Mean Percentage of the Fidelity Form Behaviors Listed Above That Teachers Were Observed Implementing	35.11	33.14	-1.97	.868	29.41	39.22	9.81	.570
Sample Size	**9 (Year 2 Only), 22 (Both Years)**							

SOURCE: Classroom observations.

Table II.13 *(continued)*

NOTE: These differences are not regression adjusted. *P-values* could not be obtained when few teachers were observed adhering to a specific behavior. This is indicated by a (.).

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the behaviors listed in this table. In addition, the calculations presented in this table are based on all behaviors on the fidelity form, not only on the behaviors in the observed windows.

[b]Value suppressed to protect teacher confidentiality.

[c]The vocabulary and fluency items have been left out of the table because developers noted they were not essential for implementation of the Read for Real intervention.

n.a. = not applicable.

Focusing on all fidelity form items, there were no statistically significant differences in implementation fidelity for the "Learn" day items between Years 1 and 2 (Table II.12). There were also no statistically significant differences in implementation fidelity for the "Learn" day items between teachers who participated in both years of the study and teachers who were new to the study in the second year (Table II.13).

Practice Days. On the Read for Real "Practice" days in Year 2, restricting to items in sections for which teachers were observed, teachers were observed engaging in 76 percent of the practices deemed important by developers for implementing the intervention (Table II.11). The highest rates of implementation in the second year were observed for teachers reading or asking students to read a selection, for teachers reminding students to write notes about the strategy, and for teachers reading or asking students to read the explanation of the During Reading strategy (100 percent, 89 percent, and 88 percent, respectively, among teachers for whom the "During Reading" section was observed) (Table II.10).

Focusing on all fidelity form items, there were no statistically significant differences in implementation fidelity for the "Practice" day items between Years 1 and 2 (Table II.12). There were also no statistically significant differences in implementation fidelity for the "Practice" day items between teachers who participated in both years of the study and teachers who were new to the study in the second year (Table II.13).

3. ReadAbout

As mentioned above, ReadAbout's developer (Scholastic) did not prescribe that teachers implement all nine teaching practices on the ReadAbout fidelity form during every lesson. In particular, teachers were told that they could conduct small group instruction in comprehension, vocabulary, or writing, but not necessarily all three in a single lesson. Therefore, the text below focuses primarily on the percentage of items observed restricting to items in sections for which teachers were observed (column 2 in Table II.14 and column 1 in Table II.15).

Restricting to items in sections for which teachers were observed in Year 2, ReadAbout teachers were observed engaging in 94 percent of the teaching practices considered important to the implementation of ReadAbout (Table II.15). The highest rates of implementation in the second year were observed for teachers providing direct instruction on comprehension or vocabulary skills (86 and 91 percent, respectively) and providing students with opportunities to apply comprehension or vocabulary skills (94 and 91 percent, respectively) (Table II.14).

Focusing on all fidelity form items, there were no statistically significant differences in implementation fidelity between the first and second years of the study (Table II.16). Teachers who participated in both years of the study were statistically significantly more likely to report using ReadAbout than teachers who were new to the study in the second year (Table II.17, 100 percent vs. 82 percent, p-value = .013). Teachers who participated in both years of the study were also statistically significantly more likely than teachers who were new to the study in the second year to be observed using the ReadAbout materials (100 percent vs. 82 percent, p-value = .013).

TABLE II.14

FIDELITY OF IMPLEMENTATION OF INDIVIDUAL TEACHING PRACTICES FOR THE READABOUT CURRICULUM IN YEAR 2

	(1) Percentage of Teachers Observed Implementing Section[a]	(2) Among Teachers Implementing Section, Percentage of Teachers Observed Implementing Behavior	(3) Among All Teachers, Percentage of Teachers Observed Implementing Behavior[a]
Part I, Section I. Comprehension			
Provided direction instruction (explain and/or model) on the strategy or skill	77.78	85.71	66.67
Provided opportunities for students to apply the skill (guided practice)		94.29	73.33
Part I, Section II. Vocabulary			
Provided direction instruction (explain and/or model) on the strategy or skill	24.44	90.91	22.22
Provided opportunities for students to apply the skill (guided practice)		90.91	22.22
Part I, Section III. Writing			
Provided students instruction on the selected 6+1 Writing Trait	11.11	40.00	—[b]
Provided opportunities to apply the 6+1 Writing Trait Model		40.00	—[b]
Part II. Use of Workstations and ReadAbout Materials			
Used the ReadAbout materials	n.a.	n.a.	95.56
Computer workstation used		n.a.	68.89
Independent workstation used		n.a.	55.56
Sample Size[c]		**46**	

SOURCE: Classroom observations.

NOTE: There are two possible explanations for why teachers were not observed implementing a section of behaviors from the fidelity form: (1) a section might not have been appropriate given the stage of the lesson; or (2) the behaviors might have been appropriate, but the teacher failed to exhibit them. Because it is not possible to determine which explanation is the reason the behaviors in that section were not observed, we report fidelity rates under both assumptions. The findings in the second column can be viewed as an "upper bound" on fidelity (corresponding to the first explanation), while the findings in the third column can be viewed as a "lower bound" (corresponding to the second explanation).

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]Value suppressed to protect teacher confidentiality.

[c]The number of teachers presented in this row is the number participating in the study.

n.a. = not applicable.

TABLE II.15

OVERALL FIDELITY OF IMPLEMENTATION FOR THE READABOUT CURRICULUM IN YEAR 2

Restricted to Behaviors in Observed Sections	
Percentage of Teachers Who Were Observed Implementing:	
80 to 100 percent of the ReadAbout fidelity form behaviors	86.05
0 to 79 percent of the ReadAbout fidelity form behaviors[a]	13.95
Mean Percentage of the ReadAbout Fidelity Form Behaviors That Teachers Were Observed Implementing	94.01
All Behaviors[b]	
Percentage of Teachers Who Were Observed Implementing:	
40 to 100 percent of the ReadAbout fidelity form behaviors[c]	73.33
0 to 39 percent of the ReadAbout fidelity form behaviors	26.67
Mean Percentage of the ReadAbout Fidelity Form Behaviors That Teachers Were Observed Implementing	45.93
Sample Size[c]	**46**

SOURCE: Classroom observations.

NOTE: There are two possible explanations for why teachers were not observed implementing a section of behaviors from the fidelity form: (1) a section might not have been appropriate given the stage of the lesson; or (2) the behaviors might have been appropriate, but the teacher failed to exhibit them. Because it is not possible to determine which explanation is the reason the behaviors in that section were not observed, we report overall fidelity rates under both assumptions. The findings in the first column can be viewed as an "upper bound" on overall fidelity (corresponding to the first explanation), while the findings in the second column can be viewed as a "lower bound" (corresponding to the second explanation).

[a]The 0 to 39 percent and 40 to 79 percent categories were combined to protect teacher confidentiality.

[b]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[c]The 40 to 79 percent and 80 to 100 percent categories were combined to protect teacher confidentiality.

[d]The number of teachers presented in this row is the number participating in the study.

TABLE II.16

FIDELITY OF IMPLEMENTATION FOR THE READABOUT CURRICULUM IN YEARS 1 AND 2

	Year 1	Year 2	Difference	*p-value*
Percentage of Teachers Who Reported Using ReadAbout	100.00	95.71	-4.29	.189
Part I, Section I. Comprehension				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provided direction instruction (explain and/or model) on the strategy or skill	69.57	68.92	-0.65	.946
Provided opportunities for students to apply the skill (guided practice)	69.57	75.23	5.66	.556
Part I, Section II. Vocabulary				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provided direction instruction (explain and/or model) on the strategy or skill	15.22	24.62	9.40	.283
Provided opportunities for students to apply skill (guided practice)	19.57	20.88	1.31	.875
Part I, Section III. Writing				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provided students instruction on the selected 6+1 Writing Trait	—[b]	—[b]	—[b]	—[b]
Provided opportunities to apply the 6+1 Writing Trait Model	—[b]	—[b]	—[b]	—[b]
Part II. Use of Workstations and ReadAbout Materials				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Used the ReadAbout materials	91.30	95.10	3.80	.493
Computer workstation used	89.13	68.56	-20.57	.010
Independent workstation used	58.70	57.74	-0.96	.919
Overall Fidelity				
Percentage of Teachers Who Were Observed Implementing:[a]				
40 to 100 percent of the fidelity form behaviors listed above[c]	76.09	73.33	–2.76	.863
0 to 39 percent of the fidelity form behaviors listed above	23.91	26.24	2.33	.801
Mean Percentage of the Fidelity Form Behaviors Listed Above That Teachers Were Observed Implementing	45.89	46.38	0.49	.882
Sample Size[d]	**53**	**46**		

SOURCE: Classroom observations.

NOTE: The reported differences are regression adjusted to account for school effects. The Year 1 mean is the raw mean, the Year 2 mean is the Year 1 mean plus the regression-adjusted difference between years. Note that this explains the difference from the raw Year 2 means reported in Table II.15.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the behaviors listed in this table. In addition, the calculations presented in this table are based on all behaviors in all sections of the fidelity form, not only on the behaviors in the observed sections.

[b]Value suppressed to protect teacher confidentiality.

[c]The 40 to 79 percent and the 80 to 100 percent categories have been combined to protect teacher confidentiality.

[d]The number of teachers presented in this row is the number participating in the study.

TABLE II.17

FIDELITY OF IMPLEMENTATION FOR THE READABOUT CURRICULUM, BY TEACHER EXPERIENCE WITH THE CURRICULUM

	Teachers in Year 2 Only	Teachers in Both Years	Difference	*p-value*
Percentage of Teachers Who Reported Using ReadAbout	81.82	100.00	18.18	.013*
Part I, Section I. Comprehension				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provided direction instruction (explain and/or model) on the strategy or skill	63.64	68.75	5.11	.762
Provided opportunities for students to apply the skill (guided practice)	72.73	75.00	2.27	.885
Part I, Section II. Vocabulary				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provided direction instruction (explain and/or model) on the strategy or skill	27.27	21.88	-5.39	.723
Provided opportunities for students to apply skill (guided practice)	27.27	21.88	-5.39	.723
Part I, Section III. Writing				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provided students instruction on the selected 6+1 Writing Trait	__[b]	__[b]	__[b]	__[b]
Provided opportunities to apply the 6+1 Writing Trait Model	__[b]	__[b]	__[b]	__[b]
Part II. Use of Workstations and ReadAbout Materials				
Percentage of Teachers Who Were Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Used the ReadAbout materials	81.82	100.0	18.18	.013*
Computer workstation used	63.64	71.88	8.24	.618
Independent workstation used	45.45	59.38	13.93	.435
Overall Fidelity				
Percentage of Teachers Who Were Observed Implementing:[a]				
80 to 100 percent of the fidelity form behaviors listed above	__[b]	__[b]	__[b]	__[b]
40 to 79 percent of the fidelity form behaviors listed above	63.64	75.00	11.36	.480
0 to 39 percent of the fidelity form behaviors listed above	27.27	25.00	-2.27	.885
Mean Percentage of the Fidelity Form Behaviors Listed Above That Teachers Were Observed Implementing	44.45	47.22	2.77	.643
Sample Size	**12**	**34**		

SOURCE: Classroom observations.

NOTE: These differences are not regression adjusted.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers (observed and not observed for fidelity) are included in these calculations. We assumed that teachers who were not implementing the curricula did not engage in the behaviors listed in this table. In addition, the calculations presented in this table are based on all behaviors in all sections of the fidelity form, not only on the behaviors in the observed sections.

[b]Value suppressed to protect teacher confidentiality.

*Statistically different at the .05 level.

4. Reading for Knowledge

As mentioned in Chapter I, Reading for Knowledge was not implemented in the second year of the study. We summarize Reading for Knowledge fidelity information from the first study year below to help readers interpret the follow-up impacts presented in Chapter IV. Like Read for Real, the Reading for Knowledge intervention involved two types of instructional days, both observed for the study. Fidelity on days 1 and 3, which involved teacher-directed instruction, was assessed based on 9 items. Fidelity on days 2 and 4, which involved students working in cooperative groups, was based on 13 items.[36]

On teacher-directed instruction days, Reading for Knowledge teachers were observed implementing 58 percent of the teaching practices deemed important by developers for Reading for Knowledge implementation (not shown in table). On the teacher-directed instruction days, Reading for Knowledge teachers had the highest rates of implementation (67 to 71 percent) on activities related to building background knowledge about the topic of the text or about a skill or strategy and explaining or reviewing the skill/strategy. Fifty-two to 57 percent of teachers were observed presenting the reading goal, awarding cooperation/improvement points, and following the recommended pacing.

On days 2 and 4, when students were working in cooperative groups, Reading for Knowledge teachers were observed implementing, on average, 65 percent of the teaching practices that developers considered important to the implementation of the intervention (not shown in table). On days 2 and 4, Reading for Knowledge teachers had the highest rates of implementation on activities related to presenting the reading goal, discussing key points about the day's skill/strategy, providing feedback and prompts to student pairs during partner reading, circulating in the classroom and monitoring team discussions, and asking team members to share with the class (76 to 88 percent).

D. READING COMPREHENSION INSTRUCTIONAL PRACTICES

In this section, we examine data from the ERC observation form, which (as described in Chapter I) was designed to gather information on the number of times treatment and control group fifth-grade teachers engaged in a set of general (non-intervention-specific) teaching practices related to reading comprehension and vocabulary instruction.[37] This is in contrast to the fidelity observation protocols discussed in the previous section, which focused on teaching practices and procedures *specific to each intervention*.

---

[36]On days 1 and 3, teachers were observed to assess whether they built background knowledge, explained a strategy, read text aloud, and helped students think of or apply a strategy. On days 2 and 4, teachers were observed to assess whether they used whole group and partner activities, provided feedback and prompts to partner pairs, charted student progress, reviewed routines, read questions aloud, circulated around the classroom, and asked teams to share with the class.

[37]These practices were selected because most of them were components of effective reading comprehension instructional interventions studied in small-scale experimental research (see Carlisle and Rice 2002; Pearson and Dole 1987).

We begin this section by presenting a description of the process used to construct teacher practice scales from the observational data. We then present findings from our examination of the differences in teaching practices between fifth-grade teachers in the intervention groups and the control group. This section also presents findings from comparisons of the instructional practices of teachers participating in the study for two years and teachers new to the study in Year 2. Finally, we present findings from comparisons of instructional practices in Years 1 and 2 for teachers participating in both years of the study.

Constructing Teacher Practice Scales. The ERC observational protocol allowed the study team to collect data from all fifth-grade classrooms (both intervention and control) using a common measure. Thus, it is possible to describe and compare teachers' instructional practices across treatment and control groups and across cohorts. With the ERC observational protocol, observers recorded the number of times treatment and control group teachers engaged in specific teaching practices. A classroom observation consisted of up to ten 10-minute intervals. For each interval, observers looked for 28 instructional practices. There were also 14 items that were completed at the end of each observation, which addressed issues such as student engagement during the lesson and teachers' management of student behavior (see Appendix Table I.1 for a list of practices on the ERC). Classroom observations were conducted for all the class periods in a given day for which the teacher indicated she would be using informational text.[38] To condense this observation data into a manageable number of variables for analysis, we developed scales based on these practices using the following three steps:

1. ***Coding tallies for each item into ordinal categories***. To support subsequent psychometric analyses—particularly the implementation of item response theory (IRT) scaling discussed in step 3 below—ordinal categories were created for the distributions of both sums and averages of tallies (or number of times teachers engaged in a specific teaching practice) across the 10-minute intervals for each item. These categories were based on an investigation of the distribution of the averages of tallies across intervals for each item. These ordered categories represented the extent to which each teacher practice was observed, where higher categories represented teachers engaging in the particular practice more frequently. For example, if the average number of tallies across intervals for a particular item ranged from 0 to 10 for all teachers, the average tally for a particular teacher might have been assigned to one of three categories (0-3, 4-6, and 7-10) depending on the average number of times across intervals the teacher was observed engaging in that practice.[39]

[38]Although classrooms, on average, were observed multiple times during the day, they were only observed for a single day, which may reduce the reliability of the teacher practice scales based on the ERC data (relative to observations conducted over multiple days). The teacher practice scales based on a single day of observations still allow us to calculate valid estimates of treatment/control differences on the scales (which are presented in this chapter).

[39]The ordered categories were then assigned numerical values. For each item, a value of zero was assigned to the lowest category. Values for subsequent categories were assigned by increasing the number of the previous category by one until the highest category was reached. In the example provided in the text, teachers in the 0-3 category were assigned a value of 0, teachers in the 4-6 category were assigned a value of 1, and teachers in the 7-10 category were assigned a value of 2.

2. ***Conducting an exploratory factor analysis.*** Exploratory factor analysis (EFA) was conducted to identify the underlying variables that best explain the ERC items.[40] This analysis enabled us to develop conceptual groupings of items that appeared related to the same underlying concept or theme. Items that contributed little to the coherence of these groupings were discarded.[41]

3. ***Estimating an item response theory model using the categorical variables formed in step 1.*** IRT scaling was performed to obtain an estimated score for each teacher. We followed this modeling approach because: (1) it allowed us to properly model the cross-loadings of items as indicated by the EFA (six items cross-loaded on two of the underlying variables explaining the ERC items that were found in step 2); (2) it maximized the amount of data we were able to use to construct the scales; and (3) it enabled us to account for the fact that some of our items have shared question stems. The IRT scaling also permitted a rigorous assessment of the psychometric properties of the items of the ERC form, as well as the unbiased estimation of scores and level of reliability for each teacher's score and the overall distribution of scores. See Appendix F for a detailed description of the IRT model used to develop teacher practice scales.

This process resulted in three scales that were used in the study's analyses.[42] The ERC items were distributed across these scales, and, as noted above, six items contribute to more than one scale. The results from the factor analysis show that items contribute to the scales with different weights, depending on the degree to which the items are related to the underlying concepts measured by the scales. (See Table II.18 for a listing of the ERC items contained in each scale.) Therefore, the study team assigned names to these scales based on the items they include and the weight that specific items take on in each scale based on the results from the factor analysis. The resulting three scales and the distinct and overlapping items included in each scale are the following:

- ***Traditional Interaction.*** This scale, which captures interactive teaching practices that have been in use for many decades in American schools (Durkin 1978-1979; Brophy and Evertson 1976), is based on 13 teaching practices (6 related to vocabulary and 7 to comprehension instruction). Unique items on this scale include practices related to teachers (a) asking questions based on material in text beyond a literal level; (b) elaborating concepts during and after reading; (c) providing definitions or explanations; (d) providing examples of multiple meanings; (e) using visuals and

[40]Factor extraction was conducted using unweighted least squares estimation; oblique rotation was used because it was expected that the underlying variables would be correlated (our analysis ultimately confirmed this expectation).

[41]The EFA methods just described were used for items on Part I of the ERC. For Part II ERC items, EFA was not necessary because there were clear groupings of items that shared similar content themes.

[42]Scale scores ranged from 408 to 551 for classrooms observed in the second year of the study. Scale scores ranged from 405 to 562 for classrooms observed in the first year of the study. See Table F.3 in Appendix F for the range of each scale.

TABLE II.18

EXPOSITORY READING COMPREHENSION ITEMS CONTAINED IN STUDY SCALES

Item	Scales		
	Traditional Interaction	Reading Strategy Guidance	Classroom Management and Student Engagement
Comprehension Items			
Teacher Explains Text Structure		√	
Students Practice Use of Text Structure		√	
Teacher Models Comprehension Strategies		√	
Teacher Explains Comprehension Strategies		√	
Students Practice Comprehension Strategies		√	
Teacher Explains How to Generate Questions	√	√	
Students Practice Generating Questions	√	√	
Teacher Explains Text Features	√	√	
Students Practice Using Text Features	√	√	
Teacher Asks Students to Justify Responses	√	√	
Teacher Asks Questions Based on Material in Text Beyond a Literal Level	√		
Teacher Elaborates Concepts During and After Reading	√		
Vocabulary Items			
Teacher Provides Definition or Explanation	√		
Teacher Provides Examples / Multiple Meanings	√		
Teacher Uses Visuals / Pictures	√		
Teacher Teaches Word-Learning Strategies	√	√	
Students Asked to Do Something Requiring Word Knowledge	√		
Student Given Chance to Apply Word-Learning Strategies	√		
Other Items			
Teacher Maximized Instruction Time			√
Teacher Managed Student Behavior			√
Student Engagement – First Half of Observation			√
Student Engagement – Second Half of Observation			√

pictures; (f) asking students to work on tasks requiring word knowledge; and (g) giving students the opportunity to apply word learning strategies.

- ***Reading Strategy Guidance.*** This scale reflects more heavily the practices entailed in research on explicit comprehension strategies (see Pearson and Dole 1987; Carlisle and Rice 2002). The scale includes 11 items. Unique items on this scale include practices related to teachers explaining and modeling (and students practicing) comprehension strategies and text structure (for example, cause-effect or compare-contrast) to improve comprehension.

- ***Classroom Management and Student Engagement.*** This scale includes one item related to how teachers manage student behavior, one item related to maximizing instructional time, and two items related to students' engagement during class.[43]

- ***Overlapping Items.*** Six items are contained in both the Traditional Interaction scale and the Reading Strategy Guidance scale because the results from the EFA (conducted to identify groupings of items related to the same underlying concept) showed that the items loaded on both scales. These items include practices related to teachers (1) explaining (and having students practice) the use of question generation and text features (for example, captions or subheadings) to improve comprehension, (2) asking students to justify their responses, and (3) teaching word-learning strategies.

We assessed the reliability of each of the three scales for classrooms observed in the second study year. The reliability of the Traditional Interaction scale was 0.65, the reliability of the Reading Strategy Guidance scale was 0.63, and the reliability of the Classroom Management scale was 0.86.[44,45]

Differences in Instructional Practices in the Second Year of the Study Between Experimental and Control Group Teachers. For two of the three scales (Classroom Management and Reading Strategy Guidance), there were no statistically significant differences in Year 2 between the treatment and control classrooms (Table II.19). However, a statistically

[43]The items in this scale were part of the set of items that were completed once at the end of each observation.

[44]For classrooms observed in Year 1, the reliability of the Traditional Interaction scale was 0.70, the reliability of the Reading Strategy Guidance scale was 0.72, and the reliability of the Classroom Management scale was 0.83. See Appendix F for additional information on the reliability, inter-rater reliability, and validity of the observation scales. Appendix F also provides figures showing how the scale score values can be interpreted and linked back to the items contained in the scales.

[45]Reliability is positively related to statistical precision. Estimates of differences between two groups based on measures with lower reliability are less precise.

TABLE II.19

DIFFERENCES IN CLASSROOM PRACTICES BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING FIFTH-GRADE TEACHERS IN YEARS 1 AND 2

	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Traditional Interaction Scale					
Year 1 (Spring 2007)					
Impact	502.36	-4.02*	-3.63	-1.77	-3.02
Effect Size		-0.62	-0.56	-0.27	-0.47
p-value		0.04	0.30	0.90	0.09
Year 2 (Spring 2008)					
Impact	500.96	-3.50*	-3.63	1.49	-2.12
Effect Size		-0.54	-0.56	0.23	-0.33
p-value		0.02	0.21	0.94	0.25
Difference Between Years 1 and 2					
Difference in Impact		0.53	0.00	3.26	0.90
Difference in Effect Size		0.08	0.00	0.51	0.14
p-value for the Difference		1.00	1.00	0.44	0.87
Reading Strategy Guidance Scale					
Year 1 (Spring 2007)					
Impact	499.09	1.29	2.31	0.22	1.21
Effect Size		0.19	0.34	0.03	0.18
p-value		1.00	0.94	1.00	0.91
Year 2 (Spring 2008)					
Impact	500.2	1.50	1.09	0.22	0.90
Effect Size		0.22	0.16	0.03	0.13
p-value		0.95	1.00	1.00	0.92
Difference Between Years 1 and 2					
Difference in Impact		0.21	-1.22	-0.01	-0.31
Difference in Effect Size		0.03	-0.18	0.00	-0.05
p-value for the Difference		1.00	1.00	1.00	0.99
Classroom Management Scale					
Year 1 (Spring 2007)					
Impact	503.94	-1.93	-10.68	0.07	-3.96
Effect Size		-0.06	-0.35	0.00	-0.13
p-value		1.00	0.66	1.00	0.92
Year 2 (Spring 2008)					
Impact	505.55	-9.38	-13.92	0.70	-8.13
Effect Size		-0.30	-0.45	0.02	-0.26
p-value		0.87	0.43	1.00	0.56
Difference Between Years 1 and 2					
Difference in Impact		-7.45	-3.24	0.63	-4.17
Difference in Effect Size		-0.24	-0.11	0.02	-0.14
p-value for the Difference		0.93	1.00	1.00	0.89
Number of Teachers in Year 1[a]	**59**	**52**	**50**	**54**	**156**
Number of Teachers in Year 2[b]	**54**	**53**	**46**	**31**	**130**

Table II.19 *(continued)*

SOURCE: Classroom observations.

NOTE: The scales presented in this table were constructed to capture the frequency of the behaviors in each instructional practice domain shown above. For each scale, the numbers reported in the column labeled "Control Group" are the average predicted value of the scale for all teachers as if they were in the control group. The numbers reported in the remaining columns are, by row: (1) the difference in means between treatment and control group, (2) the effect size, and (3) the *p-value* of the difference. The *p-values* presented in this table are adjusted for multiple-hypotheses testing. For each scale, the differences between cohort impacts are also reported. Regression-adjusted differences were calculated taking into account the clustering of teachers within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher gender, teacher age, teacher race, and district indicators. Smaller scale values represent lower levels of behaviors in the instructional practice domain, while larger values represent higher values of the behaviors. See Appendix F for more information on interpreting the scale score values.

[a]The number of teachers presented in this row is the number of fifth-grade teachers who participated in the study's first year. Some teachers taught more than one class. The calculations presented in the table are based on the number of classroom observations for which scale scores were calculated. The response rates for these calculations vary from 91 percent for CRISS classrooms to 100 percent for Read for Real classrooms.

[b]The number of teachers presented in this row is the number of fifth-grade teachers participating in the study's second year. Some teachers taught more than one class. The calculations presented in the table are based on the number of classroom observations for which scale scores were calculated. The response rates for these calculations vary from 90 percent for CRISS classrooms to 100 percent for ReadAbout classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

significant difference was found on the Traditional Interaction scale in the second year,[46] with Project CRISS teachers having lower scores on the scale than control teachers (effect size: -0.54).[47]

To assess the robustness of the statistically significant difference observed for Project CRISS on the Traditional Interaction scale scores in the study's second year, we conducted a sensitivity analysis in which scales were constructed using *sums* of tallies across intervals (instead of using *averages* of tallies across intervals as was done above). Since teachers were observed during the class periods in which they indicated they would be using informational text on the day the observations were conducted, all teachers were not observed for the same number of intervals. Thus, the sum of tallies across intervals is a substantively different indicator from the average. The analysis of Year 2 teacher practice scales based on sums was conducted in all other respects using the same method as the analysis based on averages. Differences on these scales based on sums of tallies between Year 2 treatment and control teachers were not statistically significant (see Appendix Table H.7).

As an additional sensitivity analysis, we considered a different set of teacher instructional practices scales. These scales were constructed by grouping all items pertaining to teaching comprehension to create a Teaching Comprehension scale, and all items regarding teaching vocabulary to create a Teaching Vocabulary scale. These scales were also created in two ways: using sums and using averages of tallies from the classroom observations.[48] We did not find any statistically significant differences on these scales between treatment and control teachers in the second year of the study (see Appendix Table H.8). Taken together, these sensitivity tests suggest that the statistically significant impact on Traditional Interaction scale scores is sensitive to the way in which the scale is constructed.

To further examine the Year 2 differences between Project CRISS and control teachers on the Traditional Interaction scale, we examined treatment/control differences on the ERC items on which each scale is based, both for each treatment group separately and for the combined treatment group (Tables II.20, II.21, and II.22). To ensure that the *p*-values from these analyses are comparable to the *p*-values reported in Table II.19 (where multiple comparisons adjustments

---

[46]To help interpret the treatment-control difference observed on the Traditional Interaction scale, it is useful to link the difference in scale scores to the corresponding differences in the *frequency categories* used to characterize teachers' engagement in the individual behaviors underlying each scale. Figures F.1.A and F.1.B in Appendix F relate this difference based on the scales to the underlying frequencies of the specific behaviors making up the scale. For both the treatment and control groups, the mean scale scores resulted from behaviors whose mean frequency fell within the lowest category for each of the items underlying the scale. The appendix figures show that teachers in both groups, on average, were engaging in these behaviors fewer than once during each 10-minute interval they were observed, which means that the difference between the treatment and control groups amounted to less than one time during the typical 10-minute interval.

[47]A similar effect was found for teachers observed in the first year (see Table II.19): Project CRISS teachers had statistically significantly lower scores on the Traditional Interaction scale than teachers in the control group (effect size: -0.62). The difference in these effects between Years 1 and 2 was not statistically significant.

[48]The reliability of these scales was 0.60 and 0.64, for the Teaching Comprehension and Teaching Vocabulary scales based on averages, respectively.

TABLE II.20

DIFFERENCES IN CLASSROOM PRACTICES IN THE SECOND STUDY YEAR BETWEEN TREATMENT AND CONTROL GROUP TEACHERS FOR ITEMS CONTAINED IN THE TRADITIONAL INTERACTION SCALE

		Difference Between Each of the Following and the Control Group:			
	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Comprehension Items					
Teacher Explains How to Generate Questions (Item 4b)					
Difference	0.21	0.08	-0.05	0.11	0.04
Effect Size		0.25	-0.16	0.36	0.13
p-value		0.71	0.86	0.31	0.52
Students Practice Generating Questions (Item 4c)					
Difference	0.33	0.33	-0.06	0.19	0.17
Effect Size		0.94	-0.18	0.52	0.48
p-value		0.10	0.92	0.30	0.09
Teacher Explains Text Features (Item 5b)					
Difference	0.24	-0.15*	-0.14*	0.08	-0.09*
Effect Size		-0.42	-0.38	0.23	-0.25
p-value		0.02	0.04	0.55	0.04
Students Practice Using Text Features (Item 5c)					
Difference	0.31	-0.12	-0.22*	0.09	-0.12
Effect Size		-0.24	-0.45	0.19	-0.25
p-value		0.33	0.01	0.84	0.08
Teacher Asks Students to Justify Responses (Item 6c)					
Difference	0.33	-0.05	-0.13	-0.06	-0.09
Effect Size		-0.11	-0.32	-0.13	-0.20
p-value		0.94	0.28	0.91	0.19
Teacher Asks Questions Based on Material in Text Beyond a Literal Level (Item 7c)					
Difference	1.23	-0.61*	-0.47	0.06	-0.38*
Effect Size		-0.45	-0.35	0.05	-0.28
p-value		0.03	0.15	0.98	0.05
Teacher Elaborates Concepts During and After Reading (Item 8)					
Difference	1.54	-0.52	-0.50	-0.21	-0.40
Effect Size		-0.33	-0.32	-0.13	-0.25
p-value		0.20	0.30	0.71	0.10
Vocabulary Items					
Teacher Provides Definition or Explanation (Item 1)					
Difference	0.56	-0.21	0.08	0.19	0.00
Effect Size		-0.35	0.14	0.32	0.00
p-value		0.13	0.91	0.42	1.00
Teacher Provides Examples/Multiple Meanings (Item 2)					
Difference	0.82	-0.23	0.12	0.23	0.02
Effect Size		-0.23	0.12	0.23	0.02
p-value		0.20	0.80	0.37	0.87

Table II.20 *(continued)*

	Control Group Mean	Difference Between Each of the Following and the Control Group: Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Teacher Uses Visuals/Pictures (Item 3)					
Difference	0.27	-0.10	-0.16	0.05	-0.08
Effect Size		-0.16	-0.25	0.09	-0.12
p-value		0.33	0.24	0.89	0.31
Teacher Teaches Word-Learning Strategies (Item 4)					
Difference	0.14	-0.08	-0.08	-0.02	-0.07
Effect Size		-0.29	-0.28	-0.07	-0.25
p-value		0.21	0.33	0.96	0.15
Students Asked to Do Something Requiring Word Knowledge (Item 5)					
Difference	1.53	-0.44	0.06	0.48	0.01
Effect Size		-0.29	0.04	0.32	0.01
p-value		0.20	0.99	0.30	0.97
Student Given Chance to Apply Word-Learning Strategies (Item 6)					
Difference	0.10	-0.05	-0.03	0.05	-0.02
Effect Size		-0.17	-0.11	0.19	-0.06
p-value		0.52	0.85	0.63	0.69
Number of Teachers in Year 2[a]	**54**	**53**	**46**	**31**	**130**

SOURCE: Classroom observations.

NOTE: Each item presented in this table captures the average number of times within a 10-minute interval that the behavior listed was observed throughout the observations conducted in a classroom. For each item, the number reported in the column labeled "Control Group Mean" is the actual average value of the item for the control group, not a regression-adjusted mean. The numbers reported in the remaining columns are, by row: (1) the difference in means between treatment and control group, (2) the effect size, and (3) the *p-value* of the difference. Regression-adjusted differences were calculated taking into account the clustering of teachers within schools. To ensure that the *p-values* from this table are comparable to the *p-values* reported for the differences on the Traditional Interaction scale in Table II.19 (where *p-values* were adjusted for three outcomes), each *p-value* from this table was computed taking into account differences on three outcomes. (Comparability in the approach to adjusting *p-values* is desired because the purpose of the analysis shown in this table is to better understand which specific components of the Traditional Interaction scale are driving the differences, and using a different standard of significance in this table would make that comparison more difficult.) The three outcomes are: (1) the Reading Strategy Guidance scale (see Table II.19), (2) the Classroom Management scale (see Table II.19), and (3) one of the specific items contained in the Traditional Interaction scale. For example, for the first row in this table, *p-values* are adjusted for (1) the Reading Strategy Guidance scale, (2) the Classroom Management scale, and (3) the classroom observation item listed in that row (the extent to which teachers explain how to generate questions). In addition to adjusting the *p-values* for the number of outcomes, it is necessary to adjust the *p-values* to account for the number of comparisons between groups that are being conducted. In particular, for the comparisons of each treatment group and the control group, the results are adjusted for nine comparisons because differences are estimated for each of the three intervention groups for each of the three outcomes. For the combined treatment group, the results are adjusted for three comparisons (since there is a single group being compared to the control group for each of the three outcomes). Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher gender, teacher age, teacher ethnicity and race, and district indicators.

Table II.20 *(continued)*

---

[a]The number of teachers presented in this row is the number of teachers participating in the study in Year 2. Some teachers taught more than one class. The calculations presented in the table are based on the number of classrooms observations for which scale scores were calculated. The response rates for these calculations vary from 90 percent for CRISS classrooms to 100 percent for ReadAbout classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that are adjusted for multiple-hypotheses testing.

TABLE II.21

DIFFERENCES IN CLASSROOM PRACTICES IN THE SECOND STUDY YEAR BETWEEN TREATMENT AND CONTROL GROUP TEACHERS FOR ITEMS CONTAINED IN THE READING STRATEGY GUIDANCE SCALE

		Difference Between Each of the Following and the Control Group:			
	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Comprehension Items					
Teacher Explains Text Structures (Item 2b)					
Difference	0.26	0.05	0.09	-0.10	0.04
Effect Size		0.13	0.25	-0.29	0.10
p-value		0.92	0.75	0.51	0.51
Students Practice Using Text Structures (Item 2c)					
Difference	0.41	0.11	0.18	-0.21	0.07
Effect Size		0.17	0.30	-0.34	0.11
p-value		0.75	0.51	0.26	0.41
Teacher Models Comprehension Strategies (Item 3a)					
Difference	-0.01	0.06	0.02	0.04	0.03
Effect Size		3.34	1.03	2.29	2.09
p-value		0.18	0.64	0.25	0.18
Teacher Explains Comprehension Strategies (Item 3b)					
Difference	0.74	0.08	0.57	0.75*	0.33
Effect Size		0.08	0.55	0.73	0.32
p-value		0.97	0.12	0.00	0.11
Students Practice Comprehension Strategies (Item 3c)					
Difference	1.54	0.09	0.43	0.62	0.22
Effect Size		0.04	0.20	0.28	0.10
p-value		0.99	0.58	0.14	0.43
Teacher Explains How to Generate Questions (Item 4b)					
Difference	0.21	0.08	-0.05	0.11	0.04
Effect Size		0.25	-0.16	0.36	0.13
p-value		0.71	0.86	0.31	0.52
Students Practice Generating Questions (Item 4c)					
Difference	0.33	0.33	-0.06	0.19	0.17
Effect Size		0.94	-0.18	0.53	0.48
p-value		0.10	0.92	0.30	0.09
Teacher Explains Text Features (Item 5b)					
Difference	0.24	-0.15*	-0.14*	0.08	-0.09*
Effect Size		-0.42	-0.38	0.23	-0.25
p-value		0.02	0.04	0.55	0.04
Students Practice Using Text Features (Item 5c)					
Difference	0.31	-0.12	-0.22*	0.09	-0.12
Effect Size		-0.24	-0.45	0.19	-0.25
p-value		0.33	0.01	0.84	0.08

Table II.21 *(continued)*

	Difference Between Each of the Following and the Control Group:				
	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Teacher Asks Students to Justify Response (Item 6c)					
Difference	0.33	-0.05	-0.13	-0.06	-0.09
Effect Size		-0.11	-0.32	-0.13	-0.20
p-value		0.94	0.28	0.91	0.19
Vocabulary Items					
Teacher Teaches Word-Learning Strategies (Item 4)					
Difference	0.14	-0.08	-0.08	-0.02	-0.07
Effect Size		-0.29	-0.28	-0.07	-0.25
p-value		0.21	0.33	0.96	0.15
Number of Teachers in Year 2[a]	**54**	**53**	**46**	**31**	**130**

SOURCE: Classroom observations.

NOTE: Each item presented in this table captures the average number of times within a 10-minute interval that the behavior listed was observed throughout the observations conducted in a classroom. For each item, the number reported in the column labeled "Control Group Mean" is the actual average value of the item for the control group, not a regression-adjusted mean. The numbers reported in the remaining columns are, by row: (1) the difference in means between treatment and control group, (2) the effect size, and (3) the *p-value* of the difference. Regression-adjusted differences were calculated taking into account the clustering of teachers within schools. To ensure that the *p-values* from this table are comparable to the *p-values* reported for the differences on the Reading Strategy Guidance scale in Table II.19 (where *p-values* were adjusted for three outcomes), each *p-value* from this table was computed taking into account differences on three outcomes. (Comparability in the approach to adjusting *p-values* is desired because the purpose of the analysis shown in this table is to better understand which specific components of the Reading Strategy Guidance scale are driving the differences, and using a different standard of significance in this table would make that comparison more difficult.) The three outcomes are: (1) the Traditional Interaction scale (see Table II.19), (2) the Classroom Management scale (see Table II.19), and (3) one of the specific items contained in the Reading Strategy Guidance scale. For example, for the first row in this table, *p-values* are adjusted for (1) the Traditional Interaction scale, (2) the Classroom Management scale, and (3) the classroom observation item listed in that row (the extent to which teachers explain text structure). In addition to adjusting the *p-values* for the number of outcomes, it is necessary to adjust the *p-values* to account for the number of comparisons between groups that are being conducted. In particular, for the comparisons of each treatment group and the control group, the results are adjusted for nine comparisons because differences are estimated for each of the three intervention groups for each of the three outcomes. For the combined treatment group, the results are adjusted for three comparisons (since there is a single group being compared to the control group for each of the three outcomes). Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher gender, teacher age, teacher ethnicity and race, and district indicators.

[a]The number of teachers presented in this row is the number of teachers participating in the study in Year 2. Some teachers taught more than one class. The calculations presented in the table are based on the number of classroom observations for which scale scores were calculated. The response rates for these calculations vary from 90 percent for CRISS classrooms to 100 percent for ReadAbout classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that are adjusted for multiple-hypotheses testing.

TABLE II.22

DIFFERENCES IN CLASSROOM PRACTICES IN THE SECOND STUDY YEAR BETWEEN TREATMENT AND CONTROL GROUP TEACHERS FOR ITEMS CONTAINED IN THE CLASSROOM MANAGEMENT SCALE

	Difference Between Each of the Following and the Control Group:				
	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Teacher Maximized Instruction Time (Item 10)					
Difference	3.28	-0.11	-0.06	0.09	-0.04
Effect Size		-0.15	-0.08	0.13	-0.05
p-value		0.83	0.95	0.94	0.74
Teacher Managed Student Behavior (Item 11)					
Difference	3.53	-0.27	-0.25	-0.03	-0.21*
Effect Size		-0.43	-0.41	-0.04	-0.35
p-value		0.07	0.20	1.00	0.04
Student Engagement in First Half of Lesson (Item 13)					
Difference	2.70	0.05	-0.11	0.17	0.00
Effect Size		0.12	-0.24	0.38	0.01
p-value		0.88	0.48	0.22	0.97
Student Engagement in Second Half of Lesson (Item 14)					
Difference	2.63	-0.03	-0.17	0.15	-0.05
Effect Size		-0.06	-0.34	0.29	-0.10
p-value		0.98	0.17	0.63	0.59
Number of Teachers[a]	**54**	**53**	**46**	**31**	**130**

SOURCE: Classroom observations.

NOTE: Each item presented in this table captures the average number of times within a 10-minute interval that the behavior listed was observed throughout the observations conducted in a classroom. For each item, the number reported in the column labeled "Control Group Mean" is the actual average value of the item for the control group, not a regression-adjusted mean. The numbers reported in the remaining columns are, by row: (1) the difference in means between treatment and control group, (2) the effect size, and (3) the *p-value* of the difference. Regression-adjusted differences were calculated taking into account the clustering of teachers within schools. To ensure that the *p-values* from this table are comparable to the *p-values* reported for the differences on the Classroom Management scale in Table II.19 (where *p-values* were adjusted for three outcomes), each *p-value* from this table was computed taking into account differences on three outcomes. (Comparability in the approach to adjusting *p-values* is desired because the purpose of the analysis shown in this table is to better understand which specific components of the Classroom Management scale are driving the differences, and using a different standard of significance in this table would make that comparison more difficult.) The three outcomes are: (1) the Traditional Interaction scale (see Table II.19), (2) the Reading Strategy Guidance scale (see Table II.19), and (3) one of the specific items contained in the Classroom Management scale. For example, for the first row in this table, *p-values* are adjusted for (1) the Traditional Interaction scale, (2) the Reading Strategy Guidance scale, and (3) the classroom observation item listed in that row (the extent to which teachers maximized instruction time). In addition to adjusting the *p-values* for the number of outcomes, it is necessary to adjust the *p-values* to account for the number of comparisons between groups that are being conducted. In particular, for the comparisons of each treatment group and the control group, the results are adjusted for nine comparisons because differences are estimated for each of the three intervention groups for each of the three outcomes. For the combined treatment group, the results are adjusted for three comparisons (since there is a single group being compared to the control group for each of the three outcomes).Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher gender, teacher age, teacher ethnicity and race, and district indicators.

Table II.22 *(continued)*

[a]The number of teachers presented in this row is the number participating in the study in Year 2. Some teachers taught more than one class. The calculations presented in the table are based on the number of classroom observations for which scale scores were calculated. The response rates for these calculations vary from 90 percent for CRISS classrooms to 100 percent for ReadAbout classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that are adjusted for multiple-hypotheses testing.

were made to *p*-values for differences involving three outcomes), each *p*-value from these sensitivity tests was computed taking into account an adjustment of a similar magnitude.[49]

These analyses show that the differences observed on the Traditional Interaction scale were driven by differences in teaching practices related to comprehension instruction (as opposed to vocabulary instruction). In particular, 21.4 percent (6 of 28) of the differences in teaching practices related to *comprehension* instruction were statistically significant (with lower levels for the treatment group than the control group), compared to no statistically significant differences in teaching practices related to *vocabulary* instruction (Table II.20). Statistically significant differences were found for the following *comprehension*-related teaching practices (in all cases, treatment group teachers engaged in these practices less than did control group teachers):

- Teachers explaining text features, which was statistically significant for Project CRISS, ReadAbout, and the combined treatment group (effect sizes: -0.42, -0.38, and -0.25, respectively)
- Teachers having students practice using text features, which was statistically significant for ReadAbout (effect size: -0.45)
- Teachers asking questions based on material in text beyond a literal level, which was statistically significant for Project CRISS and the combined treatment group (effect sizes: -0.45 and -0.28, respectively)

We also found differences on the individual teacher practices included in the Reading Strategy Guidance scale. While 3 of 40 (7.5 percent) differences in teacher practices related to comprehension instruction were statistically significant, none of the differences in teacher practices related to vocabulary instruction were statistically significant (Table II.21). We found statistically significant differences in the following *comprehension*-related practices that are part of the Reading Strategy Guidance scale:

- Teachers explaining comprehension strategies, which was statistically significantly higher for Read for Real teachers than for control group teachers (effect size: 0.73)

[49]Comparability in the approach to adjusting *p*-values is important because the purpose of this analysis is to better understand which specific components of each scale are driving the overall differences between the treatment and control groups, and using a different standard of significance in this analysis would make that comparison more difficult. For each of these teaching practices analyses, the three outcomes for which we adjusted include one of the specific items contained in the scale currently being analyzed and the other two scales. For example, for the first row in Table II.20, *p*-values are adjusted for (1) the classroom observation item listed in that row (item 4b: the extent to which teachers explain how to generate questions), (2) the Reading Strategy Guidance scale, and (3) the Classroom Management scale. In addition to adjusting the *p*-values for the number of outcomes, it is necessary to adjust the *p*-values to account for the number of comparisons between groups that are being conducted. In particular, for the comparisons of each treatment group and the control group, the results are adjusted for nine comparisons because models are being estimated for each of the three intervention groups for each of the three outcomes. For the combined treatment group, the results are adjusted for three comparisons (because there is a single group being compared to the control group for each of the three outcomes).

- As described above, the Reading Strategy Guidance and Traditional Interaction scales share six overlapping items. For two of these shared items, statistically significant differences were found in the number of times teachers were observed: (1) explaining text features, which was statistically significantly lower for Project CRISS teachers, ReadAbout teachers, and the combined treatment group than for control group teachers and (2) having students practice using text features, which was statistically significantly lower for ReadAbout teachers than for control group teachers (see bullets above for effect sizes of these differences).

For the teacher practices included in the Classroom Management scale (Table II.22), we found one statistically significant difference (6.3 percent of the 16 differences):

- Teachers in the combined treatment group received lower ratings of their management of student behavior than teachers in the control group (effect size: -0.35)

Instructional Practices of Teachers in Study for Two Years vs. Teachers New to Study in Year 2. We investigated whether treatment/control differences in instructional practices differ between two types of teachers during the study's second year: (1) teachers who are new to the study and (2) fifth-grade teachers who had been in the study for two consecutive years. This analysis aims to determine whether treatment/control differences in instructional practices differ between teachers who have implemented the interventions for one year and teachers who are implementing their assigned curriculum for the first time. We found the following statistically significant differences when comparing instructional practices in Year 2 of teachers participating in the study for two years and teachers new to the study (Table II.23):

- Project CRISS teachers participating in the study for two years had statistically significantly lower Traditional Interaction scale scores than control group teachers participating in the study for two years (effect size: -0.80).

- The Project CRISS/control group difference on the Traditional Interaction scale for teachers in the study for two years was statistically significantly different from the Project CRISS/control group difference on that scale for teachers new to the study (difference in effect size: -0.95). The Project CRISS/control group difference on the Traditional Interaction scale of teachers new to the study in Year 2 was not statistically significant.

- The combined treatment group/control group difference on the Traditional Interaction scale for teachers in the study for two years was statistically significantly different from the combined treatment group/control group difference on that scale for teachers new to the study (difference in effect size: -0.68). The combined treatment group/control group differences on the Traditional Interaction scale of the two groups of teachers were not statistically significant.

Instructional Practices in Years 1 and 2 of Teachers Participating in the Study for Two Years. We conducted two additional analyses to examine how treatment/control differences in

TABLE II.23

DIFFERENCES IN CLASSROOM PRACTICES IN THE SECOND STUDY YEAR BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING TEACHERS PARTICIPATING IN THE STUDY FOR TWO YEARS WITH TEACHERS NEW TO THE STUDY

	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Traditional Interaction Scale				
Teachers in Study for Two Years				
Impact	-4.60*	-3.42	1.12	-2.80
Effect Size	-0.80	-0.59	0.19	-0.49
p-value	0.01	0.18	0.90	0.06
Teachers New to the Study				
Impact	0.85	-0.37	3.45	1.13
Effect Size	0.15	-0.06	0.60	0.20
p-value	0.98	1.00	0.61	0.74
Difference Between Teachers in Study for Two Years and Teachers New to the Study				
Difference in Impact	-5.46*	-3.05	-2.33	-3.93*
Difference in Effect Size	-0.95	-0.53	-0.41	-0.68
p-value for the Difference	0.02	0.49	0.72	0.03
Reading Strategy Guidance Scale				
Teachers in Study for Two Years				
Impact	2.37	2.24	0.24	1.81
Effect Size	0.39	0.37	0.04	0.30
p-value	0.18	0.37	1.00	0.18
Teachers New to the Study				
Impact	0.61	-1.79	2.46	0.11
Effect Size	0.10	-0.30	0.41	0.02
p-value	1.00	0.99	0.89	1.00
Difference Between Teachers in Study for Two Years and Teachers New to the Study				
Difference in Impact	1.76	4.03	-2.22	1.71
Difference in Effect Size	0.29	0.67	-0.37	0.28
p-value for the Difference	0.85	0.48	0.78	0.55
Classroom Management Scale				
Teachers in Study for Two Years				
Impact	-12.83	-9.63	-2.12	-9.01
Effect Size	-0.48	-0.36	-0.08	-0.34
p-value	0.28	0.55	1.00	0.24
Teachers New to the Study				
Impact	-1.47	-9.11	16.00	0.00
Effect Size	-0.06	-0.34	0.60	0.00
p-value	1.00	0.96	0.85	1.00
Difference Between Teachers in Study for Two Years and Teachers New to the Study				
Difference in Impact	-11.35	-0.52	-18.12	-9.02
Difference in Effect Size	-0.43	-0.02	-0.68	-0.34
p-value for the Difference	0.64	1.00	0.50	0.42

TABLE II.23 (*continued*)

	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Number of Year-Two Teachers[a]	**53**	**46**	**31**	**130**
In Study for Two Consecutive Years	**35**	**34**	**22**	**91**
New to Study	**18**	**12**	**9**	**39**

SOURCE: Classroom observations.

NOTE: The scales presented in this table were constructed to capture the frequency of the behaviors in each instructional practice domain shown above. For each scale and each group of Cohort 2 teachers, the numbers reported are, by row: (1) the difference in means between treatment and control group, (2) the effect size, and (3) the *p-value* of the difference. For each scale, the differences between impacts for teachers in the study for two years and teachers new to the study are also reported. The *p-values* presented in this table are adjusted for multiple-hypotheses testing. Regression-adjusted differences were calculated taking into account the clustering of teachers within schools. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher gender, teacher age, teacher ethnicity and race, and district indicators.

[a]Counts reflect the number of teachers participating in the study in Year 2. Some teachers taught more than one class. The calculations presented in the table are based on the number of classroom observations for which scale scores were calculated. The response rates for these calculations vary from 90 percent for CRISS classrooms to 100 percent for ReadAbout classrooms.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

instructional practices changed between the first and second years of the study for teachers participating in the study both years. First, we compared the treatment/control differences in teachers' instructional practices in the first and second years of the study, as measured by the three teacher practice scales described above (see Table II.24). Second, we compared the treatment/control differences in teachers' instructional practices in the first and second years of the study, as measured by the average number of times teachers engaged in specific teaching practices included in the ERC (Table II.25). These analyses were conducted to examine whether—for teachers in the study both years—treatment/control differences in teaching practices in the second year (after teachers had a year of experience with the interventions) were larger than those in the first year.

We found one statistically significant treatment/control difference in instructional practice scales in the study's first and second years for teachers participating in the study both years (Table II.24).

- In the study's second year, Project CRISS teachers exhibited lower Traditional Interaction scale scores than control group teachers (effect size: -0.75). This difference was not statistically significantly different from the Project CRISS/control group difference on this scale in the study's first year.

In the analyses focused on individual instructional practices, we found the following statistically significant treatment/control differences in individual instructional practices in the study's second year for teachers participating in the study both years (Table II.25):

- Project CRISS teachers were observed:
  - Providing fewer explanations or definitions than control teachers
  - Asking students to do something requiring word knowledge less than control teachers
  - Teaching using outlining and/or note taking more than control teachers
  - Using graphic organizers more than control teachers
- ReadAbout teachers were observed having students practice using text features to interpret text less than control teachers
- Read for Real teachers were observed asking students to justify their responses less than control teachers. The treatment/control difference in this practice in Year 1 was statistically significantly different from the treatment/control difference in Year 2 (not shown in table).
- Teachers in the combined treatment group were observed teaching using outlining and/or note taking more than control teachers.

We hypothesized that greater experience in implementing the interventions would lead to larger treatment/control group differences in instructional practices. There is little evidence to support this hypothesis. In 93 of 100 tests conducted (Tables II.24 and II.25), there were no

TABLE II.24

DIFFERENCES IN CLASSROOM PRACTICE SCALES BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING SCALES IN YEARS 1 AND 2 FOR FIFTH-GRADE TEACHERS PARTICIPATING IN THE STUDY FOR TWO CONSECUTIVE YEARS

	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Traditional Interaction Scale				
Year 1				
Impact	-1.96	-1.86	-2.72	-1.76
Effect Size	-0.30	-0.29	-0.42	-0.27
p-value	0.45	0.65	0.23	0.20
Year 2				
Impact	-4.81*	-3.93	0.96	-2.69
Effect Size	-0.75	-0.61	0.15	-0.42
p-value	0.00	0.06	0.91	0.05
Difference Between Year 1 and Year 2				
Difference in Impact	-2.85	-2.07	3.68	-0.93
Difference in Effect Size	-0.44	-0.32	0.57	-0.14
p-value for the Difference	0.28	0.56	0.10	0.55
Reading Strategy Guidance Scale				
Year 1				
Impact	0.68	1.83	1.40	1.15
Effect Size	0.10	0.27	0.21	0.17
p-value	1.00	0.87	0.91	0.64
Year 2				
Impact	2.73	2.22	-0.01	1.68
Effect Size	0.40	0.33	0.00	0.25
p-value	0.15	0.57	1.00	0.25
Difference Between Year 1 and Year 2				
Difference in Impact	2.04	0.39	-1.42	0.53
Difference in Effect Size	0.30	0.06	-0.21	0.08
p-value for the Difference	0.65	1.00	0.76	0.77
Classroom Management Scale				
Year 1				
Impact	7.08	-2.46	7.57	4.25
Effect Size	0.23	-0.08	0.25	0.14
p-value	0.77	1.00	0.72	0.67
Year 2				
Impact	-9.94	-11.32	-2.17	-7.59
Effect Size	-0.32	-0.37	-0.07	-0.25
p-value	0.60	0.47	1.00	0.38
Difference Between Year 1 and Year 2				
Difference in Impact	-17.02	-8.86	-9.74	-11.83
Difference in Effect Size	-0.55	-0.29	-0.32	-0.38
p-value for the Difference	0.12	0.60	0.53	0.12
Number of Fifth-Grade Teachers Participating in Study for Two Consecutive Years	**35**	**34**	**22**	**91**

TABLE II.24 (*continued*)

SOURCE: Reading comprehension tests administered by study team; classroom observations.

NOTE: The scales presented in this table were constructed to capture the frequency of the behaviors in each instructional practice domain shown above. For each scale and each year, the numbers reported are, by row: (1) the difference in means between treatment and control group, (2) the effect size, and (3) the *p-value* of the difference. For each scale, the differences between impacts in Years 1 and 2 are also reported. The *p-values* presented in this table are adjusted for multiple hypotheses testing. Regression-adjusted differences were calculated taking into account the clustering of teachers within schools. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher gender, teacher age, teacher ethnicity and race, and district indicators.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE II.25

DIFFERENCES IN INDIVIDUAL CLASSROOM PRACTICES BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING INDIVIDUAL PRACTICES IN YEARS 1 AND 2 FOR FIFTH-GRADE TEACHERS PARTICIPATING IN THE STUDY FOR TWO CONSECUTIVE YEARS

		Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Part I, Comprehension					
Activates prior knowledge and/or previews text before reading					
Teacher models	Year 1	0.02	-0.01	0.01	0.01
	Year 2	-0.00	-0.00	0.00	0.00
Teacher explains, reviews, provides examples and elaborations	Year 1	-0.01	-0.23	0.22	0.01
	Year 2	0.09	0.23	0.31	0.24
Students practice	Year 1	0.13	-0.30	0.54	0.13
	Year 2	0.38	0.29	0.70	0.45
Explicit comprehension instruction that teaches students about text structure					
Teacher models	Year 1	-0.00	0.01	0.00	0.01
	Year 2	0.01	0.01	0.02	0.01
Teacher explains, reviews, provides examples and elaborations	Year 1	-0.07	0.36	0.20	0.15
	Year 2	0.08	0.13	-0.06	0.06
Students practice	Year 1	-0.07	0.41	0.17	0.18
	Year 2	0.15	0.27	-0.15	0.11
Explicit comprehension instruction that teaches students how to use comprehension strategies					
Teacher models	Year 1	0.00	0.01	0.02	0.01
	Year 2	0.08	0.03	0.02	0.05
Teacher explains, reviews, provides examples and elaborations	Year 1	-0.22	0.35	0.35	0.07
	Year 2	0.33	0.69	0.69	0.47
Students practice	Year 1	-0.20	0.52	-0.07	0.08
	Year 2	0.46	0.54	0.62	0.37
Explicit comprehension instruction that teaches students how to generate questions					
Teacher models	Year 1	-0.02	-0.02	-0.02	-0.01
	Year 2	0.01	0.00	0.02	0.01
Teacher explains, reviews, provides examples and elaborations	Year 1	-0.01	-0.14	-0.19	-0.09
	Year 2	0.07	-0.05	-0.01	0.02
Students practice	Year 1	0.03	-0.07	-0.20	-0.05
	Year 2	0.41	0.03	0.10	0.21
Explicit comprehension instruction that teaches text features to interpret text					
Teacher models	Year 1	0.01	0.01	0.00	0.01
	Year 2	-0.00	-0.01	-0.01	-0.00
Teacher explains, reviews, provides examples and elaborations	Year 1	-0.04	0.00	0.05	0.01
	Year 2	-0.13	-0.19	0.06	-0.08
Students practice	Year 1	-0.10	0.03	0.05	-0.00
	Year 2	-0.14	-0.29*	0.05	-0.15
Teacher asks students to justify their responses					
	Year 1	0.02	-0.05	-0.09	-0.02
	Year 2	-0.19	-0.15	-0.31*	-0.16

TABLE II.25 (*continued*)

		Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Teacher asks questions based on material in the text that are beyond the literal level					
	Year 1	-0.40	-0.29	-0.33	-0.35
	Year 2	-0.85	-0.79	-0.25	-0.66
Teacher elaborates, clarifies, or links concepts during and after text reading					
	Year 1	-0.62	-0.46	-0.25	-0.49
	Year 2	-0.76	-0.87	-0.31	-0.68
Part I, Vocabulary					
Teacher provides an explanation and/or a definition or asks a student to read a definition					
	Year 1	-0.43*	-0.12	-0.07	-0.21
	Year 2	-0.39*	-0.01	0.01	-0.12
Teacher provides examples, contrasting examples, multiple meanings, immediate elaborations to students' responses					
	Year 1	-0.51	-0.52	-0.54	-0.54*
	Year 2	-0.47	-0.13	0.35	-0.18
Teacher uses visuals/pictures, gestures related to word meaning, facial expressions, or demonstrations to discuss/demonstrate word meanings					
	Year 1	-0.16	-0.18	-0.20	-0.20
	Year 2	-0.29	-0.23	0.02	-0.21
Teacher teaches word-learning strategies using context clues, word parts, root meaning					
	Year 1	-0.14*	-0.04	-0.12*	-0.10*
	Year 2	-0.11	-0.12	-0.05	-0.10
Students do or are asked to do something that requires knowledge of words					
	Year 1	-0.41	-0.74	-0.61	-0.60
	Year 2	-0.79*	-0.17	0.74	-0.23
Students are given an opportunity to apply word-learning strategies using context clues, word parts, and root meaning					
	Year 1	-0.04	0.19	-0.01	0.07
	Year 2	-0.07	-0.02	0.06	-0.02
Part II, Instruction Effectiveness					
Gave inaccurate and/or confusing explanations or feedback					
	Year 1	-0.02	-0.02	-0.02	-0.02
	Year 2	0.15	0.05	0.06	0.08
Missed opportunity to correct or address error					
	Year 1	0.03	0.03	-0.04	0.01
	Year 2	0.04	0.02	0.05	0.03
Provided opportunities for most students to participate actively during teacher-led instruction					
	Year 1	0.07	0.05	0.08	0.08
	Year 2	-0.05	0.00	0.16	0.02
Paced instruction so that the length of the comprehension or vocabulary activities was appropriate for this age group					
	Year 1	0.05	0.07	0.10	0.09
	Year 2	-0.15	-0.09	-0.10	-0.10
Taught using outlining and/or note taking					
	Year 1	0.39*	0.07	0.15	0.21*
	Year 2	0.41*	0.07	0.21	0.23*

TABLE II.25 (*continued*)

	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Used graphic organizers				
Year 1	0.32*	-0.00	0.10	0.14
Year 2	0.26*	-0.04	-0.01	0.09
Kept students thinking for two or more seconds before calling on a student to respond to a complex question				
Year 1	0.06	-0.05	-0.15	-0.04
Year 2	-0.05	0.01	0.00	-0.01
Gave independent/pairs/small-group practice in answering comprehension questions or applying comprehension strategy(ies) with expected written product				
Year 1	0.34*	0.25	0.04	0.22*
Year 2	0.07	0.05	-0.08	0.03
Used writing activities in response to reading (does not include fill-in-the-blank or one-word answers)				
Year 1	0.05	-0.09	-0.04	-0.03
Year 2	0.16	0.12	0.01	0.14
Part II, Teachers' Management/Responsiveness to Students				
Teacher maximized the amount of time available for instruction				
Year 1	0.09	-0.19	0.27	0.05
Year 2	-0.33	-0.22	-0.14	-0.20
Teacher managed student behavior effectively in order to avoid disruptions and provide productive learning environments				
Year 1	0.12	-0.06	0.04	0.03
Year 2	-0.31	-0.31	-0.07	-0.24
Teacher redirected discussion if a student response was leading the group off topic/focus				
Year 1	0.22	-0.01	0.32	0.16
Year 2	-0.35	-0.13	-0.79	-0.33
Part II, Student Engagement				
Student engagement during the first half of the observation session				
Year 1	0.21	-0.07	0.14	0.08
Year 2	-0.03	-0.15	0.19	-0.03
Student engagement during the remainder of the observation session				
Year 1	0.18	-0.10	0.14	0.08
Year 2	-0.11	-0.18	0.18	-0.06
Number of Fifth-Grade Teachers Participating in Study for Two Consecutive Years	**35**	**34**	**22**	**91**

SOURCE: Reading comprehension tests administered by study team; classroom observations.

NOTE: For each item and each year, the numbers reported are the difference in means between the treatment and control groups. Variables in the regression model include baseline GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher gender, teacher age, teacher ethnicity and race, and district indicators.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

differences in instructional practices across the treatment and control groups in the second year of the study.

E. INTERVENTION AND CONTROL GROUP TEACHERS' TIME ALLOCATION

Knowing how the implementation of the interventions affected teachers' use of time during the school day is important for understanding the impacts of the curricula. This is particularly relevant in the context of this study, as the school day was not extended to facilitate implementation of the study interventions. Instead, teachers needed to use the study curricula within the confines of the existing school day, meaning that less time had to be spent engaged in other activities that could also have affected students' reading comprehension. Therefore, it is important to understand how teachers reallocated their time during the day to facilitate use of the interventions.

In the second year of the study, as described in Chapter I, we collected data to help better understand these issues. The data collected allowed for three main analyses. First, to assess whether the interventions affected the type of activities in which teachers were engaged, we compared treatment and control teachers' allocation of time during the day to various activities. Second, to assess whether the interventions affected the amount of time teachers devoted to informational text, we compared the amount of time treatment and control teachers reported using informational text with students in a typical week. Third, we asked treatment teachers to report the activities that they cut back on to make room for using the study interventions.

Findings from the first set of analyses showed three statistically significant differences between Project CRISS and control group teachers' time allocation (Table II.26). In particular, Project CRISS teachers were statistically significantly less likely than control teachers to report engaging in enrichment activities (such as art, music, or physical education), noncurricular activities (such as lunch, recess, or arrival/dismissal activities), and other activities. Similar patterns were observed for ReadAbout and Read for Real, but those differences were not statistically significant. Teachers from all three intervention groups reported spending a higher proportion of their day on reading activities relative to control teachers, but those differences were not statistically significant. The proportion of time treatment and control teachers spent in other activities did not differ significantly.

The second set of analyses showed no statistically significant differences in the average number of minutes that treatment and control teachers reported using informational text with students (Table II.27). Therefore, there is no evidence that use of the study interventions led to an increase in use of informational text. Data collected from treatment teachers on the amount of time they spent using the study interventions with students in a typical week ranged from 132 minutes for ReadAbout teachers to 192 minutes for Project CRISS teachers.

In the third set of analyses, we found that, across all three treatment groups, teachers were most likely to report reducing time spent on other reading activities to make time for the use of the study interventions (Table II.28). Twelve percent of Project CRISS teachers, 18 percent of ReadAbout teachers, and 21 percent of Read for Real teachers reported reducing time spent on other reading activities to facilitate their use of the study interventions. The reduction in time

TABLE II.26

TEACHER-REPORTED TIME ALLOCATION AS PROPORTION OF SCHOOL DAY, COHORT 2 FIFTH-GRADE CLASSROOMS

	Control Group		Project CRISS		ReadAbout		Read for Real	
Activity	Percentage of Teachers Engaging in Activity	Among All Teachers, Average Proportion of Time Spent in Activity	Percentage of Teachers Engaging in Activity	Among All Teachers, Average Proportion of Time Spent in Activity	Percentage of Teachers Engaging in Activity	Among All Teachers, Average Proportion of Time Spent in Activity	Percentage of Teachers Engaging in Activity	Among All Teachers, Average Proportion of Time Spent in Activity
Reading Activities[a]	90.5	30.1	85.4 (0.86)	38.1 (0.05)	80.7 (0.62)	35.6 (0.10)	92.6 (0.98)	35.5 (0.27)
Other Academic Activities[b]	90.4	37.3	80.9 (0.51)	34.6 (0.80)	84.0 (0.79)	39.8 (0.80)	90.0 (1.00)	35.7 (0.95)
Enrichment Activities[c]	96.1	12.3	73.0* (0.02)	11.8 (1.00)	84.0 (0.23)	10.0 (0.37)	78.8 (0.18)	10.4 (0.69)
Noncurricular Activities[d]	100.0	16.8	82.9* (0.03)	14.1 (0.12)	90.4 (0.16)	12.9 (0.07)	97.5 (0.68)	17.1 (0.99)
Other	41.2	3.5	16.2* (0.03)	1.3 (0.11)	30.1 (0.84)	1.7 (0.33)	15.6 (0.06)	1.2 (0.17)
Number of Teachers[e]	**54**		**53**		**46**		**31**	

SOURCE: Teacher Time Allocation Survey.

NOTE: For each activity and intervention group, we report the percentage of teachers who report engaging in the activity, the average proportion of time in the school day engaged in the activity, and two *p-values* corresponding to tests of whether the percentage of teachers who reported engaging in each activity and the average proportion of time spent in each activity differ from the control group. We assumed that teachers who did not report spending time in a particular activity did not engage in that activity. The average proportion of time spent in each activity is based on all teachers, including those who did not engage in the activity.

[a]This category includes the following items: (1) Separate Instruction Using Intervention Curriculum (CRISS, ReadAbout, and Read for Real); (2) Core (Basal) Reading Curriculum; (3) Supplemental Reading Curriculum (supplemental curricula other than the study interventions); (4) Comprehension; (5) Vocabulary; (6) Fluency; (7) Reading Lesson Using Fiction Materials; (8) Reading Lesson Using Nonfiction Materials; and (9) Other Language Arts Activity.

TABLE II.26 (*continued*)

[b]This category includes the following items: (1) Science Instruction (Textbooks), (2) Science Lab/Hands-on, (3) Social Studies/History, (4) News/Current Events, and (5) Computer Instruction.

[c]This category includes the following items: (1) Health or Family Life Education; (2) Library; (3) Physical Education; (4) Art, (5) Music (general music, chorus, band, or strings); and (6) Enrichment.

[d]This category includes the following items: (1) Arrival, Homeroom, Announcements; (2) Lunch; (3) Recess; and (4) Dismissal Activities.

[e]The number of teachers presented in this row is the number of teachers participating in the second year of the study.

*Statistically different at the .05 level, *p-value* adjusted for multiple-hypotheses testing.

TABLE II.27

TIME SPENT USING INFORMATIONAL TEXT AND TIME SPENT USING INTERVENTION IN COHORT 2 FIFTH-GRADE CLASSROOMS

	Control Group	Project CRISS	ReadAbout	Read for Real
Number of Minutes of Class Time That Teachers Reported Students Spent Using Informational Text in a Typical Week (Average)[a]	422.2	424.4 (1.00)	428.4 (1.00)	492.6 (0.86)
Number of Minutes of Class Time That Teachers Reported Students Spent Using Their Assigned Intervention in a Typical Week (Average)	n.a.	191.9	131.7	162.1
Number of Teachers[b]	**54**	**53**	**46**	**31**

SOURCE: Students' Use of Informational Text in Class Survey.

NOTE: Time spent using informational text refers to all time periods (reading/language arts, science, social studies, test preparation, and any other class period). It includes time spent teaching comprehension strategies or vocabulary instruction related to text, as well as time students spend reading informational text, participating in whole-class discussions involving oral answers to teachers' questions or small-group discussions of text, or completing worksheets or other written assignments about text. Time spent using their assigned intervention refers to the interventions being evaluated on this study. For example, for ReadAbout, this refers to the amount of time students who were assigned to ReadAbout spent using it in a typical week.

[a]Averages are shown for each group. Below the averages for the intervention groups, we show the *p-value* corresponding to the test of whether each intervention group average differs from the control group average. This *p-value* is not adjusted for multiple comparisons.

[b]The number of teachers presented in this row is the number of teachers participating in the second year of the study.

n.a. = not applicable.

*Statistically different at the .05 level, *p-value* adjusted for multiple-hypotheses testing.

TABLE II.28

TEACHER-REPORTED REDUCTION IN TIME SPENT ON CLASSROOM ACTIVITIES DUE TO USE OF TREATMENT CURRICULUM, COHORT 2 FIFTH-GRADE CLASSROOMS

	Project CRISS			ReadAbout			Read for Real		
Activity	Among Teachers Engaging in Activity, Percentage Who Reported Reducing or Eliminating Time Spent on Activity to Make Room for Implementing CRISS	Among Teachers Who Reported Reducing or Eliminating Time Spent on Activity, Average Reported Reduction in Minutes Spent on Activity	Percentage of Teachers Who Reported Entirely Eliminating Activity to Make Room for Implementing CRISS	Among Teachers Engaging in Activity, Percentage Who Reported Reducing or Eliminating Time Spent on Activity to Make Room for Implementing ReadAbout	Among Teachers Who Reported Reducing or Eliminating Time Spent on Activity, Average Reported Reduction in Minutes Spent on Activity	Percentage of Teachers Who Reported Entirely Eliminating Activity to Make Room for Implementing ReadAbout	Among Teachers Engaging in Activity, Percentage Who Reported Reducing or Eliminating Time Spent on Activity to Make Room for Implementing Read for Real	Among Teachers Who Reported Reducing or Eliminating Time Spent on Activity, Average Reported Reduction in Minutes Spent on Activity	Percentage of Teachers Who Reported Entirely Eliminating Activity to Make Room for Implementing Read for Real
Reading Activities,[a] Excluding Separate Instruction Using Intervention Curriculum	12.2	29.2	—[b]	17.9	20.0	8.8	21.1	52.5	—[b]
Other Academic Activities[c]	—[b]	50.0	0.0	10.3	28.3	—[b]	17.1	18.3	8.9
Enrichment Activities[d]	0.0	n.a.	0.0	0.0	n.a.	0.0	13.9	25.0	—[b]
Noncurricular Activities[e]	0.0	n.a.	0.0	—[b]	5.0	0.0	—[b]	26.7	0.0
Other	0.0	n.a.	0.0	—[b]	30.0	—[b]	0.0	n.a.	0.0
Number of Teachers[f]	**53**			**46**			**31**		

SOURCE: Teacher Time Allocation Survey.

NOTE: Teachers in the three intervention groups were asked if they reduced time spent on specific activities in order to accommodate the reading intervention assigned to them as part of this study. The reduction in time spent on activities in this table may not correspond to the impacts on average proportion of time spent in activities reported in Table II.25 because those impacts were based on differences between the treatment and control groups in reported time spent engaged in activities, whereas the time reduction reported in this table is based only on reports from teachers in the study's intervention groups. We assumed that teachers who did not report spending time in a particular activity did not engage in that activity and thus did not reduce time spent on the activity.

[a]This category includes the following items: (1) Core (Basal) Reading Curriculum, (2) Supplemental Reading Curriculum (supplemental curricula other than the study interventions), (3) Comprehension, (4) Vocabulary, (5) Fluency, (6) Reading Lesson Using Fiction Materials, (7) Reading Lesson Using Nonfiction Materials, and (8) Other Language Arts Activity.

[b]Value suppressed to protect teacher confidentiality.

[c]This category includes the following items: (1) Science Instruction (Textbooks), (2) Science Lab/Hands-on, (3) Social Studies/History, (4) News/Current Events, and (5) Computer Instruction.

[d]This category includes the following items: (1) Health or Family Life Education; (2) Library; (3) Physical Education; (4) Art; (5) Music (general music, chorus, band, or strings); and (6) Enrichment.

[e]This category includes the following items: (1) Arrival, Homeroom, Announcements; (2) Lunch; (3) Recess; and (4) Dismissal Activities.

[f]The number of teachers presented in this row is the number of teachers participating in the second year of the study.

n.a. = not applicable.

spent on other reading activities by these teachers ranged from 20 minutes for ReadAbout teachers to 53 minutes for Read for Real teachers.

Some teachers also reported reducing time spent on other non-reading academic activities such as science or social studies. Ten percent of ReadAbout teachers and 17 percent of Read for Real teachers reported reducing time spent on other academic activities to facilitate their use of the study interventions. Fourteen percent of Read for Real teachers reported reducing time spent on enrichment activities such as health, library, physical education, art, or music.

This page is intentionally left blank.

III. COMPARING POST-TEST IMPACTS FOR THE FIRST AND SECOND COHORTS OF FIFTH-GRADE STUDENTS

The analysis of impacts for the second cohort of fifth-grade students was designed to answer primary research questions about whether the reading comprehension interventions are more effective after schools and teachers have had a year of experience implementing them. The secondary questions focus on for whom and under what conditions the interventions are effective. Answers to the primary questions are expected to be of greatest interest to policymakers, since they indicate whether the interventions have the intended effect of improving reading comprehension after schools and teachers have had a year of experience with them. Addressing secondary questions can help interpret answers to the basic questions and guide future research on reading comprehension interventions. Selecting a set of primary questions on intervention effectiveness from the many questions of interest in this study is a way to limit proliferation of impact tests that could, if all were treated as core evaluation issues, just by chance yield some impacts that meet statistical standards for significance (see Schochet 2008 for a detailed discussion of multiple testing). Focusing on these core questions reduces the number of impact tests, which reduces the loss in statistical precision that occurs when we apply corrections for the multiple comparisons that are being made in this study.

EXAMINING THE EFFECTS OF EXPERIENCE

- Two primary research questions:
  - Are impacts larger after *schools* have had one year of experience with the curricula?
  - Are impacts larger after *teachers* have had one year of experience with the curricula?
- Estimating experience effects:
  - The effect of *school* experience is estimated by interacting a cohort variable with treatment variables.
  - The effect of *teacher* experience is estimated by interacting a cohort variable with a teacher experience variable and treatment variables.
  - Both models account for clustering of students within schools, use covariate adjustment to improve statistical precision, adjust p-values for multiple comparisons, and use weights that account for random assignment probabilities and nonresponse.
- Summary of findings:
  - No statistically significant impacts after *schools* have one year of experience.
  - One statistically significant impact after *teachers* have one year of experience: ReadAbout had a positive, statistically significant impact of 0.22 standard deviations on the social studies reading comprehension assessment.

We hypothesize that impacts on student test scores could be larger after schools and teachers have experience with the supplemental curricula. When a *school* has previously used a supplemental curriculum, there could be more resources for teachers to draw on within the

school to aid in their implementation of the curriculum. For example, new teachers could benefit from the experience of their colleagues who had previously used the curriculum. When a *teacher* has previously used a supplemental curriculum, she might be more effective at using it a second time. Thus, we examine whether impacts are larger after both schools and teachers have experience with the interventions. In the case of the school analysis, we look at impacts for all second-cohort students. In the case of the teacher analysis, we look at impacts only for students taught by teachers who were in the study schools in the first year (that is, those teachers who already used the curricula).

This chapter first presents information about the methods used to estimate impacts in the second year of the study (Section A). Section B then examines the comparability of the treatment and control groups. Section C focuses on primary questions of intervention effectiveness and Sections D and E focus on the secondary questions referenced above. In particular, Section C presents impacts on student test scores, focusing on results for two questions: (1) Are impacts larger after *schools* have had one year of experience with the intervention? and (2) Are impacts larger after *teachers* have had one year of experience with the intervention? Section D presents impacts for subgroups of students, defined based on characteristics of the students and their teachers, and the conditions in their schools. In Section E, we examine whether (and, if so, how) impacts are related to differences in teachers' classroom practices.

A. METHODS FOR ESTIMATING IMPACTS

The impacts presented in this chapter are based on our "benchmark" approach. This benchmark approach consists of the methods the study team deemed most appropriate for this study. In particular, the study team decided on an approach that involved accounting for clustering of students within schools (to account for the correlation between students in the same schools) and adjusting the results from statistical tests (*p*-values) for multiple comparisons (because there are multiple outcomes, multiple treatment groups being compared to a single control group, and two cohorts). Unadjusted results are presented in Appendix K.

Two types of impacts are presented. First, impacts are presented for each intervention (for example, outcomes of students in ReadAbout schools are compared with outcomes of students in the control group). These impacts provide information on the effectiveness of each intervention, which may be helpful to readers considering implementing one of the interventions included in the study. The impact of an individual intervention on student outcomes is given by the regression-adjusted difference in outcomes between students in that intervention group and students in the control group. Second, impacts are presented for the combined treatment group, based on outcomes of students in all four intervention groups and outcomes of students in the control group. These impacts provide information on the effectiveness of reading comprehension interventions more broadly (not the specific impacts of any one intervention). Impacts for the combined treatment group are presented for two reasons. First, although the details of each intervention differ, the four interventions share a set of common strategies for improving reading comprehension. As a result, examining the interventions as a group is a reasonable approach to address the question of whether the use of these types of interventions, in general, improves comprehension. Second, examining the combined treatment group gives the study more power than looking at an individual treatment group. The impact of the curricula as a whole on student outcomes is given by the regression-adjusted difference in outcomes between students in the

combined treatment group and students in the control group. The *p*-values for all of these impacts are adjusted for multiple comparisons (*p*-values that are not adjusted for multiple comparisons are presented in Appendix K).[50]

To increase the statistical precision of the study's impact estimates, we estimated impact models that controlled for student, teacher, and school characteristics. These included students' baseline GRADE and TOSCRF scores, ELL status, race, ethnicity, and an indicator for whether the student was overage for grade; teachers' race, gender, and age; and school location. Our benchmark approach also included district fixed effects to further increase statistical precision and weights that account for nonresponse and the probability of random assignment (Appendix G also contains information on the benchmark approach just described).

Estimating the effects of experience with the curricula involves assessing whether impacts of the curricula were larger after schools and teachers had one year of experience using the curricula. To estimate the effect of school experience, the study team used post-test data from the first and second cohorts of students to assess whether impacts on post-tests for the second cohort of students (whose schools all had one year of experience with the curricula) were larger than impacts on post-tests for the first cohort of students (whose schools, at that time, had no prior experience with the curricula). As summarized above (see text box), experience effects are estimated by interacting treatment variables with cohort indicator variables. In particular, the effect of school experience is estimated by interacting a cohort indicator variable with the treatment indicator variables. To determine whether the impact for the first cohort of students is statistically significantly different from the second cohort of students (which would be evidence of effects of school experience), we examine the statistical significance of the coefficients on these interaction variables.

To estimate the effect of teacher experience, the study team focused on post-test data from first and second cohort students whose teachers were in the study in both the first and second years to assess whether impacts on post-tests for the second group of students were larger than impacts for the first group. The first cohort of students was exposed to the interventions at a time when the study teachers had no prior experience with the curricula. The second cohort of students whose teachers participated in both years of the study was taught by teachers with a year of experience using the curricula. The effect of *teacher* experience is estimated by interacting a cohort variable with a teacher experience variable (which is a variable indicating whether the

[50]Our benchmark approach adjusts *p*-values *within* several domains of multiple tests (but not *across* domains). The two main impact tables in this chapter (Tables III.7 and III.8) each include eight domains. The first domain consists of 18 tests—the impact of each of three interventions (Project CRISS, ReadAbout, and Read for Real) on each of three outcome scores (GRADE, science comprehension, and social studies comprehension) for two cohorts. The second domain consists of six tests—the effect of each of the three interventions on a composite outcome for two cohorts. The third domain consists of six tests—the effect of the combined treatment group on each of three outcome measures for two cohorts. The fourth domain consists of two tests—the effect of the combined treatment group on the composite outcome for two cohorts. The fifth domain consists of nine tests – the differences in effects between Cohorts 1 and 2 of three interventions on three outcomes. The sixth domain consist of three tests—the difference in effects between Cohorts 1 and 2 of each intervention on a composite outcome. The seventh domain consists of three tests—the difference in effects between Cohorts 1 and 2 of the combined treatment group on three outcomes. The eighth domain consists of one test—the difference in effects between Cohorts 1 and 2 of the combined treatment group on the composite outcome.

teacher was in the study both years) and the treatment variables. To determine whether the impacts for the first and second cohorts of students with teachers in the study both years are statistically significantly different (which would be evidence of effects of teacher experience), we examine the statistical significance of the coefficients on these interaction variables.

As mentioned above, fewer Read for Real schools (11 of 16) agreed to continue participating in the study's second year relative to other study groups (15 of 17 Project CRISS schools, 15 of 17 ReadAbout schools, and 20 of 21 control group schools agreed to continue participating in the second year). The higher rate of attrition for Read for Real schools makes the interpretation of estimates of the effect of teacher experience with Read for Real more difficult because the schools that decided not to participate in the second year might be systematically different in unobserved ways than the full set of schools that participated in the study's first year.

As described in Chapter I, this component of the second year of the study included three of the four interventions that had been included in the first year of the study. Project CRISS, ReadAbout, and Read for Real were included in the impact analyses for the fifth-grade component of the second year of the study, but Reading for Knowledge was not because 9 of the 18 Reading for Knowledge schools elected not to continue implementing the intervention in the second year.

B. TREATMENT AND CONTROL GROUPS WERE SIMILAR AT BASELINE

Random assignment of schools yielded treatment and control groups that were similar at baseline. As mentioned in Chapter I, the baseline period for Cohort 1 students was in fall 2006 (the start of the first year of data collection) and the baseline period for Cohort 2 students was in fall 2007 (the start of the second year of data collection). We examined baseline differences for both cohorts of students. We conducted a total of 224 tests of differences in the baseline characteristics of students, teachers, and schools (including the core and supplemental reading curricula being used in study schools just prior to the start of the study) between each treatment group and the control group, and between the combined treatment group and the control group.[51] We found five differences between treatment groups and the control group: (1) fewer teachers in the ReadAbout group were female, (2) teachers in the ReadAbout group were younger, (3) teachers in the combined treatment group were younger, (4) more second cohort students in the Read for Real group were over-age for grade, and (5) fewer second cohort students in the Read for Real group were classified as ELL (see Tables III.1 through III.6).[52] The percentage of baseline differences that were statistically significant (2 percent) is less than what one would expect to occur by chance (5 percent).

---

[51]To be conservative in this analysis, we did *not* adjust *p*-values for multiple comparisons. Not adjusting for multiple comparisons is conservative in this case because an adjustment for multiple comparisons would reduce the probability of finding differences between the treatment and control groups.

[52]In addition to testing differences in school, teacher, and student characteristics, we tested whether the mean number of days between the baseline and follow-up tests differed between treatment and control groups. We did not find any statistically significant difference between the groups.

TABLE III.1

READING CURRICULA IN USE JUST BEFORE 2006–2007 SCHOOL YEAR

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Percentage of Schools That Report Using the Following Core Curriculum:[a]						
Textbook						
Most Commonly Reported Curricula[b] Fantastic Voyage,[c] Houghton Mifflin Reading,[d] Scott Foresman Reading 2000,[e] and Harcourt Trophies[f]	43	54 (0.51)	40 (0.87)	44 (0.96)	65 (0.19)	51 (0.53)
Other and None Reported[b]	57	46 (0.51)	60 (0.87)	56 (0.96)	35 (0.19)	49 (0.53)
Basal Reader Series						
Most Commonly Reported Curricula[b] Fantastic Voyage,[c] Houghton Mifflin Reading,[d] Scott Foresman Reading 2000,[e] and Harcourt Trophies[f]	40	71 (0.07)	47 (0.65)	50 (0.56)	58 (0.27)	56 (0.19)
Other and None Reported[b]	60	29 (0.07)	53 (0.65)	50 (0.56)	42 (0.27)	44 (0.19)
Special Program						
Most Commonly Reported Curricula[b] Accelerated Reader[g] and Reading Mastery[h]	25	24 (0.92)	23 (0.89)	30 (0.72)	41 (0.32)	29 (0.72)
Other	16	24 (0.54)	25 (0.49)	39 (0.12)	25 (0.49)	28 (0.25)
None Reported	59	53 (0.70)	52 (0.67)	31 (0.10)	35 (0.14)	43 (0.20)
Percentage of Schools That Report Using Supplemental Curricula in the Following Topic Areas:[i]						
Comprehension and Fluency[b]	—[j]	36 (0.09)	35 (0.11)	32 (0.14)	23 (0.34)	31 (0.10)
Vocabulary	15	30 (0.30)	23 (0.56)	25 (0.48)	29 (0.34)	26 (0.32)
Other and None Reported[b]	85	64 (0.17)	65 (0.19)	62 (0.13)	65 (0.19)	64 (0.10)
Number of Schools[k]	**21**	**17**	**17**	**16**	**18**	**68**

SOURCE: Preliminary School Information Form.

NOTE: The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. These data were collected during May-July 2006. The survey question that is the basis for this table asked principals to report what resources their school uses for its fifth-grade reading curriculum.

Table III.1 *(continued)*

[a]Columns may not sum to 100 percent due to rounding.

[b]Categories collapsed to protect school confidentiality.

[c]Schools reported using this curriculum on the study's Preliminary School Information Form. For additional information on this curriculum, please see the developer's website: http://www.pearsonschool.com/index.cfm?locator=PSZ1B7.

[d]Schools reported using this curriculum on the study's Preliminary School Information Form. For additional information on this curriculum, please see the developer's website: http://www.schooldirect.com.

[e]Schools reported using this curriculum on the study's Preliminary School Information Form. For additional information on this curriculum, please see the developer's website: <http://www.pearsonschool.com.

[f]Schools reported using this curriculum on the study's Preliminary School Information Form. For additional information on this curriculum, please see the developer's website: <https://jstore.harcourtschool.com.

[g]Schools reported using this curriculum on the study's Preliminary School Information Form. For additional information on this curriculum, please see the developer's website: <http://www.renlearn.com/ar/>.

[h]Schools reported using this curriculum on the study's Preliminary School Information Form. For additional information on this curriculum, please see the developer's website: <http://www.mcgraw-hill.co.uk/sra/readingmastery.htm>.

[i]Columns may not sum to 100 percent because schools could report using more than one supplemental curriculum.

[j]Value suppressed to protect school confidentiality.

[k]The number of schools presented in this row is the number participating in the study. One of the study schools did not fill out a Preliminary School Information Form.

TABLE III.2

BASELINE SCHOOL CHARACTERISTICS, BY TREATMENT AND CONTROL STATUS, YEAR 1

Baseline Characteristics	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Number of Students Enrolled in School	546.4	570.6 (0.77)	573.9 (0.68)	520.0 (0.63)	561.6 (0.83)	557.4 (0.84)
Number of Students Enrolled in Fifth Grade	74.0	87.3 (0.28)	76.1 (0.82)	71.4 (0.77)	80.3 (0.55)	78.7 (0.50)
Ethnicity/Race (Percentage)						
Hispanic	32	30 (0.94)	34 (0.81)	21 (0.82)	29 (0.76)	29 (0.61)
White	27	31 (0.94)	27 (0.81)	33 (0.82)	35 (0.76)	31 (0.61)
Black	38	37 (0.94)	36 (0.81)	43 (0.82)	34 (0.76)	37 (0.61)
Asian	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]
Native American	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]
Percentage of Students in School Eligible for Free or Reduced-Price Lunch	70.8	75.2 (0.48)	65.6 (0.48)	71.9 (0.89)	63.0 (0.29)	69.0 (0.75)
Percentage of Students in School Classified as English Language Learners	13.2	16.1 (0.65)	14.3 (0.84)	11.2 (0.68)	9.6 (0.43)	13.1 (0.98)
Percentage of Schools That Participated in Reading First in the 2005–2006 School Year	25	49 (0.15)	27 (0.86)	31 (0.71)	29 (0.80)	34 (0.45)
Percentage of Schools in the Following Locations:						
Urban	58	75 (0.27)	69 (0.47)	68 (0.54)	72 (0.37)	71 (0.25)
Urban fringe and rural area[b]	38	19 (0.27)	31 (0.47)	32 (0.54)	28 (0.37)	27 (0.25)
Percentage of Schools Eligible for Title I	95	100 (.)	100 (.)	94 (0.86)	89 (0.49)	96 (0.91)
Number of Schools[c]	**21**	**17**	**17**	**16**	**18**	**68**

SOURCE: Preliminary School Information Form; 2005–2006 Common Core of Data (CCD); School Information Form.

NOTE: Baseline for schools in the first year of the study was fall 2006. The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. *P-values* could not be obtained when all schools in one of the groups exhibited a given characteristic. This is indicated by a (.).

[a]Value suppressed to protect respondent confidentiality.

Table III.2 *(continued)*

[b]The urban fringe and rural area categories have been combined to protect respondent confidentiality.

[c]The number presented in this row is the number of schools participating in the first year of the study. The response rates for the calculations presented in the table vary from 67 to 100 percent, and the median response rate is 98 percent. The response rates vary because some schools did not report information on some of the items of the Preliminary School Information Form and the School Information Form.

TABLE III.3

BASELINE SCHOOL CHARACTERISTICS, BY TREATMENT AND CONTROL STATUS, SCHOOLS PARTICIPATING IN FIFTH-GRADE COMPONENT IN SECOND YEAR

Baseline Characteristics	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Number of Students Enrolled in School	553.4	622.2 (0.45)	582.2 (0.69)	524.3 (0.64)	578.0 (0.69)
Number of Students Enrolled in Fifth Grade	81.2	86.6 (0.70)	79.1 (0.85)	80.0 (0.92)	82.8 (0.88)
Ethnicity/Race (Percentage)					
Hispanic	39	38 (0.19)	41 (0.23)	37 (0.48)	41 (0.38)
White	25	30 (0.19)	24 (0.23)	25 (0.48)	27 (0.38)
Black	30	25 (0.19)	26 (0.23)	37 (0.48)	26 (0.38)
Asian	1	3 (0.19)	3 (0.23)	1 (0.48)	2 (0.38)
Native American	2	1 (0.19)	1 (0.23)	1 (0.48)	1 (0.38)
Percentage of Students in School Eligible for Free or Reduced-Price Lunch	70.9	71.8 (0.84)	65.5 (0.32)	77.1 (0.17)	71.1 (0.96)
Percentage of Students in School Classified as English Language Learners	14.7	16.5 (0.70)	16.0 (0.74)	12.1 (0.46)	16.0 (0.71)
Percentage of Schools That Participated in Reading First in the 2005–2006 School Year	27	54 (0.13)	27 (0.97)	45 (0.34)	41 (0.34)
Percentage of Schools in the Following Locations:					
Urban	53	72 (0.69)	75 (0.20)	83 (0.12)	75 (0.38)
Urban fringe	24	—[a]	—[a]	—[a]	16 (0.38)
Rural area	18	—[a]	—[a]	—[a]	7 (0.38)
Percentage of Schools Eligible for Title I	98	100 (.)	100 (.)	100 (.)	100 (.)
Number of Schools[b]	**20**	**15**	**15**	**11**	**41**

SOURCE: Preliminary School Information Form; 2005–2006 Common Core of Data (CCD); School Information Form.

NOTE: Baseline data for schools in the fifth-grade sample in Year 2 (all of whom were in the study in Year 1) are taken from fall 2006 (Year 1 of the study). The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. *P-values* could not be obtained when all schools in one of the groups exhibited a given characteristic. This is indicated by a (.).

[a]Value suppressed to protect respondent confidentiality.

[b]The number presented in this row is the number of schools participating in the second year of the study.

TABLE III.4

BASELINE TEACHER CHARACTERISTICS, BY TREATMENT AND CONTROL STATUS, YEAR 1

Baseline Characteristics	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Female (Percentage)	91	87 (0.58)	73* (0.04)	87 (0.59)	83 (0.24)	83 (0.19)
Age (Average)	45.1	41.5 (0.14)	39.9* (0.04)	40.1 (0.09)	41.7 (0.18)	40.8* (0.04)
Hispanic (Percentage)	18	14 (0.63)	15 (0.79)	16 (0.82)	13 (0.66)	15 (0.68)
Race (Percentage)						
White	83	64 (0.16)	84 (0.90)	71 (0.30)	74 (0.46)	73 (0.32)
Black	17	34 (0.16)	16 (0.90)	24 (0.30)	23 (0.46)	24 (0.32)
Asian	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]
Native American/Pacific Islander	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]
Teachers with a Master's Degree or Higher Degree (Percentage)	47	44 (0.76)	45 (0.89)	37 (0.36)	47 (0.99)	43 (0.67)
Years Teaching Experience (Average)	14.1	13.2 (0.71)	11.0 (0.12)	11.5 (0.32)	12.4 (0.39)	12.1 (0.20)
Number of Teachers[b]	**59**	**52**	**50**	**54**	**53**	**209**

SOURCE: Teacher Survey.

NOTE: Baseline for teachers in the first year of the study was fall 2006. The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. These tests account for clustering of teachers within schools.

[a]Value suppressed to protect teacher confidentiality.

[b]The number of teachers presented in this row is the number of fifth-grade teachers participating in the first year of the study. The response rates for the calculations presented in the table vary from 83 to 97 percent, and the median response rate is 91 percent. The response rates vary because some teachers did not report information on some items from the Teacher Survey.

*Statistically different from the control group at the .05 level.

TABLE III.5

BASELINE STUDENT CHARACTERISTICS, BY TREATMENT AND CONTROL STATUS, COHORT 1

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Female (Percentage)	48.0	52.0 (0.06)	51.0 (0.12)	49.0 (0.68)	48.0 (0.85)	50.0 (0.19)
Age (Average)	10.7	10.75 (0.43)	10.72 (0.80)	10.76 (0.36)	10.72 (0.67)	10.73 (0.48)
Overage (Percentage)[a]	21.0	23.0 (0.60)	23.0 (0.71)	25.0 (0.44)	23.0 (0.74)	23.0 (0.52)
Number of Days Absent in Prior School Year (Average)	12.1	10.5 (0.64)	11.4 (0.84)	14.3 (0.65)	11.3 (0.80)	11.8 (0.91)
Eligible for Free or Reduced-Price Lunch (Percentage)	60.0	59.0 (0.92)	61.0 (0.83)	58.0 (0.76)	58.0 (0.69)	59.0 (0.87)
Classified as English Language Learner (Percentage)	27.0	26.0 (0.93)	31.0 (0.70)	32.0 (0.72)	25.0 (0.85)	28.0 (0.82)
Identified as Having a Disability (Percentage)[b]	10.0	9.0 (0.69)	11.0 (0.79)	12.0 (0.60)	12.0 (0.52)	11.0 (0.77)
GRADE Score (Average)	100.0	100.7 (0.65)	99.6 (0.76)	99.2 (0.56)	101.0 (0.57)	100.1 (0.93)
TOSCRF Score (Average)	88.2	88.9 (0.53)	87.9 (0.70)	87.9 (0.68)	89.5 (0.27)	88.6 (0.65)
Number of Students[c]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

NOTE: Baseline for students in Cohort 1 was fall 2006. The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader in Cohort 1 to be overage for grade if he or she was 11 or older as of September 1, 2006.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number of Cohort 1 students participating in the study. The overall response rates for data items presented in the table vary from 69 to 96 percent, and the median response rate is 88 percent.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE III.6

BASELINE STUDENT CHARACTERISTICS, BY TREATMENT AND CONTROL STATUS, COHORT 2

	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Female (Percentage)	50.0	50.0 (0.78)	51.0 (0.47)	48.0 (0.73)	50.0 (0.69)
Age (Average)	10.7	10.7 (0.22)	10.7 (0.91)	10.8 (0.05)	10.7 (0.58)
Overage (Percentage)[a]	19.0	24.0 (0.15)	19.0 (0.93)	25.0* (0.04)	21.0 (0.46)
Number of Days Absent in Prior School Year (Average)	8.4	8.0 (0.74)	8.1 (0.82)	7.4 (0.36)	7.9 (0.68)
Eligible for Free or Reduced-Price Lunch (Percentage)	75.0	70.0 (0.47)	72.0 (0.65)	76.0 (0.92)	73.0 (0.68)
Classified as English Language Learner (Percentage)	18.0	17.0 (0.92)	16.0 (0.81)	4.0* (0.00)	16.0 (0.66)
Identified as Having a Disability (Percentage)[b]	11.0	10.0 (0.88)	10.0 (0.90)	16.0 (0.32)	11.0 (0.80)
GRADE Score (Average)	100. 4	101.6 (0.46)	100.5 (0.91)	100.2 (0.90)	100.7 (0.76)
TOSCRF Score (Average)	88.8	89.7 (0.56)	88.7 (0.91)	89.5 (0.70)	89.2 (0.79)
Number of Students[c]	**1,194**	**1,201**	**1,108**	**639**	**2,948**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

NOTE: Baseline for students in Cohort 2 was fall 2007. The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader in Cohort 2 to be overage for grade if he or she was 11 or older as of September 1, 2007.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number of Cohort 2 students participating in the study.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

While we would expect some chance differences between the treatment and control groups given the large number of tests conducted, we investigated the difference in teacher age to address the potential concern that it might indicate some systematic difference between the treatment and control groups. Specifically, we wanted to explore whether this difference might have arisen because older teachers refused to remain in the study after discovering that they were assigned to the treatment group. We examined the percentage of teachers who agreed to participate in the study and whether the difference in that percentage across the arms of the study was statistically significant. We found that 94 percent of the fifth-grade teachers in study schools agreed to participate and the difference in this percentage across the four treatment groups and the control group was not statistically significant. We included each of the variables for which statistically significant differences were observed at baseline as covariates in our benchmark impact analyses.

C. IMPACTS ON STUDENT TEST SCORES

Tables III.7 and III.8 present impact estimates for each intervention group separately as well as for the combined treatment group. For example, in the "Project CRISS" column, the estimates shown represent the regression-adjusted difference between scores of students in schools assigned to Project CRISS and scores of students assigned to the control group, while the "Combined Treatment Group" column shows the regression-adjusted difference between scores of students in schools assigned to any of the four intervention groups and scores of students assigned to the control group. When control group means are shown in report tables, they are the *regression-adjusted* control group means.

All of the analyses presented in this report focus on the *levels* of the outcome variables. The study team did not focus on *gain*s in the outcome variables from pre- to post-test (or from pre-test to follow up) because pre-test versions of the assessments were not administered for two of the study's three assessments administered at post-test and follow up.

School Experience Findings. There was no evidence that impacts were larger after schools had one year of experience using the curricula. Overall, we did not find any statistically significant impacts of the interventions on any of the three student post-test outcomes for the second cohort of fifth-grade students, whose schools had one prior year of experience using the curricula (Table III.7). This lack of statistically significant impacts was found in comparisons of Cohort 2 students in each intervention group with the control group and comparisons of the combined treatment group with the control group for the full sample of Cohort 2 students. There were also no statistically significant differences between the intervention group impacts (not shown in table). These findings provide evidence indicating that impacts of the three interventions are no larger after *schools* have had one year of experience using the curricula.

Teacher Experience Findings. Impacts for one of the curricula were statistically significantly larger after teachers had one year of experience using the curricula. We found one positive, statistically significant impact among students in the second cohort who were taught by *teachers* in the study for two years (Table III.8). (Impacts are reported as "effect sizes" to facilitate comparisons of impacts on different outcomes. The effect size is the impact divided by the standard deviation of the outcome for students in the control group. For example, an impact of 4 units on an outcome with a standard deviation of 20 would be reported as an effect size of

TABLE III.7

DIFFERENCES IN POST-TEST SCORES BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING FIFTH-GRADE COHORTS 1 AND 2

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Composite Test Score[a]					
Cohort 1 Students (Spring 2007)					
Impact	0.0	-0.01	-0.04	-0.07	-0.04
Effect Size		-0.01	-0.04	-0.08	-0.05
p-value		1	0.93	0.39	0.30
Cohort 2 Students (Spring 2008)					
Impact	0.0	0	0.05	-0.02	0.02
Effect Size		0	0.06	-0.02	0.02
p-value		1	0.81	1	0.85
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		0.01	0.09	0.05	0.06
Difference in Effect Size		0.01	0.10	0.06	0.08
p-value for the Difference		1	0.38	0.66	0.21
GRADE Score					
Cohort 1 Students (Spring 2007)					
Impact	100.6	-0.19	-0.64	-0.76	-0.60
Effect Size		-0.01	-0.05	-0.06	-0.04
p-value		1	0.99	0.92	0.65
Cohort 2 Students (Spring 2008)					
Impact	100.8	-0.28	-0.08	-0.56	-0.26
Effect Size		-0.02	-0.01	-0.04	-0.02
p-value		1	1	1	0.99
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		-0.09	0.56	0.20	0.34
Difference in Effect Size		-0.01	0.04	0.01	0.02
p-value for the Difference		1	0.99	1	0.94
Social Studies Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	500.5	-1.36	-0.38	-2.28	-1.18
Effect Size		-0.05	-0.01	-0.08	-0.04
p-value		1	1	0.75	0.81
Cohort 2 Students (Spring 2008)					
Impact	500.0	0.09	4.63	0.47	2.21
Effect Size		0	0.16	0.02	0.07
p-value		1	0.05	1	0.46
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		1.45	5.01	2.75	3.39
Difference in Effect Size		0.05	0.17	0.09	0.11
p-value for the Difference		1	0.12	0.85	0.15

TABLE III.7 (*continued*)

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Science Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	500.7	0.31	-1.07	-2.71	-1.38
Effect Size		0.01	-0.04	-0.10	-0.05
p-value		1	1	0.85	0.84
Cohort 2 Students (Spring 2008)					
Impact	501.7	0.58	1.66	-0.31	0.83
Effect Size		0.02	0.06	-0.01	0.03
p-value		1	1	1	0.99
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		0.27	2.73	2.41	2.21
Difference in Effect Size		0.01	0.10	0.09	0.08
p-value for the Difference		1	0.89	0.97	0.62
Number of Students in Cohort 1[b]	**1,362**	**1,319**	**1,245**	**1,228**	**3,792**
Number of Students in Cohort 2[c]	**1,194**	**1,201**	**1,108**	**639**	**2,948**

SOURCE: Reading comprehension tests administered by study team.

NOTE: For each outcome, the numbers reported in the column labeled "Control Group Mean" are the average predicted outcomes for all students as if they were in the control group. The numbers reported in the remaining columns are, by row: (1) the impact, (2) the effect size, and (3) the *p-value* of the impact. The *p-values* presented in this table are adjusted for multiple-hypotheses testing. Unadjusted *p-values* are presented in Appendix K. For each outcome, the differences between cohort impacts are also reported. The social studies and science reading comprehension assessments were developed by ETS. Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher sex, teacher age, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]The number of students presented in this row is the number of Cohort 1 students participating in the study. The proportion of students in each experimental condition with follow-up test scores is reported in Appendix G.

[c]The number of students presented in this row is the number of Cohort 2 students participating in the study. The proportion of students in each experimental condition with follow-up test scores is reported in Appendix G.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE III.8

DIFFERENCES IN POST-TEST SCORES BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING FIFTH-GRADE COHORT 1 AND 2 STUDENTS WITH TEACHERS IN THE STUDY FOR TWO CONSECUTIVE YEARS

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Composite Test Score[a]					
Cohort 1 Students (Spring 2007)					
Impact	0.06	-0.06	-0.06	-0.09	-0.08
Effect Size		-0.07	-0.08	-0.10	-0.09
p-value		0.87	0.70	0.38	0.15
Cohort 2 Students (Spring 2008)					
Impact	-0.04	0.02	0.09	0.03	0.05
Effect Size		0.03	0.10	0.03	0.06
p-value		1	0.39	0.99	0.41
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		0.08	0.15	0.11	0.13
Difference in Effect Size		0.10	0.18	0.14	0.15
p-value for the Difference		0.83	0.12	0.34	0.07
GRADE Score					
Cohort 1 Students (Spring 2007)					
Impact	101.4	-1.07	-0.94	-1.48	-1.14
Effect Size		-0.08	-0.07	-0.11	-0.08
p-value		0.90	0.92	0.56	0.23
Cohort 2 Students (Spring 2008)					
Impact	100.1	0.16	0.24	-0.21	0.08
Effect Size		0.01	0.02	-0.02	0.01
p-value		1	1	1	1
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		1.23	1.17	1.27	1.23
Difference in Effect Size		0.09	0.09	0.09	0.09
p-value for the Difference		0.97	0.93	0.86	0.49
Social Studies Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	502.6	-3.53	-1.78	-1.56	-2.09
Effect Size		-0.12	-0.06	-0.05	-0.07
p-value		0.89	0.96	1	0.56
Cohort 2 Students (Spring 2008)					
Impact	500.0	0.27	6.43*	3.03	3.25
Effect Size		0.01	0.22	0.10	0.11
p-value		1	0.01	0.88	0.29
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		3.80	8.21*	4.59	5.34
Difference in Effect Size		0.13	0.28	0.15	0.18
p-value for the Difference		0.99	0.01	0.84	0.12

TABLE III.8 (*continued*)

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Science Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	503.4	-0.07	-2.11	-3.04	-1.76
Effect Size		0	-0.08	-0.11	-0.06
p-value		1	1	0.72	0.84
Cohort 2 Students (Spring 2008)					
Impact	501.7	2.22	2.91	1.87	2.35
Effect Size		0.08	0.10	0.07	0.08
p-value		0.98	0.98	1	0.69
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		2.29	5.02	4.91	4.10
Difference in Effect Size		0.08	0.18	0.18	0.15
p-value for the Difference		1	0.70	0.66	0.45
Number of Students with Teachers in Study for Two Years[b]					
Cohort 1	**933**	**845**	**902**	**487**	**2,234**
Cohort 2	**949**	**775**	**815**	**478**	**2,068**

SOURCE: Reading comprehension tests administered by study team.

NOTE: For each outcome, the numbers reported in the column labeled "Control Group Mean" are the average predicted outcomes for all students as if they were in the control group. The numbers reported in the remaining columns are, by row: (1) the impact, (2) the effect size, and (3) the *p-value* of the impact. The *p-values* presented in this table are adjusted for multiple-hypotheses testing. For each outcome, the differences between cohort impacts are also reported. The social studies and science reading comprehension assessments were developed by ETS. Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher sex, teacher age, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]Counts reflect the number of students with nonmissing teacher data.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

0.20.) There was a positive, statistically significant impact of ReadAbout on the social studies reading comprehension post-test assessment (effect size: 0.22). To put this in perspective, the average gain in GRADE scores among students in the control group between pre-test and post-test was 0.44 standard deviations over a period of 245 calendar days.[53] The full school year is about 270 calendar days. Assuming a constant rate of achievement gain over time, a 0.22 standard deviation gain is equivalent to a gain of more than a third of a school year (0.22/(0.44*270/245) = 0.45). To provide additional context, for a student at the 50th percentile, an effect size of 0.10 represents about 4 percentile points, an effect size of 0.15 represents about 6 percentile points, and an effect size of 0.20 represents about 8 percentile points. A meta-analysis by Rosenshine and Meister (1994) provides additional perspective; they found an average effect size of 0.32 across nine studies examining the impact of multiple reading comprehension strategy instruction on standardized test scores (this meta-analysis focused on reciprocal teaching, which involves the use of guided practice and dialogue between students and teachers to teach students about four comprehension strategies including question generation, summarization, prediction, and clarification). Another meta-analysis by Rosenshine, Meister, and Chapman (1996) found an average effect size of 0.36 across 13 studies examining the impact of question generation on standardized test scores.

The impact of ReadAbout on the social studies reading comprehension post-test assessment for the second cohort of students was statistically significantly greater than the impact of ReadAbout on this outcome for the first cohort of students taught by the same teachers in the first year of the study (effect size difference: 0.28). ReadAbout's impacts on the other post-test assessments (GRADE and science comprehension) were not statistically significant. There were also no statistically significant differences between the intervention group impacts (not shown in table). These findings provide some evidence indicating that impacts of ReadAbout (for one of three assessments) are statistically significantly larger after teachers have had one year of experience using the curriculum.[54]

Sensitivity Tests to Assess the Robustness of the Impact Findings. We assessed the robustness of these impacts through the following sensitivity tests (see Appendix G for more information): (1) excluding covariates, (2) using an alternative weighting approach, and (3) focusing only on students with both a pre- and a post-test. None of the findings presented above were sensitive to these changes in estimation approach. Specifically, all of the school experience findings remained statistically insignificant and the statistically significant impact of ReadAbout found in the teacher experience analysis remained statistically significant in each of these sensitivity analyses.

---

[53]A similar comparison cannot be made using the ETS test because the ETS test was not administered at baseline.

[54]The impact of ReadAbout on social studies scores after *teachers* had one year of experience with the curricula (effect size: 0.22) is similar to the impact observed after *schools* had one year of experience with the curricula (effect size: 0.16), but the latter impact was not statistically significant (*p*-value: .053).

D. SIX OF 288 DIFFERENCES IN STUDENT SUBGROUP IMPACTS ARE STATISTICALLY SIGNIFICANT

The study team also conducted a series of subgroup analyses to investigate secondary research questions related to whether impacts of the interventions might vary for second cohort students with different characteristics. Most of these subgroups are formed using characteristics observed at the beginning of each implementation year (fall 2006 for Cohort 1 and fall 2007 for Cohort 2), so the analyses preserve the properties of random assignment because the intervention could not have influenced these characteristics and thus there should be no systematic differences in unobserved characteristics of students in these subgroups between the treatment and control groups. Consequently, most of these findings allow for causal conclusions to be drawn about the impact of the interventions for these subgroups. The three exceptions are the subgroups defined by teachers' self-reported past professional development, teaching efficacy, and school professional culture (all of which are based on data collected through the Teacher Survey, which was administered by the study team in August through November 2006, at the start of the study's first year of data collection[55]). Both the number and composition of teachers in the treatment group who reported receiving past professional development and who reported a given level of teacher efficacy or school professional culture could have been affected by the product-specific training received in the summer before the first implementation year (in particular, teachers may have reported the training as professional development, and the training may have affected teachers' responses to survey questions on their teaching efficacy and the professional culture in their schools). Because this potential shift in the size and composition of these subgroups affected only the treatment group but not the control group, analyses of these subgroups do *not* maintain the properties of random assignment and, therefore, do *not* allow for causal conclusions to be drawn about the impact of the interventions for these subgroups.

Our main approach to creating subgroups was to split the student sample into two groups of roughly equal size at the median level of each relevant characteristic *for the study sample*. For the subgroups based on baseline student test scores, we used a different approach, in which the two subgroups were created in five different ways: (1) by splitting the sample at the average score on the GRADE and TOSCRF tests *for the norm sample,* (2) by splitting the sample at the median score on the GRADE and TOSCRF tests *for the study sample,* (3) by comparing students in the top and bottom thirds of the GRADE and TOSCRF distributions, (4) by comparing students in the middle and bottom thirds of the GRADE and TOSCRF distributions, and (5) by comparing students in the top and middle thirds of the GRADE and TOSCRF distributions.[56] For the subgroups based on teacher experience, we used an approach in which the two subgroups

[55]Teacher surveys were not administered to teachers new to the study sample in fall 2007; therefore, new teachers were not included in these teacher subgroup analyses.

[56]For both the GRADE and TOSCRF, the average score for the norm sample was 100. The median values *for our study sample* were 100.5 for the GRADE and 89 for the TOSCRF.

were created in two ways: (1) by splitting the sample at the sample median (11 years) and (2) by splitting the sample at 5 years.[57] Three types of student subgroups were created, as follows:

1. ***Subgroups of students based on characteristics of the students themselves:*** fluency (baseline TOSCRF), comprehension (baseline GRADE), and English language learner (ELL) status. These subgroups were selected because they may be observed by teachers and could be used as the basis for targeting the interventions to specific students (for example, if it is found that students with below-average fluency levels respond better to a particular intervention).

2. ***Subgroups of students based on characteristics of their teachers:*** teachers' years of experience, hours of professional development in past 12 months, and self-reported efficacy. These subgroups were selected because they are characteristics that might be used by teachers and principals to target interventions to specific circumstances (for example, certain interventions might be more effective for teachers with below-average years of experience).

3. ***Subgroups of students based on conditions of the schools they attend:*** professional culture in the school, concentration of students eligible for free or reduced-price lunch, and concentration of ELL students in the school. These subgroups were selected because they are conditions that might be used by principals to target interventions to specific settings (for example, certain interventions might be more effective in schools with above-average concentrations of English language learners).

We report subgroup impacts based on the difference in impacts between subgroups among students in the second cohort (for example, the difference in impacts between ELL and non-ELL students in the second cohort). These differences are reported in Appendix Tables L.1-L.4 (with adjustments for multiple hypothesis testing) and L.9-L.12 (without adjustments for multiple hypothesis testing). Below, we focus on the findings that are statistically significant with adjustments for multiple hypothesis testing.[58]

There was no clear pattern to the six statistically significant subgroup differences observed in the second year (288 subgroup differences were examined). Greater impacts were observed of:

---

[57]We examined a five-year teacher experience cut-point (in addition to using the sample median as a cut-point), because Ingersoll (2002) found that as many as 39 percent of teachers leave teaching altogether in the first five years of their careers.

[58]These adjustments are conducted in four domains for each subgroup (we do not adjust for multiple comparisons *between* subgroups, only *within* subgroups). The first domain consists of nine tests—the test described above for each of three interventions (Project CRISS, ReadAbout, and Read for Real) on each of three outcome scores (GRADE, science comprehension, and social studies comprehension). The second domain consists of three tests—the test described above for each intervention on a composite outcome. The third domain consists of three tests—the test described above for the combined treatment group on each of three outcome measures. The fourth domain consists of one test— the test described above for the combined treatment group on the composite outcome.

- Project CRISS on the composite test score for second-cohort students who scored in the top third of the pre-test GRADE distribution (effect size: 0.06) than those who scored in the middle third (effect size: -0.11) (Table L.1)

- Read for Real on the composite test score for second-cohort students classified as ELL (effect size: 0.46) than for students not classified as ELL (effect size: -0.08) (Table L.1)

- The combined treatment group on the composite test score for second-cohort students who were taught by teachers below the median of teacher-reported efficacy (effect size: 0.14) than for students taught by teachers above the median (effect size: -0.03) (Table L.1)

- ReadAbout on the social studies reading comprehension test for second-cohort students who were in schools where the professional culture scale was below the median (effect size: 0.45) than for students in schools where the professional culture scale was above the median (effect size: 0.04) (Table L.3)

- The combined treatment group on the social studies reading comprehension test for second-cohort students who were in schools where the professional culture scale was below the median (effect size: 0.27) than for students in schools where the professional culture scale was above the median (effect size: -0.01) (Table L.3)

- Read for Real on the science reading comprehension test score for second-cohort students who were taught by teachers below the median of teacher-reported efficacy (effect size: 0.27) than for students who were taught by teachers above the median (effect size: -0.26) (Table L.4)

E. NONE OF THE TEACHER PRACTICES SUBGROUP DIFFERENCES ARE STATISTICALLY SIGNIFICANT

As a secondary analysis, we also investigated the relationship between intervention impacts and classroom practices (see Chapter II for more information on the three teacher practice scales the study team constructed). [59] We did this by conducting analyses of post-test scores for students in classrooms with different levels of observed teaching practices (as with the subgroup analyses described above, we split the sample at the median levels of teacher practices observed). These relationships must be interpreted cautiously because the interventions may have affected the extent to which treatment teachers engage in specific practices or the types of treatment teachers who choose to engage in those practices. As a result, treatment and control teachers who engage in teaching practices to the same degree may differ in unmeasurable ways.[60] Therefore,

[59]See Appendix Figures F.1A through F.3 for information on how the frequency of specific teacher practices corresponds to different scale scores.

[60]If the intervention affected teacher practices, then that impact on teacher practices might explain the overall impact on student test scores. However, it is not possible to make causal statements about that relationship (causal statements would require a different study design than the one we used on this study, such as one in which teachers or schools were randomly assigned to implement the interventions to different degrees or amounts).

these estimates of the relationship between intervention impacts and teacher practices *cannot* be interpreted as providing rigorous impact estimates and do *not* allow causal conclusions to be drawn about the impact of the interventions for these subgroups.

We report teacher subgroup effects based on the same types of subgroup differences described in Section D, using the same approach to adjusting for multiple comparisons. The findings are reported in Appendix Tables L.1-L.4 and L.9-L.12. There were no statistically significant differences in these analyses.

IV. COMPARING POST-TEST AND FOLLOW-UP IMPACTS FOR THE FIRST COHORT OF FIFTH-GRADE STUDENTS

Similar to the analyses presented in Chapter III, the analysis of impacts of the interventions on follow-up test scores of the first cohort of students was designed to answer primary and secondary research questions. The primary research question focuses on whether the reading comprehension interventions had an impact on outcomes of first cohort students roughly one year after the implementation of the interventions ended. There are two main reasons for examining this research question: (1) it is possible that impacts of the interventions could emerge in the second year even after the intervention implementation has ended and (2) to examine whether the negative effects of Reading for Knowledge observed in the first year continued into the second year. To facilitate this examination, this analysis also investigates whether the impacts of the reading comprehension interventions in the second year of the study (when students from the first cohort were in sixth grade and no longer receiving the interventions) were different than the impacts in the first year of the study (when first cohort students were in fifth grade and receiving the interventions). The secondary research questions center on for whom and under what conditions the reading comprehension interventions have impacts one year after the end of implementation. Similar to the analysis of impacts for the second cohort of fifth-grade students presented in Chapter III, the answer to the primary question presented in this chapter is expected to be of most interest to policymakers since it indicates whether the interventions have an impact on the first cohort of students one year after the end of the implementation of the interventions. Answering secondary questions can be helpful in explaining the results from the primary analyses and shaping an agenda for future research on reading comprehension interventions.

EXAMINING IMPACTS ONE YEAR AFTER THE END OF THE INTERVENTION IMPLEMENTATION

- One primary research question:
  - Do the curricula have impacts on students one year after the end of the intervention implementation?
- Estimating impacts one year after the end of the intervention implementation:
  - The impact of the interventions at follow up (one year after the end of the intervention implementation for first cohort students, which is the second study year) is estimated by regressing follow-up test scores on treatment variables and other covariates.
  - The model accounts for clustering of students within schools, uses covariate adjustment to improve statistical precision, adjusts *p*-values for multiple comparisons, and uses weights that account for random assignment probabilities and nonresponse.
- Summary of findings:
  - No statistically significant impacts of the interventions at follow up for first cohort students.

Information about the methods used to estimate impacts of the interventions one year after the end of the implementation of the interventions is presented in Section A. Section B examines the comparability of the experiences of treatment and control group students from the first cohort in the second year of the study. Section C addresses the primary questions on impacts of the interventions one year after the end of the intervention implementation and Sections D and E focus on the secondary questions described above. Specifically, Section C presents impacts on follow-up test scores for Cohort 1 students, as well as information on whether the follow-up impacts differ from post-test impacts (which were measured at the end of the first year of the study). Section D presents impacts for subgroups of sixth-grade students from the first cohort, based on the characteristics of the students and their teachers, and on the conditions in their schools. Section E investigates whether, and, if so, how, follow-up impacts are related to differences in teachers' classroom practices.

A. METHODS FOR ESTIMATING FOLLOW-UP IMPACTS

Similar to the impacts presented in Chapter III, the impacts presented in this chapter are based on our "benchmark" approach. This approach involves (1) accounting for clustering of students within schools to account for the correlation between students in the same schools, (2) adjusting the results from statistical tests (*p*-values) for multiple comparisons since there are multiple outcomes, multiple treatment groups being compared to a single control group, and two assessments (post-test and follow up), (3) controlling for district fixed effects and student, teacher, and school characteristics to increase statistical precision of the impact estimates, and (4) including weights that account for nonresponse and the probability of random assignment. Chapter III and Appendix G also contain information on the benchmark approach just described. Our benchmark approach to estimate impacts of the interventions one year after the end of the intervention implementation, adjusts *p*-values within (not across) several domains of multiple tests.[61]

As in Chapter III, this chapter presents two types of impacts. First, impacts at follow up for each intervention are presented to provide information on the effects of each intervention one year after the end of the intervention implementation. Second, impacts at follow-up are presented for the combined treatment group, based on outcomes at follow up of students in all four intervention groups and outcomes of students in the control group.

---

[61]The main impact table in this chapter (Table IV.3) includes eight domains. The first domain consists of 24 tests—the impact of each of four interventions (Project CRISS, ReadAbout, Read for Real, and Reading for Knowledge) on each of three outcome scores (GRADE, science comprehension, and social studies comprehension) for two assessments (post-test and follow up). The second domain consists of eight tests—the effect of each of the four interventions on a composite outcome for two assessments. The third domain consists of six tests—the effect of the combined treatment group on each of three outcome measures for two assessments. The fourth domain consists of two tests—the effect of the combined treatment group on the composite outcome for two assessments. The fifth domain consists of 12 tests – the differences in effects between follow up and post-test of four interventions on three outcomes. The sixth domain consists of four tests—the difference in effects between follow up and post-test of each intervention on a composite outcome. The seventh domain consists of three tests—the difference in effects between follow up and post-test of the combined treatment group on three outcomes. The eighth domain consists of one test—the difference in effects between follow up and post-test of the combined treatment group on the composite outcome.

All four interventions (including Reading for Knowledge) were included in the impact analyses for the sixth-grade component of the second year of the study, as this component involved tracking the first cohort of students through the end of the 2007-2008 school year and administering follow-up tests (unlike the fifth-grade component, this component did not require the interventions to be implemented by sixth-grade teachers).

Students in the study's sixth-grade component were classified according to their treatment status from the study's first year. For example, students who attended Read for Real schools in the study's first year are in the Read for Real group in the analyses for the study's sixth-grade component, regardless of the school they attended in the study's second year. Likewise, students who attended control schools in the study's first year are in the control group for the analyses of the study's sixth-grade component. This allows the study team to assess the longer-term effectiveness of the single year of curricula implementation provided to students in the first year of the study.

In the second year of the study, the only way in which Cohort 1 sixth-grade students could have a teacher who received training in one of the study interventions is if one of the fifth-grade treatment group teachers who was trained in the first year of the study became a sixth-grade teacher. We found that this occurred to a very limited extent (affecting 1 percent of Cohort 1 students). (In addition, just one of these students was both (1) in the control group in Year 1 and (2) enrolled in a classroom in sixth grade in which the teacher had been in a treatment group school in the first year.) Note that for these Cohort 1 sixth-grade students who were enrolled in a sixth-grade classroom with a teacher who had been trained in the use of one of the study interventions in the prior year, the interventions were *not* implemented in sixth grade and intervention materials were *not* provided to sixth-grade classrooms.

B. EXPERIENCES OF THE FIRST COHORT OF TREATMENT AND CONTROL STUDENTS WERE SIMILAR DURING THE SECOND YEAR OF THE STUDY

In the second year of the study, first cohort students attended sixth grade in 252 schools.[62] This is a larger number of schools than the 89 that participated in the first study year, as many students went on to new schools when they entered sixth grade. Because of the way in which multiple elementary schools fed into a single middle school serving sixth-grade students, first-cohort students from the treatment group could attend school (and even be in the same class) with first-cohort students from the control group in sixth grade.

In this chapter, we report on the experiences of sixth graders from the first cohort.[63] In the first year of the study, one would not expect to observe any systematic differences in the experiences of first cohort students in the treatment and control groups due to random

[62]58 of the 252 schools were schools that were randomly assigned to one of the interventions or to the control group at the beginning of the study. These 58 schools served sixth-grade students, so they are included in this component of the study.

[63]Baseline characteristics of first cohort students and their teachers and schools in the study's first year are discussed in Chapter III (see Tables III.1, III.2, III.4, and III.5).

assignment. In the second year of the study, however, now that the treatment and control students from the first cohort have spread out across a larger number of schools to attend sixth grade and treatment students could attend school (and be in the same class) with control students during sixth grade, the experiences of those groups in sixth grade could by chance be different.[64] For this reason, we conducted analyses specifically designed to assess the similarity of the experiences of the treatment and control groups in sixth grade. In particular, we examined school and teacher characteristics and conducted a total of 60 tests of differences in those characteristics between each treatment group and the control group, and between the combined treatment group and the control group.[65]

We did not find any statistically significant differences in the characteristics of schools attended by Cohort 1 treatment and control students in the second year of the study (Table IV.1). In addition, we did not find any statistically significant differences in the characteristics of the teachers who taught Cohort 1 treatment and control students in the second year of the study (Table IV.2). This is an important finding, as it suggests that any observed differences between the treatment and control groups at follow up are the result of the implementation of the interventions in fifth grade, not due to differences in the experiences of students in the treatment and control groups in sixth grade.

C. IMPACTS ON FOLLOW-UP TEST SCORES OF SIXTH-GRADE COHORT 1 STUDENTS

Table IV.3 presents impact estimates at post-test (first year of study) and follow-up (second year of study) for each intervention group separately as well as for the combined treatment group of first cohort students. For example, in the "Follow up" panes and the "ReadAbout" column, the estimates shown represent the regression-adjusted difference between follow-up scores of students who, in the first year of the study, attended schools assigned to ReadAbout and follow-up scores of students who, in the first year of the study, attended schools assigned to the control group. The control group means shown in Table IV.3 are *regression-adjusted* control group means.

---

[64]Recall that the design of the sixth-grade component of the study did not call for implementing the interventions in sixth grade. Rather, it called for following fifth-grade students for one additional year (through the end of sixth grade) to examine the longer-term effects of the interventions that were implemented in fifth grade. Therefore, there was no intention to attempt to control the schools in which students would enroll in sixth grade. Students attended the schools they would have attended in the absence of the study. Because students were not randomly assigned to schools in sixth grade, it was important to examine the experiences of treatment and control students in sixth grade to assess whether they were similar. If they were *not* similar, one might be concerned that the impacts on sixth-grade students reflect not only the intervention implementation in fifth grade but also differences in students' experiences in sixth grade. If the experiences of the groups were similar in sixth grade (as was found in this study), one would have more confidence that the impacts on sixth-grade students reflect only the intervention implementation in fifth grade.

[65]Similar to the baseline analysis presented in Chapter III, we did not adjust *p*-values for multiple comparisons to be conservative in the analysis of experiences of sixth graders in the second year of the study. Not adjusting for multiple comparisons is conservative in this case because an adjustment for multiple comparisons would reduce the probability of finding differences between treatment and control groups.

TABLE IV.1

CHARACTERISTICS OF SCHOOLS ATTENDED BY SIXTH-GRADE STUDENTS IN YEAR 2, BY TREATMENT AND CONTROL STATUS OF STUDENTS

School Characteristics	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Number of Students Enrolled in School	742.2	768.2 (0.76)	758.0 (0.83)	741.0 (0.99)	737.8 (0.95)	751.6 (0.87)
Number of Students Enrolled in Sixth Grade	179. 9	208.2 (0.46)	185.9 (0.87)	220.3 (0.25)	185.8 (0.86)	199.3 (0.48)
Ethnicity/Race (Percentage)						
Hispanic	39.0	33.0 (0.84)	40.0 (0.90)	32.0 (0.75)	37.0 (0.81)	36.0 (0.80)
White	27.0	34.0 (0.84)	27.0 (0.90)	34.0 (0.75)	31.0 (0.81)	31.0 (0.80)
Black	29.0	30.0 (0.84)	28.0 (0.90)	29.0 (0.75)	28.0 (0.81)	28.0 (0.80)
Asian	2.0	2.0 (0.84)	2.0 (0.90)	2.0 (0.75)	2.0 (0.81)	2.0 (0.80)
Native American	1.0	1.0 (0.84)	1.0 (0.90)	1.0 (0.75)	1.0 (0.81)	1.0 (0.80)
Percentage of Students in School Eligible for Free or Reduced-Price Lunch	65.3	68.5 (0.46)	64.4 (0.82)	63.6 (0.72)	65.4 (0.99)	65.5 (0.96)
Percentage of Schools in the Following Locations:						
Urban	58.0	71.0 (0.21)	65.0 (0.76)	52.0 (0.25)	71.0 (0.43)	65.0 (0.78)
Urban fringe	24.0	15.0 (0.21)	16.0 (0.76)	19.0 (0.25)	15.0 (0.43)	16.0 (0.78)
Rural area	15.0	6.0 (0.21)	11.0 (0.76)	29.0 (0.25)	6.0 (0.43)	13.0 (0.78)
Percentage of Schools Eligible for Title I	89.0	92.0 (0.64)	91.0 (0.84)	84.0 (0.59)	88.0 (0.90)	89.0 (0.98)
Number of Students[a]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**

SOURCE: 2005–2006 Common Core of Data (CCD).

NOTE: The numbers reported in this table represent the experiences of the average student from each treatment and control group. Analyses were conducted at the student level because students from treatment and control groups can attend the same school in sixth grade. The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. These tests account for clustering of students within the schools in which students were enrolled at the time of random assignment.

[a]The number of students presented in this row is the number of Cohort 1 students participating in the study. The overall response rates for data items presented in the table vary from 87 to 91 percent, and the median response rate is 89 percent. The number of schools is not reported by treatment and control groups because sixth-grade students from more than one group can attend the same school.

TABLE IV.2

CHARACTERISTICS OF TEACHERS WHO TAUGHT SIXTH-GRADE STUDENTS IN YEAR 2, BY TREATMENT AND CONTROL STATUS OF STUDENTS

Teacher Characteristics	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Female (Percentage)	86.0	72.0 (0.16)	78.0 (0.35)	79.0 (0.34)	75.0 (0.23)	76.0 (0.17)
Age (Average)	41.0	45.4 (0.10)	43.8 (0.30)	42.3 (0.61)	44.1 (0.26)	43.8 (0.21)
Hispanic (Percentage)	17.0	10.0 (0.47)	11.0 (0.46)	20.0 (0.80)	7.0 (0.06)	12.0 (0.41)
Race (Percentage)						
White	70.0	77.0 (0.39)	60.0 (0.55)	64.0 (0.09)	71.0 (0.78)	69.0 (0.98)
Black	23.0	21.0 (0.39)	36.0 (0.55)	22.0 (0.09)	24.0 (0.78)	25.0 (0.98)
Asian	3.0	3.0 (0.39)	5.0 (0.55)	0.0 (0.09)	0.0 (0.78)	2.0 (0.98)
Native American/Pacific Islander	0.0	0.0 (0.39)	0.0 (0.55)	0.0 (0.09)	3.0 (0.78)	1.0 (0.98)
Teachers with a Master's Degree or Higher Degree (Percentage)	48.0	39.0 (0.52)	66.0 (0.20)	55.0 (0.53)	51.0 (0.80)	52.0 (0.69)
Years Teaching Experience (Average)	11.2	13.3 (0.28)	14.4 (0.10)	12.7 (0.41)	13.5 (0.16)	13.4 (0.07)
Number of Students[a]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**

SOURCE: Teacher Survey administered to sixth-grade teachers in fall 2007.

NOTE: The numbers reported in this table represent the experiences of the average student from each treatment and control group. Analyses were conducted at the student level because students from treatment and control groups can attend the same school in sixth grade (and, therefore, can have the same teachers in sixth grade). The treatment and control group means presented in this table are weighted means. The weight is determined by random assignment probabilities, which were unequal when the number of schools in a district was not evenly divisible by 5. The *p-values* from statistical tests of differences in treatment and control group weighted means are presented in parentheses. These tests account for clustering of students within the schools in which students were enrolled at the time of random assignment.

[a]The number of students presented in this row is the number of Cohort 1 students participating in the study. The overall response rates for data items presented in the table vary from 51 to 62 percent, and the median response rate is 55 percent. The number of teachers is not reported by treatment and control groups because the same sixth-grade teacher can teach students from more than one group.

TABLE IV.3

DIFFERENCES IN POST-TEST AND FOLLOW-UP TEST SCORES BETWEEN TREATMENT AND CONTROL GROUPS, COHORT 1 STUDENTS

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Composite Test Score[a]						
Post-Test (Spring 2007)						
Impact	0.01	-0.01	-0.03	-0.06	-0.11*	-0.07*
Effect Size		-0.02	-0.04	-0.06	-0.13	-0.08
p-value		1	0.99	0.81	0.04	0.03
Follow Up (Spring 2008)						
Impact	-0.06	-0.01	0	0.06	0.05	0.02
Effect Size		-0.01	0	0.07	0.06	0.03
p-value		1	1	0.63	0.77	0.61
Difference Between Post-Test and Follow Up						
Difference in Impact		0.01	0.03	0.12	0.17*	0.09*
Difference in Effect Size		0.01	0.04	0.13	0.18	0.1
p-value for the Difference		1	0.94	0.12	0.02	0.03
GRADE Score						
Post-Test (Spring 2007)						
Impact	100.96	-0.44	-0.68	-0.74	-1.45	-1.01
Effect Size		-0.03	-0.05	-0.05	-0.11	-0.07
p-value		1	1	0.98	0.31	0.09
Follow Up (Spring 2008)						
Impact	96.04	-0.75	-0.14	0.52	0.31	-0.04
Effect Size		-0.05	-0.01	0.04	0.02	0
p-value		0.97	1	1	1	1
Difference Between Post-Test and Follow Up						
Difference in Impact		-0.31	0.54	1.25	1.76	0.97
Difference in Effect Size		-0.02	0.04	0.09	0.13	0.07
p-value for the Difference		1	1	0.61	0.13	0.27
Social Studies Reading Comprehension Assessment Score						
Post-Test (Spring 2007)						
Impact	500.4	-0.67	-0.36	-1.38	-1.91	-1.36
Effect Size		-0.02	-0.01	-0.05	-0.06	-0.05
p-value		1	1	1	0.96	0.70
Follow Up (Spring 2008)						
Impact	498.15	1.42	-0.65	1.70	3.22	1.08
Effect Size		0.05	-0.02	0.06	0.11	0.04
p-value		1	1	1	0.92	0.93
Difference Between Post-Test and Follow Up						
Difference in Impact		2.09	-0.29	3.08	5.13	2.44
Difference in Effect Size		0.07	-0.01	0.10	0.17	0.08
p-value for the Difference		0.99	1	0.83	0.41	0.37

TABLE IV.3 (*continued*)

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Science Reading Comprehension Assessment Score						
Post-Test (Spring 2007)						
Impact	500.61	0.94	-0.42	-1.14	-5.43*	-1.92
Effect Size		0.03	-0.02	-0.04	-0.20	-0.07
p-value		1	1	1	0.05	0.54
Follow Up (Spring 2008)						
Impact	497.27	1.37	1.92	3.18	1.35	2.23
Effect Size		0.04	0.06	0.1	0.04	0.07
p-value		1	1	0.83	1	0.52
Difference Between Post-Test and Follow Up						
Difference in Impact		0.43	2.33	4.31	6.78	4.15
Difference in Effect Size		0.01	0.08	0.14	0.24	0.14
p-value for the Difference		1	0.96	0.75	0.20	0.12
Number of Cohort 1 Students[b]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**

SOURCE: Reading comprehension tests administered by study team.

NOTE: For each outcome, the numbers reported in the column labeled "Control Group Mean" are the average predicted outcomes for all students as if they were in the control group. The numbers reported in the remaining columns are, by row: (1) the impact, (2) the effect size, and (3) the *p-value* of the impact. The *p-values* presented in this table are adjusted for multiple-hypotheses testing. For each outcome, the differences between impacts for the post-test and follow up are also reported. The social studies and science reading comprehension assessments were developed by ETS. Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]The number of students presented in this row is the number of Cohort 1 students participating in the study. The proportion of students in each experimental condition with post-test and follow-up test scores is reported in Appendix G.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

Follow-Up Impact Findings. There were no statistically significant impacts of the interventions on any of the three follow-up test score outcomes (GRADE, social studies reading comprehension assessment, and science reading comprehension assessment) for sixth-grade students from the first cohort. We did not find any statistically significant impacts in comparisons of follow-up test scores of students in each intervention group with follow-up test scores of students in the control group or in comparisons of test scores of students in the combined treatment group with test scores of the students in the control group. There were also no statistically significant differences between the intervention group follow-up impacts (not shown in table). In particular, the statistically significant negative impacts observed in the study's first year (a negative impact of Reading for Knowledge on the post-test composite and science reading comprehension assessment scores, and statistically significant negative impact of the combined treatment group on the post-test composite scores) were not found in the second year of the study. These findings provide evidence that the four reading comprehension interventions did not have impacts on test scores outcomes of first cohort students one year after the end of the implementation of the interventions.

Differences Between Post-Test and Follow-Up Impacts.[66] We found two statistically significant differences in post-test and follow-up impacts for first cohort students. (Differences on post-test and follow-up impacts are reported in differences in effect sizes.) There were statistically significant differences between the follow-up and post-test impacts of Reading for Knowledge and the combined treatment group on the composite test score (effect size difference: 0.18 for Reading for Knowledge and 0.10 for the combined treatment group). In both cases, the impacts at post-test were negative and statistically significant, and the follow-up impacts were positive but not statistically significant (*p*-values of 0.77 and 0.61, respectively). These findings provide some evidence that the impacts of Reading for Knowledge and the combined treatment changed over time. However, since the impacts of Reading for Knowledge and the combined treatment at follow up are not statistically significant, these findings cannot be interpreted as evidence of positive impacts of Reading for Knowledge and the combined treatment group in the second year of the study.

Sensitivity Tests to Assess the Robustness of the Impact Findings. We assessed the robustness of these findings through the following sensitivity tests (see Tables H.1, H.2, and H.3 in Appendix H for more information): (1) excluding covariates, (2) using an alternative weighting approach, (3) estimating impacts using a hierarchical linear model (HLM) approach, and (4) focusing only on students with both a pre- and post-test. None of the findings presented above were sensitive to these changes in estimation approach. Specifically, all of the follow-up impacts for first cohort students remained statistically insignificant in each of these sensitivity analyses.

[66]Statistical tests for differences in impacts between post-test and follow up were conducted by estimating a stacked regression model that allowed for the calculation of cross-equation covariance terms.

D. EIGHT OF 360 SUBGROUP ANALYSES YIELD STATISTICALLY SIGNIFICANT IMPACTS

We conducted a series of subgroup analyses to examine secondary research questions related to whether the impacts of the interventions in the second year of the study (at follow up for the first cohort students) vary for sixth grade students with different characteristics. Since these subgroups are formed using characteristics of first cohort students observed at the beginning of the first implementation year (fall 2006), the interventions could not have influenced these student characteristics and thus there should be no systematic differences in unobserved characteristics in these subgroups between the treatment and control groups. Therefore, most of the subgroup analyses preserve the properties of random assignment and the findings allow for causal conclusions to be drawn about the impact of the interventions for these subgroups. As reported in Chapter III, the three exceptions are the subgroups defined by teachers' self-reported past professional development, teaching efficacy, and school professional culture (all of which are based on data collected through the study's first year Teacher Survey, which was administered by the study team in August through November 2006). The number and composition of teachers in the intervention groups who reported receiving past professional development and who reported a given level of teacher efficacy or school professional culture could have been affected by the product-specific training received in the summer before the first implementation year.[67] Since that potential shift in the size and composition of those subgroups affected only the treatment group and not the control group, analyses of those subgroups do *not* maintain the properties of random assignment and, thus, do *not* allow for causal conclusions to be drawn about the impact of the interventions for those subgroups.

To create the subgroups analyzed in this chapter, we followed the same approach used in Chapter III, which was generally to split the student sample into two subgroups of roughly equal size at the median level of each relevant characteristic for the sample of first cohort students. See Chapter III for more information on the subgroups examined.

Similar to the subgroup impacts reported in Chapter III, the subgroup impacts reported in this chapter are based on the difference in follow-up impacts between subgroups among first cohort students (for example, the difference in follow-up impacts between ELL and non-ELL first cohort students). These subgroup impacts are reported in Appendix L in Tables L.5-L.8 (with adjustments for multiple hypothesis testing) and L.13-L.16 (without adjustments for multiple hypothesis testing). In the text that follows, our focus is on the findings that are statistically significant with adjustments for multiple hypothesis testing.[68]

---

[67]In particular, as mentioned in Chapter III, teachers may have reported the training as professional development, and the training may have affected teachers' responses to survey questions on their teaching efficacy and the professional culture in their schools.

[68]These adjustments are conducted in four domains for each subgroup (we do not adjust for multiple comparisons *between* subgroups, only *within* subgroups). The first domain consists of 12 tests—the test for the difference described above for each of four interventions on each of three outcome scores (GRADE, science comprehension, and social studies comprehension). The second domain consists of four tests—the test for the difference described above for each intervention on a composite outcome. The third domain consists of three tests—the test for the difference described above for the combined treatment group on each of three outcome measures.

Subgroup Findings. Eight of the 360 subgroup differences at follow up were statistically significant (one would expect 18 significant findings [5 percent of 360] by chance). For first cohort students, we observed greater impacts of:

1. Project CRISS on composite follow-up scores for students who scored in the bottom third of the pre-test TOSCRF distribution (effect size: 0.09) than for those who scored in the top third (effect size: -0.04) (Table L.5).
2. Project CRISS on GRADE follow-up scores for students with pre-test TOSCRF scores below the national norm sample average (effect size: -0.02) than for those who scored above that average (effect size: -0.20) (Table L.6).
3. Project CRISS on GRADE follow-up scores for students who scored in the bottom third of the pre-test TOSCRF distribution (effect size: 0.05) than for those who scored in the top third (effect size: -0.09) (Table L.6).
4. ReadAbout on composite follow-up scores for students who scored in the middle third of the pre-test TOSCRF distribution (effect size: 0.09) than for those who scored in the bottom third (effect size: -0.09) (Table L.5).
5. The combined treatment group on composite follow-up scores for students who scored in the bottom third of the pre-test GRADE distribution (effect size: 0.12) than for those who scored in the middle third (effect size: -0.04) (Table L.5).
6. The combined treatment group on composite follow-up scores for students not classified as ELL (effect size: 0.05) than for those classified as ELL (effect size: -0.06) (Table L.5).
7. The combined treatment group on GRADE follow-up scores for students with pre-test TOSCRF scores below the national norm sample average (effect size: 0.02) than for those who scored above that average (effect size: -0.08) (Table L.6).
8. The combined treatment group on the social studies reading comprehension assessment follow-up scores for students who scored in the bottom third of the pre-test GRADE distribution (effect size: 0.11) than for those who scored in the middle third (effect size: -0.06) (Table L.7).

E. FIVE OF 60 TEACHER PRACTICES SUBGROUP DIFFERENCES ARE STATISTICALLY SIGNIFICANT

Similar to the analysis presented in Chapter III, we investigated the relationship between intervention impacts at follow up (one year after intervention implementation ended) for first cohort students and classroom practices[69] during the year when the interventions were

(continued)

The fourth domain consists of one test—the test for the difference described above for the combined treatment group on the composite outcome.

[69]See Chapter II for more information on the three teacher practice scales the study team constructed.

implemented with these students (that is, during the first year of the study when the first cohort students were in fifth grade). We did this by conducting analyses of follow-up test scores for first cohort students in classrooms with different levels of observed teaching practices (as with the subgroup analyses described above, we split the sample at the median levels of teacher practices observed). As described in Chapter III, these relationships must be interpreted cautiously because the research design did not randomly assign different levels of teacher practices to teachers. Therefore, estimates of the relationship between intervention impacts and teacher practices cannot be interpreted as providing rigorous impact estimates and do *not* allow causal conclusions to be drawn about the impact of the interventions for these subgroups.

We report teacher subgroup impacts based on the same types of subgroup differences described in Section D, using the same approach to adjusting for multiple comparisons. The findings are reported in Appendix Tables L.5 through L.8 and L.13 through L.16.

Five of 60 teacher practice subgroup differences at follow up were statistically significant (one would expect three significant findings [5 percent of 60] by chance). For first cohort students, we found greater impacts of:

1. ReadAbout on composite follow-up scores for students whose classrooms in the first year of the study had Classroom Management Scale scores above the sample median (effect size: 0.07) than for students whose classrooms had scores below the sample median (effect size: -0.07) (Table L.5).

2. The combined treatment group on composite follow-up scores for students whose classrooms in the first year of the study had Classroom Management Scale scores above the sample median (effect size: 0.07) than for students whose classrooms had scores below the sample median (effect size: -0.02) (Table L.5).

3. ReadAbout on GRADE follow-up scores for students whose classrooms in the first year of the study had Classroom Management Scale scores above the sample median (effect size: 0.07) than for students whose classrooms had scores below the sample median (effect size: -0.09) (Table L.6).

4. The combined treatment group on GRADE follow-up scores for students whose classrooms in the first year of the study had Classroom Management Scale scores above the sample median (effect size: 0.04) than for students whose classrooms had scores below the sample median (effect size: -0.05) (Table L.6).

5. Reading for Knowledge on the social studies reading comprehension assessment follow-up scores for students whose classrooms in the study's first year had Traditional Interaction Scale scores below the sample median (effect size: 0.26) than for students whose classrooms had scores above the sample median (effect size: -0.04) (Table L.7).

V. ADDITIONAL DESCRIPTIVE AND NONEXPERIMENTAL ANALYSES

This chapter presents additional descriptive and nonexperimental analyses that go beyond the study's original research questions and, in most cases, are not related to the experimental design (the exception being Section E). These analyses do not answer questions of intervention effectiveness. Instead, they are correlational and descriptive analyses that are intended to inform future research and program development efforts. In summary, these analyses are not experimental and do not support causal conclusions. Therefore, findings from these analyses should be interpreted cautiously.

The chapter begins with an examination of the descriptive data from the ERC classroom observations (Section A) and how they relate to student reading comprehension achievement (Section B). We then turn to an examination of how other teacher characteristics, specifically self-reported teaching efficacy and past professional development, relate to student test scores (Section C). In Section D we explore whether reading comprehension achievement is related to (1) the time teachers reported spending on reading activities with students on a given day and (2) the time teachers reported that students used informational text in a typical week. Finally, in Section E we examine correlations between intervention impacts for each block (or groups of schools within which random assignment was conducted) and the average characteristics of schools in those blocks.

A. DESCRIPTIVE INFORMATION ON CLASSROOM PRACTICES

The observational data gathered from the Expository Reading Comprehension (ERC) observation protocol (described above in Chapter II) allows us to provide a snapshot of the nature of reading comprehension instruction in 270 fifth-grade classrooms in the United States.[70] The districts and schools participating in the study were not randomly selected from the universe of districts and schools in the United States. While the findings from the ERC data collected in study classrooms do not generalize statistically to the broader population of classrooms serving fifth-grade students in the United States, an examination of this data can still contribute to the literature addressing the extent to which teachers provide instruction to students on how to make sense of text. Durkin (1978) noted that teachers tended to ask students questions and tell them whether their answers were right or wrong, but provided little guidance in how to think through solutions to problems and answers to questions. A more recent study by Connor et al. (2004) of third-grade teachers indicated that, on average, only about a minute of the daily language arts instructional time was spent on explicit instruction of reading comprehension strategies.

As described above in Chapter II, the ERC observation form enabled the study team to tally the number of times instructional practices were seen in treatment and control classrooms during

[70] 270 classrooms were included the first year of the study and 190 classrooms were included in the second year of the study. The data discussed in this chapter are based on the 270 and 190 classrooms in Years 1 and 2, respectively.

the study's classroom observations (recall that these non-curricula-specific observations were conducted in both treatment and control classrooms and were designed to examine the extent to which specific teaching practices related to vocabulary and comprehension instruction were observed). Two types of information are presented in the tables in this section: (1) the number of times practices were observed in Years 1 and 2 and (2) differences in the number of times practices were observed in the two years.

Recall that observations included any time during the day that students and teachers worked with informational text. This could include parts of the reading/language arts lesson, history lesson, science lesson, and, for the intervention classrooms, the time during the day devoted to the intervention. Therefore, the amount of time devoted to instruction using informational text varied across classrooms. Thus, for Part I items (see Table V.1a) related to vocabulary and comprehension instruction, the frequency with which instructional practices were observed was constructed in two ways: (1) based on sums of activities across all observation intervals during the course of an entire school day and (2) based on averages of activities across the observation intervals. These two methods of constructing frequencies provide two important pictures of instructional practices: the total number of times students are exposed to a particular practice in a given day and the average number of times they are exposed to a particular practice in a 10-minute time period.

In the text that follows, we focus on describing the most frequently implemented instructional practices, as well as describing instances when the Year 1 and Year 2 teacher practices were statistically significantly different. We examine differences between the two years to assess whether teaching practices of the full set of teachers participating in the study changed between the years. Because data from Year 1 and Year 2 are not based on the same set of teachers, the observed differences could reflect compositional changes as well as changes in teacher practices. The comparisons are still useful, however, for understanding whether the overall teaching practices experienced by Cohort 1 students differed from the overall teaching practices experienced by Cohort 2 students.

The most frequently observed reading comprehension instructional practice (see first pane of Table V.1a), based on the sum of activities across all observation intervals in a school day, was students practicing using reading comprehension strategies (on average, 15.34 total times during the observations conducted in Year 2). The least frequently observed reading comprehension instructional practices, based on the total number of times the practices were observed during an entire school day, were those involving teacher modeling (which ranged from 0.03 to 0.18 times during the observations conducted in Year 2). Three items were observed significantly more often in Year 2 relative to Year 1:

1. Teacher explains, reviews, and provides examples and elaborations of activating prior knowledge and/or previewing text before reading (6.5 times in Year 2 vs. 4.6 times in Year 1, *p*-value = .03)

2. Teacher explains, reviews, and provides examples and elaborations when providing explicit comprehension instruction that teaches students about text structure (2.9 times in Year 1 vs. 1.8 times in Year 2, *p*-value =.03)

TABLE V.1a

DESCRIPTIVE STATISTICS FOR EXPOSITORY READING COMPREHENSION CLASSROOM OBSERVATION INSTRUMENT ITEMS

	Total Number of Times Observed[a]				Average Number of Times Observed[b]			
	Year 1	Year 2	Difference	*p-value*	Year 1	Year 2	Difference	*p-value*
Part I, Comprehension[c]								
Activates prior knowledge and/or previews text before reading								
Teacher models	0.06	0.04	-0.02	0.40	0.01	0.00	-0.00	0.18
Teacher explains, reviews, provides examples and elaborations	4.55	6.45	1.90*	0.03	0.61	0.71	0.10	0.32
Students practice	7.95	9.86	1.91	0.12	1.07	1.06	-0.01	0.96
Explicit comprehension instruction that teaches students about text structure								
Teacher models	0.03	0.04	0.02	0.59	0.00	0.00	0.00	0.96
Teacher explains, reviews, provides examples and elaborations	1.84	2.90	1.06*	0.03	0.24	0.32	0.08	0.18
Students practice	2.79	4.69	1.90*	0.02	0.34	0.50	0.17	0.06
Explicit comprehension instruction that teaches students how to use comprehension strategies								
Teacher models	0.10	0.18	0.09	0.42	0.01	0.01	0.00	0.69
Teacher explains, reviews, provides examples and elaborations	8.20	9.01	0.81	0.48	1.22	1.03	-0.19	0.26
Students practice	12.66	15.34	2.68	0.11	1.75	1.78	0.04	0.87
Explicit comprehension instruction that teaches students how to generate questions								
Teacher models	0.05	0.05	0.00	0.92	0.00	0.00	-0.00	0.88
Teacher explains, reviews, provides examples and elaborations	1.93	2.68	0.74	0.18	0.24	0.27	0.02	0.64
Students practice	3.44	4.64	1.20	0.20	0.43	0.47	0.04	0.62
Explicit comprehension instruction that teaches text features to interpret text								
Teacher models	0.01	0.03	0.02	0.21	0.00	0.00	0.00	0.86
Teacher explains, reviews, provides examples and elaborations	1.41	1.63	0.21	0.49	0.19	0.17	-0.02	0.63
Students practice	1.83	2.17	0.34	0.44	0.24	0.22	-0.02	0.71
Teacher asks students to justify their responses	1.89	2.55	0.67	0.09	0.24	0.27	0.03	0.39

TABLE V.1a (*continued*)

	Total Number of Times Observed[a]				Average Number of Times Observed[b]			
	Year 1	Year 2	Difference	*p-value*	Year 1	Year 2	Difference	*p-value*
Teacher asks questions based on material in the text that are beyond the literal level	8.10	7.85	-0.26	0.79	0.96	0.90	-0.07	0.57
Teacher elaborates, clarifies, or links concepts during and after text reading	10.44	10.04	-0.40	0.72	1.29	1.17	-0.12	0.41
Part I, Vocabulary[c]								
Teacher provides an explanation and/or a definition or asks a student to read a definition	5.35	5.05	-0.30	0.60	0.71	0.54	-0.17*	0.02
Teacher provides examples, contrasting examples, multiple meanings, immediate elaborations to students' responses	6.90	7.26	0.36	0.66	0.87	0.80	-0.06	0.55
Teacher uses visuals/pictures, gestures related to word meaning, facial expressions, or demonstrations to discuss/demonstrate word meanings	1.91	1.93	0.03	0.95	0.23	0.21	-0.01	0.81
Teacher teaches word-learning strategies using context clues, word parts, root meaning	0.66	0.80	0.14	0.42	0.09	0.09	-0.00	0.86
Students do or are asked to do something that requires knowledge of words	11.23	13.41	2.18	0.13	1.39	1.47	0.08	0.64
Students are given an opportunity to apply word-learning strategies using context clues, word parts, and root meaning	0.77	0.95	0.18	0.47	0.12	0.10	-0.02	0.59

SOURCE: Classroom observations.

[a]The number reported is the total number of times each behavior was observed across the full day.

[b]The number reported is the average number of times each practice was observed across all 10-minute observation intervals.

[c]For items in this pane of the table, observers recorded tallies for the number of times each behavior was observed.

*Statistically different at the .05 level.

3. Students practice working with text structure (4.7 times in Year 2 vs. 2.8 times in Year 1, *p*-value =.02)

There were no statistically significant differences between Years 1 and 2 for the other total frequencies with which reading comprehension items were observed, or for frequencies based on averages.

The most frequently observed vocabulary item (see second pane of Table V.1a) was students doing—or being asked to do—something that requires knowledge of words such as providing a definition or an example, or using the word in a sentence (13.4 total times during the observations conducted in Year 2). There were no statistically significant differences between Years 1 and 2 with one exception—the average frequency with which teachers were observed providing an explanation and/or definition or asking students to read a definition was statistically significantly lower in Year 2 than in Year 1 (0.71 times in Year 1 vs. 0.54 times in Year 2, *p*-value = .02).

There were no statistically significant differences between Years 1 and 2 in the type of grouping arrangements observed in classrooms. In both years, teachers were observed working with the whole class most frequently (in 82 and 85 percent of the 10-minute intervals observed in Years 1 and 2, respectively) (Table V.1b). Teachers were seen working with small groups of three to six students in 21 and 16 percent of the 10-minute intervals observed in Years 1 and 2, respectively.

Teachers were observed most frequently implementing supported oral reading of connected text (observed in 39 and 46 percent of the 10-minute intervals observed in Years 1 and 2, respectively) (Table V.1b). Teachers used independent or buddy oral reading in 32 and 21 percent of the 10-minute intervals observed in Years 1 and 2, respectively. The difference between the years was statistically significant (*p*-value = .00). We also observed a statistically significant difference in the extent to which teachers read aloud with students following along silently (observed in 16 and 24 percent of the 10-minute intervals observed in Years 1 and 2, respectively, *p*-value = .01).

For Part II items, which were recorded once per observation, the study team used the average value recorded across observations, to provide an overall picture for the day's instruction. The first nine items (Table V.1c) were yes/no items, and the remaining items (Table V.1d) used Likert scales.

In Years 1 and 2, teachers were most frequently observed providing opportunities for most students to participate actively during teacher-led instruction (87 percent in Year 1 and 82 percent in Year 2) and pacing instruction so that the length of the comprehension or vocabulary activities was appropriate for the age group (88 percent in Year 1 and 82 percent in Year 2) (Table V.1c). There was a statistically significant difference between the two years for the latter item (*p*-value = .03). Teachers were observed giving inaccurate and/or confusing explanations or feedback in 3 percent of observations in Year 1 and 8 percent in Year 2, a difference that was statistically significant (*p*-value = .01). There was also a statistically significant difference between the two years in the percentage of teachers observed keeping

TABLE V.1b

DESCRIPTIVE STATISTICS FOR EXPOSITORY READING COMPREHENSION CLASSROOM OBSERVATION INSTRUMENT ITEMS

	Average Number of Times Observed[a]			
	Year 1	Year 2	Difference	*p-value*
Part I, Grouping Arrangements and Text Reading[b]				
Teacher is working with:				
Whole class (≥75% of class)	0.82	0.85	0.03	0.14
Large group (> 6 students, < 75% of class)	0.02	0.02	0.00	0.93
Small groups (3-6 students)	0.21	0.16	-0.05	0.06
Pairs	0.09	0.08	-0.01	0.67
An individual	0.04	0.05	0.01	0.33
No direct student contact	0.01	0.01	-0.00	0.77
Text reading (applies to reading-connected text)				
Supported oral reading (includes choral and round-robin reading)	0.39	0.46	0.07	0.05
Independent silent reading	0.25	0.22	-0.03	0.35
Independent or buddy oral reading	0.32	0.21	-0.11*	0.00
Teacher reads aloud	0.17	0.12	-0.04	0.11
Teacher reads aloud with students following along silently	0.16	0.24	0.08*	0.01
Text not present	0.05	0.07	0.01	0.46
Text present but not being read	0.23	0.25	0.02	0.47

SOURCE: Classroom observations.

[a]The number reported is the average number of times each practice was observed across all 10-minute observation intervals.

[b]For items in this pane of the table, observers selected all items that they observed (more than one category could be selected).

*Statistically different at the .05 level.

TABLE V.1c

DESCRIPTIVE STATISTICS FOR EXPOSITORY READING COMPREHENSION CLASSROOM OBSERVATION INSTRUMENT ITEMS

	Average Number of Times Observed[a]			
	Year 1	Year 2	Difference	*p-value*
Part II, Instruction Effectiveness[b]				
Gave inaccurate and/or confusing explanations or feedback	0.03	0.08	0.05*	0.01
Missed opportunity to correct or address error	0.05	0.09	0.04	0.10
Provided opportunities for most students to participate actively during teacher-led instruction	0.87	0.82	-0.06	0.07
Paced instruction so that the length of the comprehension or vocabulary activities was appropriate for this age group	0.88	0.82	-0.07*	0.03
Taught using outlining and/or note taking	0.32	0.26	-0.06	0.07
Used graphic organizers	0.33	0.29	-0.04	0.28
Kept students thinking for two or more seconds before calling on a student to respond to a complex question	0.62	0.49	-0.13*	0.01
Gave independent/pairs/small-group practice in answering comprehension questions or applying comprehension strategy(ies) with expected written product	0.56	0.47	-0.09	0.07
Used writing activities in response to reading (does not include fill-in-the-blank or one-word answers)	0.39	0.34	-0.05	0.23

SOURCE: Classroom observations.

[a]The number reported is the average number of times each practice was observed across all observations.

[b]For items in this pane of the table, observers recorded "Yes" or "No" for each item.

*Statistically different at the .05 level.

TABLE V.1d

DESCRIPTIVE STATISTICS FOR EXPOSITORY READING COMPREHENSION CLASSROOM OBSERVATION INSTRUMENT ITEMS

	Average Number of Times Observed[a]			
	Year 1	Year 2	Difference	*p-value*
Part II, Teachers' Management/Responsiveness to Students[b]				
Teacher maximized the amount of time available for instruction	3.25	3.26	0.01	0.94
Teacher managed student behavior effectively in order to avoid disruptions and provide productive learning environments	3.40	3.39	-0.01	0.87
Teacher redirected discussion if a student response was leading the group off topic/focus	3.30	3.12	-0.17	0.20
Part II, Student Engagement[c]				
Student engagement during the first half of the observation session	2.64	2.72	0.08	0.12
Student engagement during the remainder of the observation session	2.58	2.61	0.03	0.57

SOURCE: Classroom observations.

[a]The number reported is the average value recorded across all observations.

[b]For items in this pane of the table, observers record a "1" for minimal/poor, "2" for fair, "3" for good, or "4" for excellent.

[c]For items in this pane of the table, observers could record a "1" for few engaged, "2" for many engaged, or "3" for most engaged.

*Statistically different at the .05 level.

students thinking for two or more seconds before calling on a student to respond to a complex question (62 percent in Year 1 vs. 49 percent in Year 2, *p*-value = .01).

The items from Part II of the ERC on teachers' management of the classroom and responsiveness to students were recorded using a 1 to 4 Likert scale, with 1 meaning "Minimal/Poor" and 4 meaning "Excellent." All items in this section were rated above 3 for Years 1 and 2, and there were no statistically significant differences between the two years (Table V.1d). The items from Part II of the ERC on student engagement were recorded using a 1 to 3 Likert scale, with 1 meaning "Few Engaged," 2 meaning "Many Engaged," and 3 meaning "Most Engaged." All items in this section were rated above 2 for Years 1 and 2, and there were no statistically significant differences between the two years.

B. RELATIONSHIP BETWEEN CLASSROOM PRACTICES AND TEST SCORES

Studies using observational data and correlational analysis have indicated a positive and statistically significant relationship between certain interactive teaching practices related to reading comprehension and student reading comprehension outcomes. For example, Connor et al. (2004) observed 43 third-grade classrooms and found that children achieved greater reading comprehension growth when more time was spent in explicit, interactive instruction in reading comprehension strategies. Stallings (1975), who observed 171 third-grade teachers for three days as part of the Follow Through study, found that interactive teaching practices (such as presenting information, asking students questions, and providing immediate corrective feedback) were associated with higher reading comprehension and vocabulary scores. And Denham and Lieberman (1980), who observed close to 300 second- and fifth-grade teachers extensively as part of the six-year Beginning Teacher Evaluation Study, found that substantive interactive instruction (the combination of teacher explanations, questioning of students, and provision of feedback) was associated with significantly high levels of academic engagement (r = .45) (reported in Rosenshine 1980, p. 121), which in turn was significantly positively associated with student comprehension and vocabulary performance (Borg 1980, p. 52).

Given the positive relationship between interactive teaching practices in reading and student outcomes that has been reported in prior observational research, we thought it was important to examine the relationship between the ERC observational data and student reading outcomes in this study.[71] The ERC classroom observations (as described in Chapters I and II) were designed to gather information on the number of times treatment and control group teachers engaged in a set of general, non-intervention-specific teaching practices related to reading comprehension and vocabulary instruction.

In this section we examine whether these practices as measured by the ERC are associated with student reading comprehension outcomes.[72] We first examine whether the three ERC scales

[71]The ERC differs from some of the observational measures developed in the 1970s and 1980s in that it also includes items that examine the extent to which teachers think aloud or model use of comprehension strategies.

[72] Although the ERC was used to observe classrooms multiple times during the day, classrooms were only observed for a single day, which may reduce the reliability of the teacher practice scales based on the ERC data

(Traditional Interaction, Reading Strategy Guidance, and Classroom Management and Student Engagement) are associated with students' post-test scores, both with and without regression adjustment for students' pre-test scores and other characteristics of teachers and students.[73] Second, we examine which of the individual ERC items are associated with student test scores. Finally, we examine how the association between ERC items and student test scores relates to the frequency with which those practices are observed.

Correlations between the ERC scales and reading comprehension post-tests are presented in Table V.2. The numbers in this table are the standardized coefficients on each scale in a regression of a post-test on the scale and other covariates. Below each regression coefficient is a *p*-value. The regression coefficients in the first column come from a regression model without any other covariates. The second column adds pre-test scores from the GRADE and TOSCRF as covariates. The third column adds student ethnicity and race, student English language learner status, school location, and teacher race. Finally, the fourth column adds the remaining ERC scales (for example, it reports the effect of the Traditional Interaction scale adjusted for the other two scales). Adjusting for the other scales shows the correlation between each scale and test scores, holding the other scales constant. For example, this analysis indicates how test scores are expected to change when the Classroom Management scale is increased while holding both Reading Strategy Guidance and Traditional Interaction scales constant (for example, the classroom management scale has a statistically significant correlation of 0.04 with the GRADE when holding the Reading Strategy Guidance and Traditional Interaction scales constant).[74] All variables in these regressions are standardized.[75]

This table shows that the Classroom Management scale has the most consistently positive and statistically significant association with post-test scores. For three of the four outcomes presented in this table, the relationship between the Classroom Management scale and the post-test is positive and statistically significant even after adjusting for baseline covariates and the

---

(continued)

(relative to observations conducted over multiple days). When the teacher practice scales based on a single day of observations are used in the correlational analyses in this chapter, the correlations may be attenuated.

[73]Treatment and control students from Cohorts 1 and 2 are included in the analyses presented in Sections B through D of this chapter. The analyses focus on post-test (end of fifth grade) data that were collected in spring 2007 for Cohort 1 and spring 2008 for Cohort 2 (this was after treatment students were exposed to one year of intervention implementation). Note that teachers of *Cohort 1* treatment students were implementing the study interventions for the first time, while teachers of *Cohort 2* treatment students had one year of experience using the interventions as part of the first year of the study.

[74]As shown in Table II.18, several items are common to both the Traditional Interaction and the Reading Strategy Guidance scales (there is no overlap between the Classroom Management scale and either of the other two scales). This overlap in items could lead to a high correlation between the Reading Strategy Guidance and Traditional Interaction scales, which could confound our findings reported in the last column of Table V.2. To investigate this issue, we also estimated these regressions excluding the Traditional Interaction scale. We found that the sign and statistical significance of the correlations reported in the last column of Table V.2 do not change when the Traditional Interaction scale is excluded.

[75]To standardize each variable used in the regressions, we subtracted the mean from the variable and then divided by its standard deviation.

TABLE V.2

REGRESSION-ADJUSTED CORRELATION BETWEEN EXPOSITORY READING COMPREHENSION (ERC) SCALES AND STUDENT POST-TEST SCORES

	Covariate Adjustment			
	None	Pretest	Pretest and Other Covariates	Pretest, Other Covariates, and ERC Scales
GRADE Score				
Traditional Interaction	0.01 (0.56)	0.01 (0.67)	-0.01 (0.45)	-0.01 (0.45)
Reading Strategy Guidance	0.05 (0.06)	0.03* (0.03)	0.02* (0.01)	0.01 (0.43)
Classroom Management	0.09* (0.00)	0.05* (0.00)	0.04* (0.00)	0.04* (0.00)
Social Studies Reading Comprehension Assessment Score				
Traditional Interaction	0.01 (0.58)	0.01 (0.72)	-0.02 (0.29)	-0.01 (0.58)
Reading Strategy Guidance	0.05* (0.05)	0.04* (0.05)	0.03 (0.08)	0.02 (0.39)
Classroom Management	0.07* (0.00)	0.04* (0.03)	0.02 (0.19)	0.02 (0.18)
Science Reading Comprehension Assessment Score				
Traditional Interaction	0.04 (0.11)	0.03 (0.07)	0.01 (0.64)	0.01 (0.65)
Reading Strategy Guidance	0.07* (0.02)	0.05* (0.01)	0.04* (0.02)	0.03 (0.25)
Classroom Management	0.13* (0.00)	0.10* (0.00)	0.08* (0.00)	0.07* (0.00)
Composite Test Score[a]				
Traditional Interaction	0.02 (0.35)	0.02 (0.33)	-0.01 (0.55)	-0.004 (0.72)
Reading Strategy Guidance	0.06* (0.04)	0.04* (0.01)	0.03* (0.00)	0.02 (0.18)
Classroom Management	0.11* (0.00)	0.07* (0.00)	0.05* (0.00)	0.05* (0.00)

SOURCE: Classroom observations and reading comprehension tests administered by study team.

NOTE: Standardized regression coefficients and *p-values* (in parentheses) are reported for each ERC scale. The *p-values* presented in this table are adjusted for clustering of students within schools but not for multiple-hypotheses testing. The outcome for each regression is indicated by the pane labels (in gray shading). Other covariates in the regression models vary by column. The first column includes no additional covariates (so it is simply regressing test scores on each of the ERC scales). The second column includes the GRADE and TOSCRF tests administered in the fall of each cohort year. In addition to those two tests, the third column also includes student ethnicity and race, student English language learner status, school location, and teacher race. In addition to those covariates, the fourth column adds the remaining ERC scales (for example, it reports the effect of the traditional interaction scale adjusted for the other two scales).

TABLE V.2 (*continued*)

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

other scales (see the last column in Table V.2). No other scale has a statistically significant association with post-test scores after adjusting for that set of covariates. The Reading Strategy Guidance scale does have a positive and statistically significant association with three out of four post-test scores when adjusting for baseline covariates but not the other scales (see the third column in Table V.2). The Traditional Interaction scale was not statistically significantly related to any of the four test scores.

The association between individual ERC instrument items and student test scores is shown in Table V.3. The instrument items that have the most statistically significant correlations with test scores after adjusting for baseline covariates are:

- The three items under the heading "Explicit comprehension instruction that teaches students how to use comprehension strategies" (8 of 12 correlations are statistically significant, all of which show that the more teachers were observed implementing these practices, the higher were student test scores)
- The three items under the heading "Part II, Teachers' Management/Responsiveness to Students" (10 of 12 correlations are statistically significant, all of which show that the higher teachers' scores were on these items during the observation, the higher were student test scores)
- The two items under the heading "Part II, Student Engagement" (6 of 8 correlations are statistically significant, all of which show that the more engaged students were during the observation, the higher were student test scores)

One interesting question that can be examined with study data is whether behaviors that are observed most frequently are those for which there is a statistically significant relationship with test scores. To investigate this issue, we examined the correlation between the number of times ERC items were observed (focusing on behaviors in Table V.1a, which are items for which observers marked tallies each time the item was observed) and the extent to which there was a statistically significant correlation between the item and post-test scores (Table V.3).

We found that, among the ERC comprehension and vocabulary items, there is a positive, statistically significant correlation between the number of times items were observed and the extent to which there was a statistically significant correlation between the item and post-test scores. For each of these 24 items, we calculated the correlation between the average number of times it was observed per interval (the average of the fifth and sixth columns in Table V.1a) and the number of statistically significant correlations for that item in Table V.3 (only counting the columns that are regression-adjusted for baseline covariates, meaning a maximum of four statistically significant correlations per item). The correlation was 0.58 and is statistically significant, meaning that the more commonly implemented teaching practices are more likely to be correlated with higher post-test scores.

TABLE V.3

REGRESSION-ADJUSTED CORRELATION BETWEEN EXPOSITORY READING COMPREHENSION (ERC) INSTRUMENT ITEMS AND STUDENT POST-TEST SCORES

	Outcomes and Covariate Adjustment							
	Composite[a]		GRADE		Social Studies Reading Comprehension		Science Reading Comprehension	
	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates
Part I, Comprehension[b]								
Activates prior knowledge and/or previews text before reading								
Teacher models	0.017 (0.56)	-0.002 (0.75)	0.017 (0.52)	0.000 (0.99)	-0.001 (0.95)	-0.007 (0.47)	0.028 (0.36)	0.009 (0.47)
Teacher explains, reviews, provides examples and elaborations	0.025 (0.40)	-0.007 (0.50)	0.032 (0.22)	-0.001 (0.95)	0.014 (0.59)	-0.014 (0.25)	0.024 (0.33)	-0.013 (0.23)
Students practice	0.007 (0.84)	-0.012 (0.31)	0.016 (0.58)	-0.003 (0.77)	-0.014 (0.62)	-0.027 (0.05)	0.008 (0.80)	-0.013 (0.40)
Explicit comprehension instruction that teaches students about text structure								
Teacher models	-0.019 (0.39)	0.009 (0.28)	-0.018 (0.36)	0.004 (0.38)	-0.013 (0.48)	0.018 (0.09)	-0.025 (0.30)	0.001 (0.97)
Teacher explains, reviews, provides examples and elaborations	0.003 (0.90)	0.019 (0.07)	-0.002 (0.94)	0.014 (0.13)	0.014 (0.52)	0.021 (0.08)	-0.005 (0.87)	0.008 (0.58)
Students practice	-0.005 (0.85)	0.018 (0.11)	-0.006 (0.80)	0.016 (0.08)	-0.004 (0.84)	0.009 (0.44)	-0.002 (0.95)	0.012 (0.50)
Explicit comprehension instruction that teaches students how to use comprehension strategies								
Teacher models	-0.035* (0.01)	0.008 (0.16)	-0.041* (0.00)	-0.004 (0.39)	-0.014 (0.34)	0.020* (0.01)	-0.015 (0.14)	0.009 (0.42)
Teacher explains, reviews, provides examples and elaborations	0.113* (0.00)	0.041* (0.01)	0.092* (0.00)	0.026* (0.04)	0.091* (0.00)	0.033 (0.07)	0.107* (0.00)	0.057* (0.00)
Students practice	0.116* (0.00)	0.054* (0.00)	0.094* (0.00)	0.038* (0.00)	0.101* (0.00)	0.049* (0.00)	0.108* (0.00)	0.060* (0.00)
Explicit comprehension instruction that teaches students how to generate questions								
Teacher models	-0.015 (0.13)	-0.001 (0.89)	-0.009 (0.24)	0.004 (0.49)	-0.006 (0.45)	0.002 (0.83)	-0.012 (0.52)	0.000 (0.99)
Teacher explains, reviews, provides examples and elaborations	0.039 (0.11)	-0.003 (0.80)	0.036 (0.10)	0.000 (0.99)	0.017 (0.51)	-0.021 (0.25)	0.053* (0.02)	0.017 (0.30)

Table V.3 *(continued)*

	Outcomes and Covariate Adjustment							
	Composite[a]		GRADE		Social Studies Reading Comprehension		Science Reading Comprehension	
	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates
Students practice	0.039	-0.001	0.037	0.003	0.023	-0.009	0.032	0.003
	(0.32)	(0.96)	(0.29)	(0.82)	(0.56)	(0.68)	(0.33)	(0.87)
Explicit comprehension instruction that teaches text features to interpret text								
Teacher models	0.026	0.003	0.025	0.005	0.008	-0.005	0.035	0.011
	(0.26)	(0.87)	(0.27)	(0.80)	(0.65)	(0.73)	(0.11)	(0.57)
Teacher explains, reviews, provides examples and elaborations	0.033	0.000	0.036	0.005	0.006	-0.026	0.036	-0.001
	(0.28)	(1.00)	(0.17)	(0.71)	(0.87)	(0.36)	(0.17)	(0.93)
Students practice	0.021	0.025	0.022	0.026	-0.008	-0.009	0.035	0.027
	(0.47)	(0.20)	(0.39)	(0.10)	(0.81)	(0.74)	(0.16)	(0.13)
Teacher asks students to justify their responses	0.044	-0.004	0.033	-0.007	0.029	-0.009	0.061	0.008
	(0.20)	(0.78)	(0.26)	(0.57)	(0.33)	(0.60)	(0.07)	(0.69)
Teacher asks questions based on material in the text that are beyond the literal level	0.006	0.002	-0.007	-0.010	-0.009	-0.007	0.051	0.038*
	(0.85)	(0.91)	(0.81)	(0.38)	(0.73)	(0.72)	(0.08)	(0.03)
Teacher elaborates, clarifies, or links concepts during and after text Reading	0.032	0.008	0.028	0.003	0.009	0.000	0.054*	0.028
	(0.25)	(0.57)	(0.29)	(0.83)	(0.73)	(0.98)	(0.05)	(0.13)
Part I, Vocabulary[b]								
Teacher provides an explanation and/or a definition or asks a student to read a definition	0.022	-0.026	0.021	-0.013	0.021	-0.027*	0.009	-0.028
	(0.44)	(0.05)	(0.41)	(0.25)	(0.39)	(0.04)	(0.75)	(0.09)
Teacher provides examples, contrasting examples, multiple meanings, immediate elaborations to students' responses	0.035	0.000	0.027	-0.004	0.026	-0.004	0.035	0.012
	(0.39)	(0.97)	(0.45)	(0.70)	(0.46)	(0.76)	(0.34)	(0.36)
Teacher uses visuals/pictures, gestures related to word meaning, facial expressions, or demonstrations to discuss/demonstrate word meanings	-0.020	0.002	-0.026	-0.003	-0.010	0.002	-0.027	0.003
	(0.65)	(0.92)	(0.51)	(0.87)	(0.82)	(0.94)	(0.54)	(0.86)
Teacher teaches word-learning strategies using context clues, word parts, root meaning	0.073*	0.003	0.061*	0.001	0.072*	0.009	0.063*	0.000
	(0.01)	(0.78)	(0.01)	(0.91)	(0.00)	(0.45)	(0.02)	(0.99)
Students do or are asked to do something that requires knowledge of Words	0.028	0.004	0.017	-0.005	0.026	-0.003	0.040	0.021
	(0.39)	(0.74)	(0.57)	(0.60)	(0.39)	(0.86)	(0.18)	(0.17)
Students are given an opportunity to apply word-learning strategies using context clues, word parts, and root meaning	0.033	-0.006	0.022	-0.009	0.028	-0.009	0.034	-0.003
	(0.06)	(0.46)	(0.18)	(0.18)	(0.05)	(0.28)	(0.07)	(0.79)

Table V.3 *(continued)*

	Outcomes and Covariate Adjustment							
	Composite[a]		GRADE		Social Studies Reading Comprehension		Science Reading Comprehension	
	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates
Part I, Grouping Arrangements and Text Reading[c]								
Teacher is working with:								
Whole class (≥75% of class)	-0.034	-0.002	-0.042	-0.012	-0.006	0.017	-0.029	-0.003
	(0.26)	(0.86)	(0.11)	(0.29)	(0.81)	(0.20)	(0.35)	(0.85)
Large group (> 6 students, < 75% of class)	0.012	-0.002	0.017	0.001	0.010	-0.001	-0.005	-0.010
	(0.62)	(0.73)	(0.39)	(0.91)	(0.56)	(0.87)	(0.87)	(0.57)
Small groups (3-6 students)	-0.006	0.012	-0.002	0.016	-0.023	-0.009	-0.002	0.018
	(0.84)	(0.35)	(0.95)	(0.17)	(0.45)	(0.60)	(0.96)	(0.24)
Pairs	0.035	0.000	0.039	-0.003	0.023	0.002	0.018	0.005
	(0.18)	(0.97)	(0.09)	(0.78)	(0.37)	(0.92)	(0.48)	(0.73)
An individual	-0.037	-0.012	-0.029	-0.009	-0.031	-0.014	-0.028	-0.008
	(0.20)	(0.21)	(0.24)	(0.28)	(0.35)	(0.35)	(0.22)	(0.56)
No direct student contact	-0.003	-0.008	-0.011	-0.016	0.003	0.002	0.006	0.002
	(0.90)	(0.38)	(0.62)	(0.14)	(0.91)	(0.87)	(0.83)	(0.82)
Text reading (applies to reading-connected text)								
Supported oral reading (includes choral and round-robin reading)	-0.025	-0.021	-0.013	-0.005	-0.013	-0.020	-0.041	-0.043*
	(0.40)	(0.12)	(0.62)	(0.66)	(0.61)	(0.19)	(0.20)	(0.02)
Independent silent reading	0.020	0.021	0.020	0.019	0.010	0.026	0.006	0.012
	(0.52)	(0.16)	(0.46)	(0.10)	(0.72)	(0.22)	(0.83)	(0.39)
Independent or buddy oral reading	0.039	0.031*	0.042	0.030*	0.001	0.000	0.041	0.041*
	(0.14)	(0.02)	(0.08)	(0.01)	(0.98)	(0.98)	(0.13)	(0.02)
Teacher reads aloud	0.001	0.010	-0.004	0.007	0.014	0.016	-0.011	0.011
	(0.96)	(0.36)	(0.85)	(0.50)	(0.55)	(0.19)	(0.66)	(0.52)
Teacher reads aloud with students following along silently	-0.004	-0.015	-0.011	-0.019	0.021	0.006	0.003	-0.014
	(0.90)	(0.29)	(0.67)	(0.14)	(0.40)	(0.64)	(0.91)	(0.43)
Text not present	-0.009	-0.020	0.000	-0.014	-0.003	-0.015	-0.038	-0.030
	(0.78)	(0.16)	(0.99)	(0.24)	(0.93)	(0.29)	(0.35)	(0.22)
Text present but not being read	-0.054	-0.004	-0.042	-0.003	-0.033	-0.001	-0.041	-0.005
	(0.13)	(0.77)	(0.15)	(0.81)	(0.31)	(0.96)	(0.18)	(0.79)

Table V.3 *(continued)*

	Outcomes and Covariate Adjustment							
	Composite[a]		GRADE		Social Studies Reading Comprehension		Science Reading Comprehension	
	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates
Part II, Instruction Effectiveness[d]								
Gave inaccurate and/or confusing explanations or feedback	-0.024 (0.35)	-0.005 (0.72)	-0.016 (0.42)	-0.005 (0.66)	-0.022 (0.42)	-0.012 (0.59)	-0.021 (0.50)	-0.017 (0.37)
Missed opportunity to correct or address error	-0.044* (0.05)	-0.021 (0.22)	-0.036* (0.04)	-0.019 (0.09)	-0.032 (0.18)	-0.014 (0.55)	-0.045 (0.06)	-0.026 (0.19)
Provided opportunities for most students to participate actively during teacher-led instruction	0.005 (0.88)	0.012 (0.38)	-0.001 (0.96)	0.008 (0.45)	-0.013 (0.66)	-0.002 (0.90)	0.009 (0.76)	0.027 (0.14)
Paced instruction so that the length of the comprehension or vocabulary activities was appropriate for this age group	0.017 (0.63)	0.006 (0.64)	0.012 (0.66)	0.011 (0.27)	-0.013 (0.69)	-0.016 (0.38)	0.011 (0.73)	0.012 (0.50)
Taught using outlining and/or note taking	0.042 (0.17)	-0.004 (0.77)	0.028 (0.30)	-0.010 (0.39)	0.025 (0.39)	-0.007 (0.64)	0.052 (0.10)	0.016 (0.41)
Used graphic organizers	-0.040 (0.21)	-0.004 (0.78)	-0.032 (0.25)	0.007 (0.52)	-0.032 (0.27)	-0.004 (0.80)	-0.044 (0.18)	-0.011 (0.53)
Kept students thinking for two or more seconds before calling on a student to respond to a complex question	-0.003 (0.92)	0.014 (0.33)	-0.009 (0.75)	0.010 (0.34)	-0.022 (0.48)	0.002 (0.91)	0.001 (0.96)	0.024 (0.15)
Gave independent/pairs/small-group practice in answering comprehension questions or applying comprehension strategy(ies) with expected written product	0.051 (0.11)	0.014 (0.30)	0.043 (0.12)	0.013 (0.21)	0.037 (0.17)	0.008 (0.64)	0.049 (0.11)	0.012 (0.48)
Used writing activities in response to reading (does not include fill-in-the-blank or one-word answers)	0.071* (0.01)	0.024 (0.14)	0.058* (0.02)	0.021 (0.14)	0.073* (0.00)	0.035* (0.03)	0.062* (0.02)	0.016 (0.41)
Part II, Teachers' Management/Responsiveness to Students[e]								
Teacher maximized the amount of time available for instruction	0.081* (0.01)	0.045* (0.00)	0.059* (0.03)	0.032* (0.00)	0.060* (0.03)	0.027 (0.12)	0.105* (0.00)	0.071* (0.00)
Teacher managed student behavior effectively in order to avoid disruptions and provide productive learning environments	0.097* (0.00)	0.064* (0.00)	0.071* (0.01)	0.047* (0.00)	0.069* (0.01)	0.040* (0.02)	0.126* (0.00)	0.091* (0.00)
Teacher redirected discussion if a student response was leading the group off topic/focus	0.041 (0.36)	0.044* (0.03)	0.030 (0.41)	0.039* (0.01)	-0.001 (0.97)	-0.010 (0.60)	0.092 (0.08)	0.092* (0.00)

Table V.3 *(continued)*

	Outcomes and Covariate Adjustment							
	Composite[a]		GRADE		Social Studies Reading Comprehension		Science Reading Comprehension	
	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates	None	Pretest and Other Covariates
Part II, Student Engagement[f]								
Student engagement during the first half of the observation session	0.105*	0.037*	0.089*	0.031*	0.076*	0.010	0.121*	0.052*
	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.52)	(0.00)	(0.01)
Student engagement during the remainder of the observation session	0.105*	0.040*	0.088*	0.036*	0.064*	0.001	0.125*	0.065*
	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)	(0.95)	(0.00)	(0.01)

SOURCE: Classroom observations and reading comprehension tests administered by study team.

NOTE: Standardized regression coefficients and *p-values* (in parentheses) are reported for each ERC instrument item. The *p-values* presented in this table are adjusted for clustering of students within schools but not for multiple-hypotheses testing. The outcome for each regression is indicated by the row labels. Other covariates in the regression models vary by column. The first column for each outcome includes no additional covariates (so it is simply regressing test scores on each of the ERC instrument items). The second column includes the GRADE and TOSCRF tests administered in the fall of each cohort year, student ethnicity and race, student English language learner status, school location, and teacher race as covariates in the regression.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]For items in this pane of the table, observers recorded tallies for the number of times each behavior was observed.

[c]For items in this pane of the table, observers selected all items that they observed (more than one category could be selected).

[d]For items in this pane of the table, observers recorded "Yes" or "No" for each item.

[e]For items in this pane of the table, observers recorded a "1" for minimal/poor, "2" for fair, "3" for good, or "4" for excellent.

[f]For items in this pane of the table, observers recorded a "1" for few engaged, "2" for many engaged, or "3" for most engaged.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

C. RELATIONSHIP BETWEEN TEACHER EFFICACY AND PROFESSIONAL DEVELOPMENT AND TEST SCORES

Teachers' self-reported efficacy (Hoy and Woolfolk 1993) and hours of professional development were collected through a survey of teachers (described in Chapter 1). In this section we examine whether efficacy and professional development are correlated with students' reading comprehension test scores among teachers and students in our sample. Previous research has found mixed evidence of correlations between various measures of teacher efficacy and professional development and students' reading comprehension, as follows:

Self-Efficacy

- Ross (1994) conducted a literature review of 88 teacher self-efficacy studies, 5 of which reported significant correlations between teacher efficacy and student achievement in language-oriented subjects (reading, language arts, and social studies).
- Goddard, Hoy, and Hoy (2000) surveyed a sample of 47 elementary schools and found that a one standard deviation increase in a school's collective teacher efficacy index (as opposed to individual teacher self-efficacy) was correlated with a 0.43 standard deviation increase in a combined measure of students' second-, third-, and fifth-grade math and reading test scores.

Professional Development

- Angrist and Lavy (1998) determined that additional training in reading received by teachers in non-religious schools led to a significant improvement in fourth-grade students' reading test scores by 0.62 standard deviation.
- McCutchen et al. (2002) found that additional training in literacy instruction was associated with a 60 percent growth in first-grade reading comprehension scores after a year of reading instruction, significantly larger than the growth in scores for students whose teachers did not receive the training.
- Jacob and Lefgren (2004) found that increases in in-service training had no statistically significant effect on reading achievement for students in third through sixth grade.

Correlations between teacher efficacy and professional development hours and students' scores on reading comprehension post-tests are presented in Table V.4. The numbers in this table are the standardized coefficients on each variable from a regression of a post-test on the variables of interest and other covariates. Below each regression coefficient is a *p*-value. Teacher efficacy and professional development hours are analyzed in separate regression models (for example, the correlation between efficacy and test scores is not adjusted for professional development hours). The regression coefficients in the first column come from a regression model without any other covariates (note that the regression of test scores on professional development includes all of the professional development indicator variables shown in the table with "no professional

TABLE V.4

REGRESSION-ADJUSTED CORRELATION BETWEEN TEACHER TRAITS AND STUDENT POST-TEST SCORES

	Covariate Adjustment		
	None	Pretest	Pretest and Other Covariates
GRADE Score			
Teacher Efficacy Scale	0.06	0.03	0.02
	(0.10)	(0.07)	(0.18)
Hours of Professional Development in Reading Instruction			
1 to 8	0.07*	0.04*	0.02
	(0.04)	(0.04)	(0.20)
9 to 16	0.02	0.03	-0.005
	(0.65)	(0.22)	(0.77)
17 to 32	0.11*	0.05*	0.02
	(0.02)	(0.03)	(0.40)
33 or more	0.06	0.02	0.001
	(0.11)	(0.38)	(0.96)
Social Studies Reading Comprehension Assessment Score			
Teacher Efficacy Scale	0.06	0.04	0.00
	(0.09)	(0.11)	(0.82)
Hours of Professional Development in Reading Instruction			
1 to 8	0.05	0.03	0.01
	(0.20)	(0.36)	(0.77)
9 to 16	-0.01	0.01	-0.03
	(0.85)	(0.82)	(0.29)
17 to 32	0.07	0.04	0.00
	(0.06)	(0.14)	(0.92)
33 or more	0.05	0.03	0.02
	(0.13)	(0.33)	(0.50)
Science Reading Comprehension Assessment Score			
Teacher Efficacy Scale	0.05	0.03	0.01
	(0.16)	(0.27)	(0.51)
Hours of Professional Development in Reading Instruction			
1 to 8	0.08	0.05	0.03
	(0.06)	(0.06)	(0.23)
9 to 16	0.01	0.01	-0.02
	(0.87)	(0.70)	(0.47)
17 to 32	0.12*	0.07*	0.04
	(0.01)	(0.01)	(0.08)
33 or more	0.05	0.02	0.01
	(0.22)	(0.60)	(0.81)

TABLE V.4 (*continued*)

	Covariate Adjustment		
	None	Pretest	Pretest and Other Covariates
Composite Test Score[a]			
Teacher Efficacy Scale	0.07	0.03	0.01
	(0.09)	(0.06)	(0.31)
Hours of Professional Development in Reading Instruction			
1 to 8	0.07*	0.04*	0.03
	(0.04)	(0.05)	(0.19)
9 to 16	0.01	0.02	-0.02
	(0.86)	(0.45)	(0.42)
17 to 32	0.11*	0.06*	0.02
	(0.01)	(0.01)	(0.30)
33 or more	0.06	0.02	0.01
	(0.09)	(0.29)	(0.76)

SOURCE: Teacher survey and reading comprehension tests administered by study team.

NOTE: Standardized regression coefficients and *p-values* (in parentheses) are reported for each variable. The *p-values* presented in this table are adjusted for clustering of students within schools but not for multiple-hypotheses testing. The outcome for each regression is indicated by the pane labels (in gray shading). Other covariates in the regression models vary by column. The first column includes no additional covariates (so it is simply regressing test scores on the variable of interest—note that for professional development, all professional development indicator variables are included in the regression model with the omitted category being "no professional development"). The second column includes the GRADE and TOSCRF tests administered in the fall of each cohort year. In addition to those two tests, the third column also includes student ethnicity and race, student English language learner status, school location, and teacher race.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different from zero at the .05 level.

development" as the omitted category). The second column adds pre-test scores from the GRADE and TOSCRF as covariates. The third column adds student ethnicity and race, student English language learner status, school location, and teacher race. All variables in these regressions are standardized.

We find no evidence of a statistically significant correlation between teacher self-efficacy and reading comprehension test scores, regardless of whether regression adjustment is used. We do find a statistically significant correlation between hours of professional development and test scores in the first and second columns of the table, but this correlation disappears when we adjust for additional student, teacher, and school characteristics.

D. RELATIONSHIP BETWEEN READING TIME AND TEST SCORES

Daily time that teachers spend in reading activities and weekly time that students spend using informational text were both collected from teachers using two surveys (described in Chapter 1). In this section we examine whether daily time in reading activities and weekly time using informational text are correlated with students' reading comprehension test scores among teachers and students in our sample. Previous research has found mixed evidence of the relationship between various measures of reading time and students' achievement, as follows:

- Reutzel and Hollingsworth (1991) found that greater reading skill instruction and reading time was associated with a statistically significant increase of 16.5 percent in test scores on a criterion-referenced reading comprehension test for fourth graders.
- Anderson et al. (1988) found that additional minutes spent reading books was associated with a statistically significant increase in fifth-grade reading comprehension scores by 8.1 percentile points.
- Connor et al. (2004) found that time spent on teacher-managed explicit instruction predicted greater growth in students' reading comprehension.
- Taylor et al. (2000) found a significant correlation between K – 3 schools with higher scores on several measures of reading achievement (including reading comprehension) and the presence of teachers who spent more time on independent reading in those schools.
- Taylor et al. (1990) found that minutes of reading per day during class time increased scores on a reading comprehension subtest by a magnitude of 0.11.
- Seago-Tufaro (2002) found that additional time for independent reading made no significant difference in reading comprehension scores.

Correlations between daily time in reading activities and weekly time using informational text and reading comprehension post-tests are presented in Table V.5. The numbers in this table are the standardized coefficients on each variable in a regression of a post-test on the variables of interest and other covariates. Below each regression coefficient is a *p*-value. The regression coefficients in the first column come from a regression model without any other covariates. The

TABLE V.5

REGRESSION-ADJUSTED CORRELATION BETWEEN TIME DEVOTED TO READING INSTRUCTION AND POST-TEST SCORES

	Covariate Adjustment		
	None	Pretest	Pretest and Other Covariates
GRADE Score			
Time in reading activities (daily)[a]	-0.04	-0.01	-0.01
	(0.21)	(0.71)	(0.47)
Time using informational text (weekly)[b]	-0.03	0.00	-0.01
	(0.62)	(0.95)	(0.43)
Social Studies Reading Comprehension Assessment Score			
Time in reading activities (daily)[a]	-0.08*	-0.04	-0.04
	(0.05)	(0.15)	(0.10)
Time using informational text (weekly)[b]	0.00	0.02	0.01
	(0.95)	(0.65)	(0.78)
Science Reading Comprehension Assessment Score			
Time in reading activities (daily)[a]	-0.04	-0.01	-0.02
	(0.26)	(0.57)	(0.53)
Time using informational text (weekly)[b]	-0.02	0.00	-0.04
	(0.70)	(0.95)	(0.18)
Composite Test Score[c]			
Time in reading activities (daily)[a]	-0.06	-0.02	-0.03
	(0.14)	(0.41)	(0.21)
Time using informational text (weekly)[b]	-0.02	0.01	-0.01
	(0.72)	(0.80)	(0.53)

SOURCE: Teacher Survey; reading comprehension tests administered by study team.

NOTE: Standardized regression coefficients and *p-values* (in parentheses) are reported for each variable. The *p-values* presented in this table are adjusted for clustering of students within schools but not for multiple-hypotheses testing. The outcome for each regression is indicated by the pane labels (in gray shading). Other covariates in the regression models vary by column. The first column includes no additional covariates (so it is simply regressing test scores on the variable of interest). The second column includes the GRADE and TOSCRF tests administered in the fall of each cohort year. In addition to those two tests, the third column also includes student ethnicity and race, student English language learner status, school location, and teacher race.

[a]This variable is the number of minutes spent each day in any of the following activities: (1) Separate Instruction Using Intervention Curriculum (CRISS, ReadAbout, and Read for Real); (2) Core (Basal) Reading Curriculum; (3) Supplemental Reading Curriculum (other than the study interventions) focused on comprehension, vocabulary, or fluency; (4) Reading Lesson Using Fiction Materials; (5) Reading Lesson Using Nonfiction Materials; and (6) Other Language Arts Activity.

[b]This variable is the number of minutes of class time that teachers reported students spent using informational text in a typical week.

TABLE V.5 (*continued*)

---

[c]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different from zero at the .05 level.

second column adds pre-test scores from the GRADE and TOSCRF as covariates. The third column adds student ethnicity and race, student English language learner status, school location, and teacher race. As above, all variables in these regressions are standardized.

We see no evidence of positive correlations between time teachers report spending on reading instruction each day and test scores. There is one statistically significant negative correlation between daily time spent in reading activities and the social studies reading comprehension test, but this relationship is not statistically significant after adjusting for pre-test scores. There were no statistically significant correlations between time spent in a typical week using informational text and test scores.

E. CORRELATION OF IMPACTS AND SCHOOL CHARACTERISTICS

As described in detail in Appendix A, random assignment of schools to treatment and control groups was conducted within blocks of similar schools. In this section, we examine the correlations between regression-adjusted, block-level impacts and the average characteristics of the schools in the blocks that are involved in the calculation of each impact. For example, there were 13 blocks for which an impact of Project CRISS was calculated. For each of those 13 blocks, we calculate the impact of Project CRISS and the average characteristics of all the schools in that block that were either assigned to Project CRISS or the control group. We then calculate the correlation between those 13 impacts and the average characteristics of the 13 blocks of schools. Given the small number of blocks, statistical power for these correlations is limited.

These exploratory correlations are intended to inform future research and development of programs and must be interpreted cautiously. A statistically significant correlation between block-level impacts and block characteristics should not be confused with a statistically significant impact of a curriculum. There are too few schools within each block to calculate the statistical significance of block-specific impacts, so none of the block-specific impacts can be characterized as statistically significant. Furthermore, the correlation between block-level impacts and block characteristics could be affected by variation across blocks in unobserved characteristics that are correlated with the observed characteristics. See Chapter III for findings on the experimental impacts of these programs.

Correlations are presented in Table V.6. The block characteristics included in the table are the racial and ethnic makeup of the schools, the percentage of students receiving free or reduced-price lunch, and the percentage of students classified as ELL. We observe the following statistically significant correlations:

- A negative correlation of -0.64 between impacts of Project CRISS and the percentage of black students in a school.

Recall that the analyses presented in this chapter are not experimental and do not support causal conclusions. Therefore, findings from these analyses should be interpreted cautiously.

TABLE V.6

CORRELATION COEFFICIENTS BETWEEN BLOCK-LEVEL TEST SCORE IMPACTS AND BLOCK-LEVEL MEANS OF SCHOOL CHARACTERISTICS

Block Characteristic	Project CRISS	ReadAbout	Read for Real
Composite Test Score[a]			
Percentage of Hispanic Students	0.29	0.04	-0.14
	(0.33)	(0.89)	(0.71)
Percentage of White Students	0.10	0.12	-0.06
	(0.73)	(0.71)	(0.87)
Percentage of Black Students	-0.36	-0.17	0.14
	(0.22)	(0.60)	(0.71)
Percentage of FRPL Students	-0.20	-0.27	0.24
	(0.52)	(0.39)	(0.53)
Percentage of ELL Students	0.32	-0.16	-0.13
	(0.28)	(0.62)	(0.73)
GRADE Score			
Percentage of Hispanic Students	0.18	-0.02	0.10
	(0.57)	(0.94)	(0.80)
Percentage of White Students	-0.07	0.20	-0.06
	(0.83)	(0.53)	(0.87)
Percentage of Black Students	-0.22	-0.16	-0.05
	(0.47)	(0.61)	(0.90)
Percentage of FRPL Students	-0.18	-0.15	0.15
	(0.56)	(0.64)	(0.70)
Percentage of ELL Students	0.27	-0.22	0.11
	(0.38)	(0.50)	(0.78)
Social Studies Reading Comprehension Assessment Score			
Percentage of Hispanic Students	0.40	-0.19	-0.60
	(0.18)	(0.56)	(0.09)
Percentage of White Students	0.40	-0.11	0.09
	(0.18)	(0.73)	(0.81)
Percentage of Black Students	-0.64*	0.27	0.42
	(0.02)	(0.40)	(0.26)
Percentage of FRPL Students	-0.51	0.02	0.29
	(0.08)	(0.94)	(0.45)
Percentage of ELL Students	0.40	-0.30	-0.59
	(0.18)	(0.35)	(0.09)

TABLE V.6 (*continued*)

Block Characteristic	Project CRISS	ReadAbout	Read for Real
Science Reading Comprehension Assessment Score			
Percentage of Hispanic Students	0.01	0.27	-0.14
	(0.97)	(0.40)	(0.71)
Percentage of White Students	0.21	0.09	0.06
	(0.49)	(0.78)	(0.88)
Percentage of Black Students	-0.06	-0.38	0.12
	(0.84)	(0.22)	(0.76)
Percentage of FRPL Students	0.06	-0.52	0.13
	(0.85)	(0.08)	(0.74)
Percentage of ELL Students	-0.04	0.07	-0.18
	(0.90)	(0.83)	(0.64)

SOURCE: School Information Form, 2005–2006 Common Core of Data (CCD); reading comprehension tests administered by study team.

NOTE: Schools were randomly assigned to treatment or control groups within blocks of similar schools. For the second cohort, there were 13 blocks with CRISS schools, 12 blocks with ReadAbout schools, and 9 blocks with Read for Real Schools. The numbers reported are the correlations between regression-adjusted block-level impacts and the average characteristics of the schools in the blocks that are involved in the calculation of each impact. Numbers in parentheses are the *p-values* for these correlations. Regression-adjusted impacts were calculated taking into account pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, whether students were overage for grade, teacher sex, teacher age, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

ELL = English language learners; FRPL = free or reduced-price lunch; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

This page is intentionally left blank.

VI. SUMMARY

This study used a rigorous experimental design to assess the effects of four reading comprehension curricula on reading comprehension among fifth-grade students in selected districts across the country. All four curricula were included in the first year of the study and in the sixth-grade component of the study's second year, and three of the four curricula were included in the fifth-grade component of the study's second year. Consistent with the study's focus on schools serving low-income students, the districts and schools that the study team targeted—and that agreed to participate in the study—had above-average poverty levels, and were larger and more urban, on average, than districts and schools in the United States.

The key findings from the second year of the study are as follows:

Implementation Findings

- **During summer and early fall 2007, 50 to 91 percent of treatment teachers were trained to use the curricula.** Fifty percent of Read for Real teachers, 89 percent of Project CRISS teachers, and 91 percent of ReadAbout teachers were trained in the use of the curricula.

- **In the spring of the second year of the study, over 80 percent (83 to 96 percent) of treatment teachers reported using their assigned curriculum.** Eighty-three percent of Read for Real teachers, 92 percent of Project CRISS teachers, and 96 percent of ReadAbout teachers reported using their assigned curriculum. The percentage of teachers who reported using each of the three interventions did not differ significantly between the first and second years.

- **Classroom observation data from the second year of intervention implementation showed that teachers implemented 65 to 94 percent of the behaviors deemed important by the developers for implementing each curriculum.** Project CRISS and ReadAbout teachers implemented, on average, 65 and 94 percent of such behaviors, respectively, and Read for Real teachers implemented 75 and 76 percent of the behaviors deemed important for the two types of instructional days that are part of that curriculum. There were no statistically significant differences in average fidelity levels between the first and second study years.[76]

[76]The fidelity levels reported for ReadAbout and Read for Real are based on fidelity form behaviors that fell within a window observed by the study's classroom observers. The fidelity levels reported for Project CRISS are based on all behaviors on the CRISS fidelity form. See Chapter II for more information.

Findings on Intervention Effectiveness

- **The curricula did not have an impact on students one year after the end of their implementation.** In the second year, after the first cohort of students was no longer using the interventions, there were no statistically significant impacts of any of the four curricula. (In the first year, a statistically significant negative impact of Reading for Knowledge was observed for the first cohort of students.)

- **Impacts were not statistically significantly larger after *schools* had one year of experience using the curricula.** Impacts for the second cohort of students (who attended schools that had one prior year of experience using the study curricula) were not statistically significantly different from zero or from the impacts for the first cohort of students. (Treatment students in the *second* cohort attended schools that had one prior year of experience using the study curricula, while treatment students in the *first* cohort attended schools with no prior experience using the study curricula. Reading for Knowledge was not implemented with the second cohort of students.)

- **The impact of one of the curricula (ReadAbout) was statistically significantly larger after *teachers* had one year of experience using the curricula.** There was a positive, statistically significant impact of ReadAbout on the social studies reading comprehension assessment for second-cohort students taught by teachers who were in the study both years (effect size: 0.22). This impact was statistically significantly larger than the impact for first cohort of students taught by the same teachers in the first year of the study.[77]

Findings on the Effectiveness of the Interventions for Subgroups of Students

- **The curricula did not have differential impacts on fifth-grade post-test scores for most Cohort 2 student subgroups (282 of 288).** Statistically significantly greater impacts were observed for Project CRISS students scoring in the top third of the pre-test GRADE distribution, for Read for Real students classified as ELL or taught by teachers with below-median teaching efficacy, for ReadAbout students in schools with below-median School Professional Culture scores, and for combined treatment group students taught by teachers below the median efficacy level or in schools with below-median School Professional Culture scores. All of these findings have a causal interpretation—with the exception of the teaching efficacy and School Professional Culture subgroup findings—because the subgroups were formed using characteristics observed at the beginning of the study's implementation year.

- **The curricula did not have differential impacts on sixth-grade follow-up scores for most Cohort 1 student subgroups (352 of 360).** Statistically significantly greater impacts were observed for Project CRISS students scoring in the bottom third of the pre-test TOSCRF distribution or scoring below the TOSCRF national norm sample

[77]This is similar to the impact observed after *schools* had one year of experience with the curricula, but that impact was not statistically significant (*p*-value: .053).

average, for ReadAbout students scoring in the middle third of the pre-test TOSCRF distribution, and for combined treatment group students (1) scoring in the bottom third of the pre-test TOSCRF distribution, (2) scoring below the TOSCRF national norm sample average, (3) not classified as ELL, or (4) scoring in the bottom third of the pre-test GRADE distribution.

This page is intentionally left blank.

REFERENCES

Adams, A., Carnine, D., and Gersten, R. (1982). Instructional Strategies for Studying Content Area Texts in the Intermediate Grades. *Reading Research Quarterly, 18*, 27-55.

Adams, R.J., Wilson, M., and Wang, W.C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1-23.

Anderson, R.C., Wilson, P.T., and Fielding, L.G. (1988). Growth in Reading and How Children Spend Their Time Outside of School. *Reading Research Quarterly*, *23*(3), 285-303.

Anderson, V., and Roit, M. (1993). Planning and Implementing Collaborative Strategy Instruction for Delayed Readers in Grades 6-10. *The Elementary School Journal*, *94*(2) (Special Issue: Strategies Instruction), 121-137.

Angrist, J.D., and Lavy, V. (1988). Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools. National Bureau of Economic Research, Working Paper 6781. Washington, DC: National Bureau of Economic Research.

Baumann, J.F. (1984). The Effectiveness of a Direct Instruction Paradigm for Teaching Main Idea Comprehension. *Reading Research Quarterly*, *20*(1), 93-115.

Baumann, J.F., and Bergeron, B.S. (1993). Story Map Instruction Using Children's Literature: Effects on First Graders' Comprehension of Central Narrative Elements. *Journal of Reading Behavior, 25*(4), 407-437.

Borg, W.R. (1980). Time and School Learning. In C. Denham and A. Lieberman (Eds.), *Time to Learn,* (pp. 33-72). Washington, DC: National Institute of Education.

Brophy, J., and Evertson, C. (1976). *Learning from Teaching: A Developmental Perspective*. Boston, MA: Allyn and Bacon.

Brown, A.L., and Day, J.D. (1983). Macrorules for Summarizing Text: The Development of Expertise. *Journal of Verbal Learning and Verbal Behavior, 22*, 1-14.

Brown, R., Pressley, M., Van Meter, P., and Schuder, T. (1996). A Quasi-Experimental Validation of Transactional Strategies Instruction with Low-Achieving Second-Graders. *Journal of Educational Psychology, 88*, 18-37.

Carlisle, J. (2003). Teacher's QUEST: Self-Administered Questionnaire. Ann Arbor, MI: Regents of the University of Michigan.

Carlisle, J., and Rice, M. (2002). *Improving Reading Comprehension: Research-Based Principles and Practices*. Baltimore, MD: York Press.

Chall, J. (1983). *Stages of Reading Development.* Fort Worth, TX: Harcourt-Brace.

Chromy, J.R. (1979). Sequential Sample Selection Methods. *Proceedings of the American Statistical Association, Survey Research Methods Section.* 401–406.

Connor, C.M., Morrison, F.J., and Petrella, J.N. (2004). Effective Reading Comprehension Instruction: Examining Child by Instruction Interactions. *Journal of Educational Psychology, 96*(4), 682-698.

Consortium on Chicago School Research. (1999). Improving Chicago's Schools: The Teachers' Turn, 1999; Elementary School Teacher Survey, 1999. Chicago: CCSR, 1999. Retrieved from http://www.consortium-chicago.org.

Cooley, W.W., and Leinhardt, G. (1980). The Instructional Dimensions Study. *Educational Evaluation and Policy Analysis, 2*, 7-25.

Crawford, L.W., Martin, C.E., and Philbin, M.M. (2005). *Read for Real: Nonfiction Strategies for Reading Results*. Columbus, OH: Zaner-Bloser.

Darch, C., and Gersten, R. (1986). Direction Setting Activities in Reading Comprehension: A Comparison of Two Approaches. *Learning Disabilities Quarterly*, *9*(3), 235-243.

Darch, C., and Kame'enui, E. (1987). Teaching LD Students Critical Reading Skills: A Systematic Replication. *Learning Disability Quarterly*, *10*, 82-91.

Denham, C., and Lieberman, A. (eds.). (1980). *Time to Learn.* Washington, DC: National Institute of Education.

Desimone, L. (2002). How Can Comprehensive School Reform Models Be Successfully Implemented? *Review of Educational Research, 72*, 433-479.

Duffy, G.G., Roehler, L.R., Sivan, E., Rackliffe, G., Book, C., Meloth, M.S., Vavrus, L.G., Wesselman, R., Putnam, J., and Bassiri, D. (1987). Effects of Explaining the Reasoning Associated with Using Reading Strategies. *Reading Research Quarterly, 23*, 347-386.

Duke, N.K., and Pearson, P.D. (2002). Effective Practices for Developing Reading Comprehension. In A.E. Farstrup and S.J. Samuels (Eds.), *What Research Has to Say About Reading Instruction (Third Edition),* (pp. 205-242). Newark, DE: International Reading Association.

Dunnett, C.W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association, 50,* 1096-1121.

Durkin, D. (1978-1979). What Classroom Observations Reveal About Reading Comprehension Instruction. *Reading Research Quarterly, 14*(4), 481-533.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., Means, B., Murphy, N., Penuel, W., Javitz, H., Emery, D., and Sussex, W. (2007). Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Educational Testing Service. (2007a). *Science Reading Comprehension Assessment* (unpublished). Princeton, NJ: ETS.

Educational Testing Service. (2007b). *Social Studies Reading Comprehension Assessment* (unpublished). Princeton, NJ: ETS.

Gersten, R., Baker, S., and Lloyd, J.W. (2000). Designing High Quality Research in Special Education: Group Experimental Design. *Journal of Special Education, 34,* 2-18.

Gersten, R., Fuchs, L.S., Compton, D., Coyne, M., Greenwood, C., and Innocenti, M.S. (2005). Quality Indicators for Group Experimental and Quasi-experimental Research in Special Education. *Exceptional Children, 71*, 149-164.

Gersten, R., Fuchs, L., Williams, J., and Baker, S. (2001). Teaching Reading Comprehension Strategies to Students with Learning Disabilities. *Review of Educational Research*, *71,* 279-320.

Gibson, S., and Dembo, M.H. (1984). Teacher Efficacy: A Construct Validation. *Journal of Educational Psychology, 76,* 569-582.

Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Lugo-Gil, J., Grider, M., Britton, E., and Ali, M. (2008). Impacts of Comprehensive Teacher Induction: Results From the First Year of a Randomized Controlled Study. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S., and Myers, D. (2004). Assessing the Effectiveness of Education Interventions: Issues and Recommendations for the Title I Evaluation. Washington, DC: Mathematica Policy Research.

Goddard, R.D., Hoy, W.K., and Hoy, A.W. (2000). Collective Teacher Efficacy: Its Meaning, Measure, and Impact on Student Achievement. *American Education Research Journal*, *37*(2), 479-507.

Guthrie, J.T., Cox, K.E., Anderson, E., Harris, K., Mazzoni, S., and Rach, L. (1998). Principles of Integrated Instruction for Engagement in Reading. *Educational Psychology Review, 10*(2), 177-199.

Guthrie, J.T., Shafer, W.D., Von Secker, C., and Alban, T. (2000a). Contributions of Integrated Reading Instruction and Text Resources to Achievement and Engagement in a Statewide School Improvement Program. *Journal of Educational Research, 93*, 211-226.

Guthrie, J.T., Wigfield, A., and Von Secker, C. (2000b). Effects of Integrated Instruction on Motivation and Strategy Use in Reading. *Journal of Educational Psychology, 92*(2), 331-341.

Hammill, D., Wiederholt, J., and Allen, E. (2006). *Test of Silent Contextual Reading Fluency (TOSCRF), Examiner's Manual*. Austin, TX: PRO-ED, Inc.

Hare, V.C., and Borchardt, K.M. (1984). Direct Instruction in Summarization Skills. *Reading Research Quarterly, 20*(1), 62-78.

Hart, B., and Risley, T.R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Brooks.

Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346-363.

Hoy, W.K., and Woolfolk, A.E. (1993). Teachers' Sense of Efficacy and the Organizational Health of Schools. *Elementary School Journal, 93,* 355-372.

Ingersoll, R. Holes in the Teacher Supply Bucket. *The School Administrator, 59*(3), 42-43.

Jacob, B.A., and Lefgren, L. (2004). The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago. *The Journal of Human Resources*, *39*(1), 50-79.

James-Burdumy, S., Myers, D., Deke, J., Mansfield, W., Gersten, W., Dimino, J., Dole, J., Liang, L., Vaughn, S., and Edmonds, M. (2006). The National Evaluation of Reading Comprehension Interventions: Design Report. Final report submitted to the U.S. Department of Education. Princeton, NJ: Mathematica Policy Research.

James-Burdumy, S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., Douglas, A., Gersten, R., Newman-Gonchar, R., Dimino, J., and Faddis, B. (2009). Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students (NCEE 2009-4032). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Jones, M.P. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*, *91*(433), 222-230.

Klinger, J.K, Vaughn, S., and Shay Schumm, J. (1998). Collaborative Strategic Reading During Social Studies in Heterogeneous Fourth-Grade Classrooms. *Elementary School Journal*, *99,* (1), 3-22.

Levin, H.M., and McEwan, P.J. (2001). *Cost-Effectiveness Analysis: Methods and Applications.* Second edition. Thousand Oaks, CA: Sage.

Liang, L.A., and Dole, J.A. (2006). Help with Reading Comprehension: Comprehension Instructional Frameworks. *The Reading Teacher, 58,* 2-13.

Linacre, J.M. (2006). *Winsteps (Version 3.61.2).* Computer software. Chicago: Winsteps.com.

Lloyd, J., Cullinan, D., Heins, E., and Epstein, M. (1980). Direct Instruction: Effects on Oral and Written Language Comprehension. *Learning Disabilities Quarterly*, *3,* 70-76.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M., and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Madden, N.A., and Crenson, V. (2006). *Reading for Knowledge.* Baltimore, MD: Success for All Foundation.

McCutchen, D., Abbott, R.D., Green, L.B., Beretvas, S.N., Cox, S., Potter, N.S., Quiroga, T., and Gray, A.L. (2002). Beginning Literacy: Links Among Teacher Knowledge, Teacher Practice, and Student Learning. *Journal of Learning Disabilities*, *35*(1), 69-86.

Martin, V.L., and Pressley, M. (1991). Elaborative-Interrogation Effects Depend on the Nature of the Question. *Journal of Educational Psychology, 83,* 113-119.

Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47,* 149-174.

Masters, G.N., and Wright, B.D. (1997). The Partial Credit Model. In W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, (pp. 101-121). New York: Springer-Verlag.

Moats, L. (1999). *Teaching Reading Is Rocket Science*. Washington, DC: American Federation of Teachers.

National Center for Education Statistics. (2009). *Common Core of Data, Public Elementary/Secondary School Universe Survey: School Year 2006-07, October 2008.* Retrieved January 12, 2009, from http://nces.ed.gov/ccd/.

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.) Washington, DC: U.S. Government Printing Office.

Nunnally, J.C., and Bernstein, I.H. (1994). *Psychometric Theory. Third Edition.* New York: McGraw-Hill, Inc.

O'Donnell, C. L. (2008). Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K-12 Curriculum Intervention Research. *Review of Educational Research, 78,* 33-84.

Palincsar, A.S., and Brown, A.L. (1984). Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities. *Cognition and Instruction, 2,* 117-175.

Patching, W., Kame'enui, E., Carnine, D., Gersten, R., and Colvin, G. (1983). Direct Instruction in Critical Reading. *Reading Research Quarterly*, *18,* 406-418.

Pearson, P.D., and Dole, J.A. (1987). Explicit Comprehension Instruction: A Review of Research and a New Conceptualization of Instruction. *Elementary School Journal, 88,* 151-165.

Pearson, P.D., and Fielding, L. (1991). Comprehension Instruction. In R. Barr, M.L. Kamil, P. Mosenthal, and P. Mosenthal (Eds.), *Handbook of Reading Research*, *Volume II* (pp. 815-860). Mahwah, NJ: Lawrence Erlbaum.

Pearson, P.D., Roehler, L.R. Dole, J.A., and Duffy, G.G. (1992). Developing Expertise in Reading Comprehension. In S.J. Samuels and A.E. Farstrup (Eds.), *What Research Has to Say About Reading Instruction (Second Edition)* (pp. 145-199). Newark, DE: International Reading Association.

Pressley, M. (2002). Comprehension Strategies Instruction: A Twentieth Century Report. In C.C. Block and M. Pressley (Eds.), *Comprehension Instruction: Research-Based Best Practices,* (pp. 11-27). New York: Guilford Press.

Pressley, M. (1998). *Reading Instruction That Works: The Case for Balanced Teaching.* New York: Guilford.

Pressley, M. (2000). What Should Comprehension Instruction Be the Instruction of? In M. Kamil, P. Mosenthal, P.D. Pearson, and R. Barr (Eds.), *Handbook of Reading Research*, *Volume III* (pp. 545-562). Mahwah, NJ: Erlbaum.

RAND Reading Study Group. (2000). *Reading for Understanding: Toward an R&D Program in Reading Comprehension.* Washington, DC: Office of Educational Research and Improvement.

Raphael, T.E., and Pearson, P.D. (1985). Increasing Students' Awareness of Sources of Information for Answering Questions. *American Educational Research Journal*, *22,* 217-235.

Renninger, K.A., Hidi, S., and Krapp, A. (eds.). (1992). *The Role of Interest in Learning and Development.* Hillsdale, NJ: Erlbaum.

Reutzel, D.R., and Hollingsworth P.M. (1991). Reading Time in School: Effect on Fourth Graders' Performance on a Criterion-Referenced Comprehension Test. *Journal of Educational Research*, *84*(3), 170-176.

Rosenshine, B. (1980). How Time Is Spent in Elementary Classrooms. In C. Denham and A. Lieberman (Eds.), *Time to Learn* (pp. 107-126). Washington, DC: National Institute of Education.

Rosenshine, B., and Meister, C. (1994). Reciprocal Teaching: A Review of the Research. *Review of Educational Research, 64*(4), 479-530.

Rosenshine, B., Meister, C., and Chapman, S. (1996). Teaching Students to Generate Questions: A Review of the Intervention Studies. *Review of Educational Research*, *66*(2) 181-221.

Rosenshine, B., and Stevens, R. (1986). Teaching Functions. In M. Wittrock (Ed.), *Handbook of Research on Teaching, Third Edition* (pp. 376-391). New York: Macmillan.

Ross, J.A. (1994). Beliefs That Make a Difference: The Origins and Impacts of Teacher Efficacy. Paper presented at the annual meeting of the Canadian Association for Curriculum Studies.

Santa, C.M., Havens, L.T., and Valdes, B.J. (2004). *Project CRISS: Creating Independence through Student-Owned Strategies* (3rd ed.). Dubuque, IA: Kendall/Hunt Publishing.

Seago-Tufaro, C. (2002). The Effects of Independent Reading on Oral Reading Fluency and Comprehension (EDRS 463 553). M.A. Research Project, Kean University, 2002.

Schochet, P.Z. (2008). Guidelines for Multiple Testing in Impact Evaluations of Educational Interventions. Final report submitted to the U.S. Department of Education. Princeton, NJ: Mathematica Policy Research.

Scholastic. (2005). *ReadAbout: The Personal Reading Coach for Every Student.* New York: Scholastic.

Schraw, G., Bruning, R., and Zosvoboa, C. (1995). Source of Situational Interest. *Journal of Reading Behavior, 27,* 1-17.

Shany, M.T., and Biemiller, A. (1995). Assisted Reading Practice: Effects on Performance for Poor Readers in Grades 3 and 4. *Reading Research Quarterly*, *30*(3) 382-395.

Snow, C.E. (2002). *Reading for Understanding: Toward a Research and Development Program in Reading Comprehension.* Santa Monica, CA: RAND.

Snow, C.E., and Biancarosa, G. (2003). *Adolescent Literacy and the Achievement Gap: What Do We Know and Where Do We Go From Here?* New York: Carnegie Corporation of New York.

Sparks, G.M. (1988). Teachers' Attitudes Toward Change and Subsequent Improvements in Classroom Teaching. *Journal of Educational Psychology, 80,* 111-117.

Stallings, J. (1975). Implementation and Child Effects of Teaching Practices in Follow Through Classrooms. *Monographs of the Society for Research in Child Development, 40*(163), 7-8.

Taylor, B.M., and Beach, R.W. (1984). The Effects of Text Structure Instruction on Middle-Grade Students' Comprehension and Production of Expository Prose. *Reading Research Quarterly*, *19,* 134-136.

Taylor, B.M., Frye, B.J., and Maruyama, G.M. (1990). Time Spent Reading and Reading Growth. *American Educational Research Journal*, *27*(2), 351-362.

Taylor, B.M., Pearson, D.P., Clark, K., and Walpole, S. (2000). Effective Schools and Accomplished Teachers: Lessons About Primary-grade Reading Instruction in Low-income Schools. *The Elementary School Journal*, *101*(2), 121-165.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2007). *National Assessment of Educational Progress (NAEP), 2007 Reading Assessments.* Retrieved from http://nationsreportcard.gov/reading_2007/r0003.asp.

Williams, K.T. (2001). *Group Reading Assessment and Diagnostic Evaluation (GRADE) Technical Manual.* Circle Pines, MN: American Guidance Service, Inc.

Wood, E., Pressley, M., and Winne, P.H. (1990). Elaborative Interrogation Effects on Children's Learning of Factual Content. *Journal of Educational Psychology, 82,* 741-748.

Wright, B.D., and Linacre, J.M. (1994). Reasonable Mean-square Fit Values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B.D., and Stone, M.H. (1979). *Best Test Design.* Chicago: MESA.

Wu, M.L., Adams, R.J., Wilson, M.R., and Haldane, S.A. (2007). *ACER ConQuest, version 2.0.* Computer software. Victoria, Australia: ACER Press.

APPENDIX A

RANDOM ASSIGNMENT

This page is intentionally left blank.

Random assignment was conducted to ensure that the estimated impacts of the interventions could be attributed to the interventions and not to other factors. The random assignment method used was designed to ensure an even distribution of the interventions overall and within each school district. Schools, not teachers, were randomly assigned due to concerns about the potential for contamination of control group teachers that could arise if teachers randomly assigned to treatment and control status were working within the same schools.

Random assignment of schools (conducted prior to the 2006-2007 school year) was carried out within school districts, and, whenever possible, within blocks of schools formed in each district based on baseline reading scores in participating schools.[78] Random assignment within districts helped to ensure that each treatment group was represented in each district. Conducting random assignment within blocks of schools in each district avoided the possibility of a "bad draw"—a situation in which all the schools with high (or low) baseline reading scores might be assigned to one of the study's five arms (four treatment and one control).[79]

Two different methods were used to form blocks of schools. The first method—explicit blocking—was generally used when the number of schools within a district was a multiple of five. The second method—implicit blocking—was generally used when the number of schools was not a multiple of five.

In explicit blocking, the study team formed two groups or blocks of schools, and then conducted random assignment within those blocks. For example, in a district with 10 schools, two blocks of 5 schools were formed where the schools in each block had similar baseline reading achievement levels. Random assignment was then conducted separately within those two blocks. This resulted in one school from each block being assigned to each of the five arms of the study (and, overall, two schools assigned to each of the five study arms).

When the blocked experimental design was not possible, implicit ordering through a modified Chromy selection procedure was implemented (Chromy 1979). This modified procedure ordered schools within districts based on baseline reading scores, and then the curricula were randomly assigned to the ordered list of schools to achieve an approximate balance in both baseline scores in each study arm and the number of times each intervention appeared overall.

The treatment and control statuses of schools and students participating in the second year of the study were based on the random assignment conducted prior to the first year of the study. In particular, schools participating in the fifth-grade component of the study's second year were in the same treatment or control group in the second year as in the first year. Students in the study's sixth-grade component were classified according to their treatment status from the study's first

[78]In one district, blocks were formed based on magnet school status, as that district had five participating schools that were regular schools and five participating schools that were magnet schools.

[79]Another factor we considered when conducting the random assignment was the desire to have at least two control schools in each district, so that impacts for that district could still be estimated even if one of the control schools dropped out of the study.

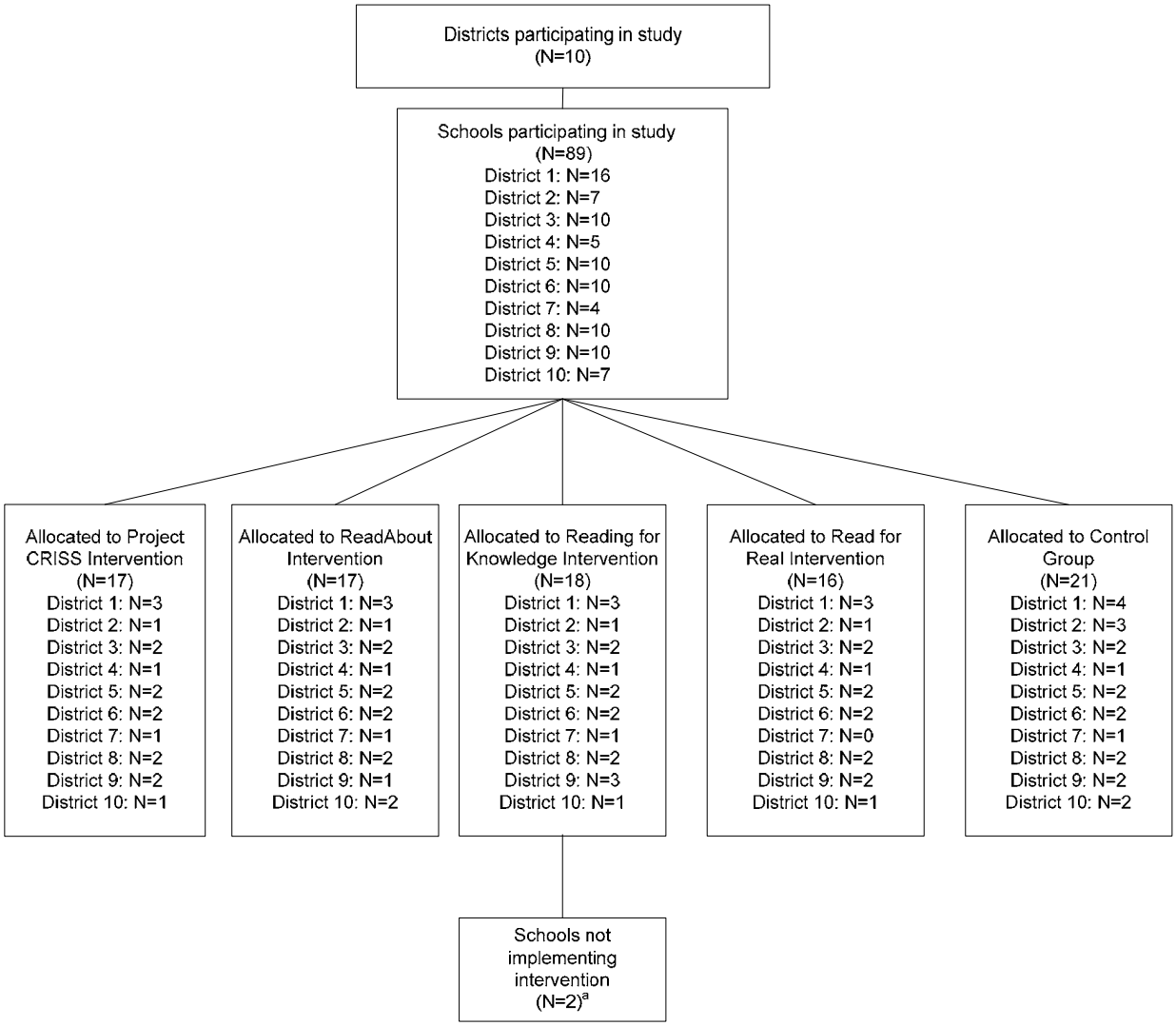
year. For example, students who attended Read for Real schools in the study's first year are in the Read for Real group in the analyses for the study's sixth-grade component, regardless of the school they attended in the study's second year. Likewise, students who attended control schools in the study's first year are in the control group for the analyses of the study's sixth-grade component. This allows the study team to assess the effects of the single year of curricula implementation provided to students in the first year of the study.

APPENDIX B

FLOW OF SCHOOLS AND STUDENTS THROUGH THE STUDY

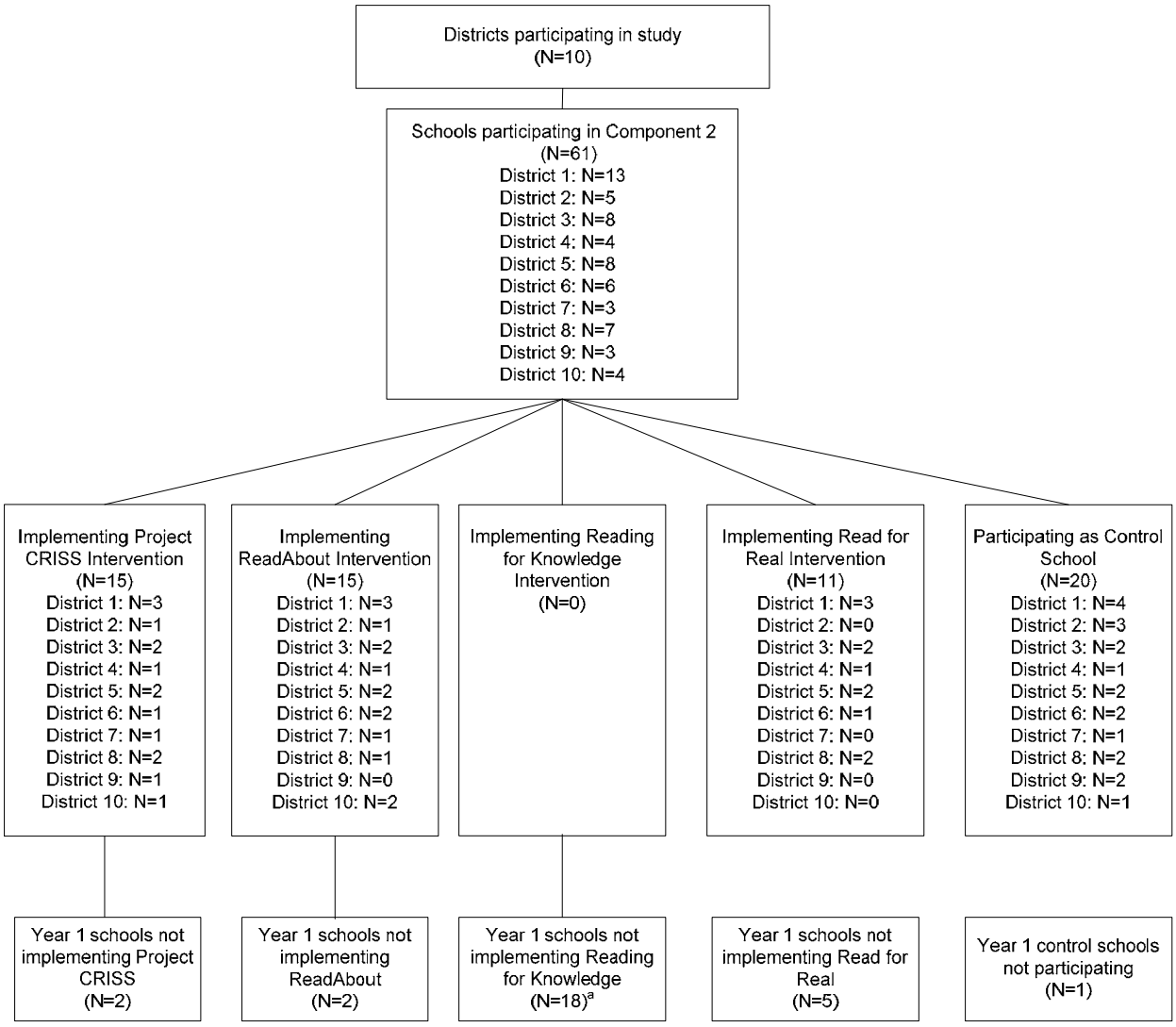
This page is intentionally left blank.

TABLE B.1
FLOW OF SCHOOLS THROUGH STUDY
Year 1, Cohort 1, Grade Five



[a]One school in District 5 stopped implementing the intervention early in the school year when the only teacher who attended training discontinued using the program. One school in District 7 never implemented the program after teachers were trained; the school said its schedule could not accommodate the required 45 minutes of instructional time. Follow-up data collection was conducted in both of these schools.

TABLE B.1a
FLOW OF SCHOOLS THROUGH STUDY
Year 2, Cohort 2, Grade Five



[a]Reading for Knowledge was not included in the fifth-grade component of the second year of the study.

TABLE B.1b
FLOW OF SCHOOLS THROUGH STUDY
Year 2, Cohort 1, Grade Six, Follow Up

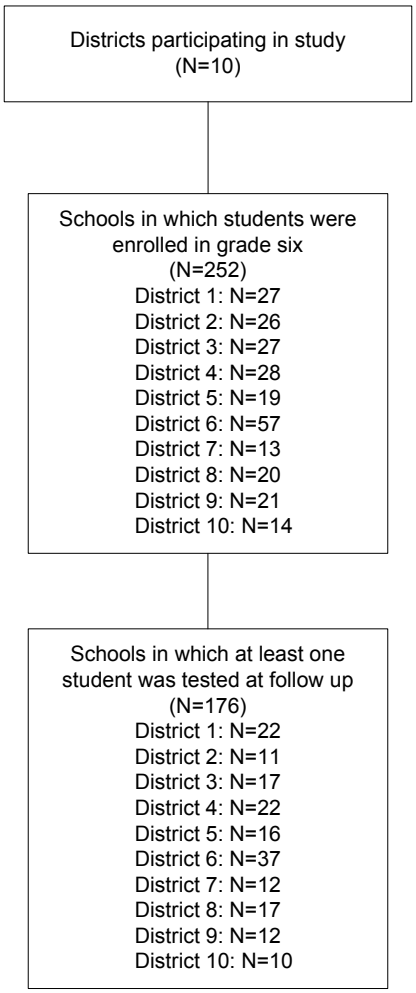
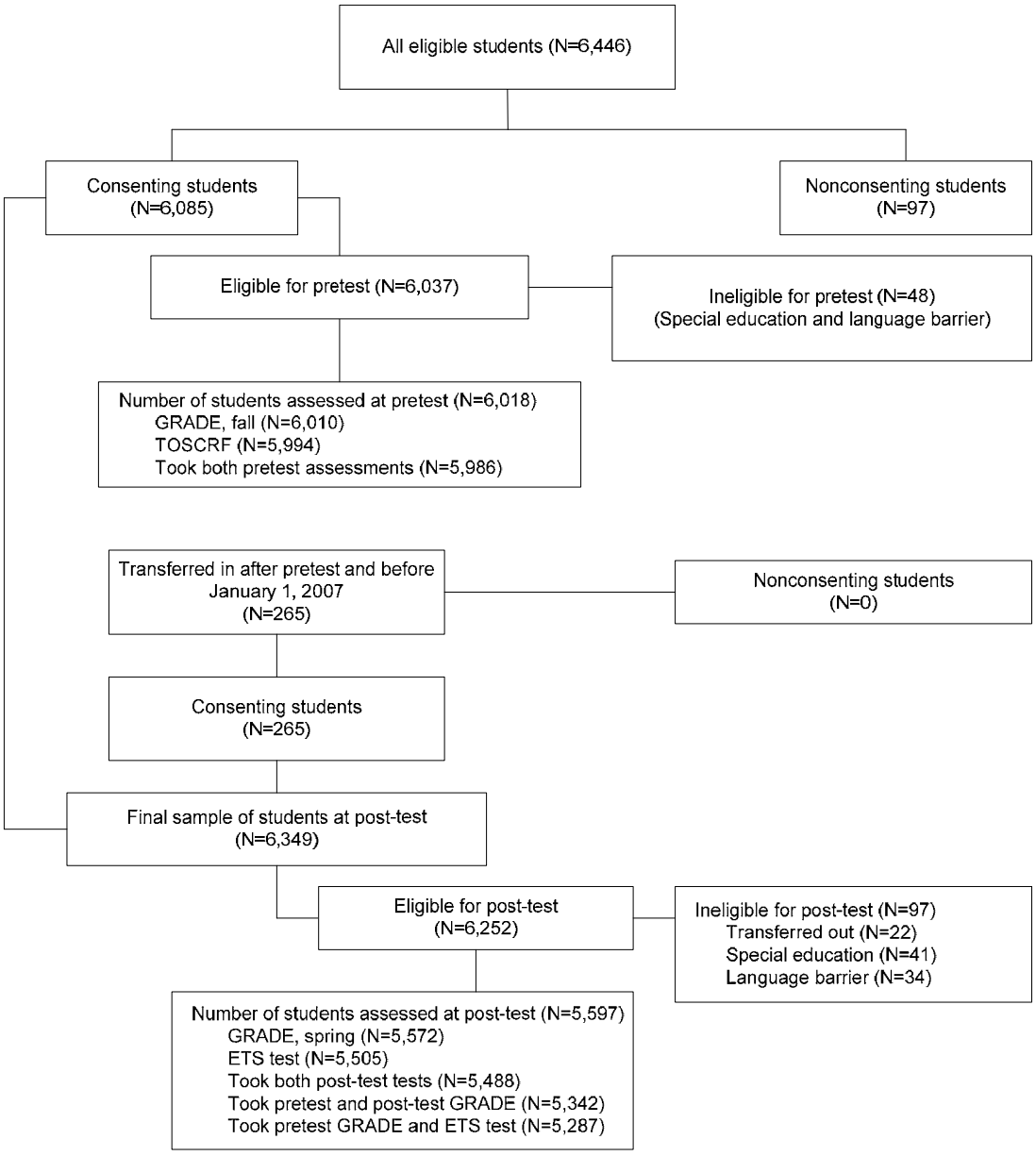
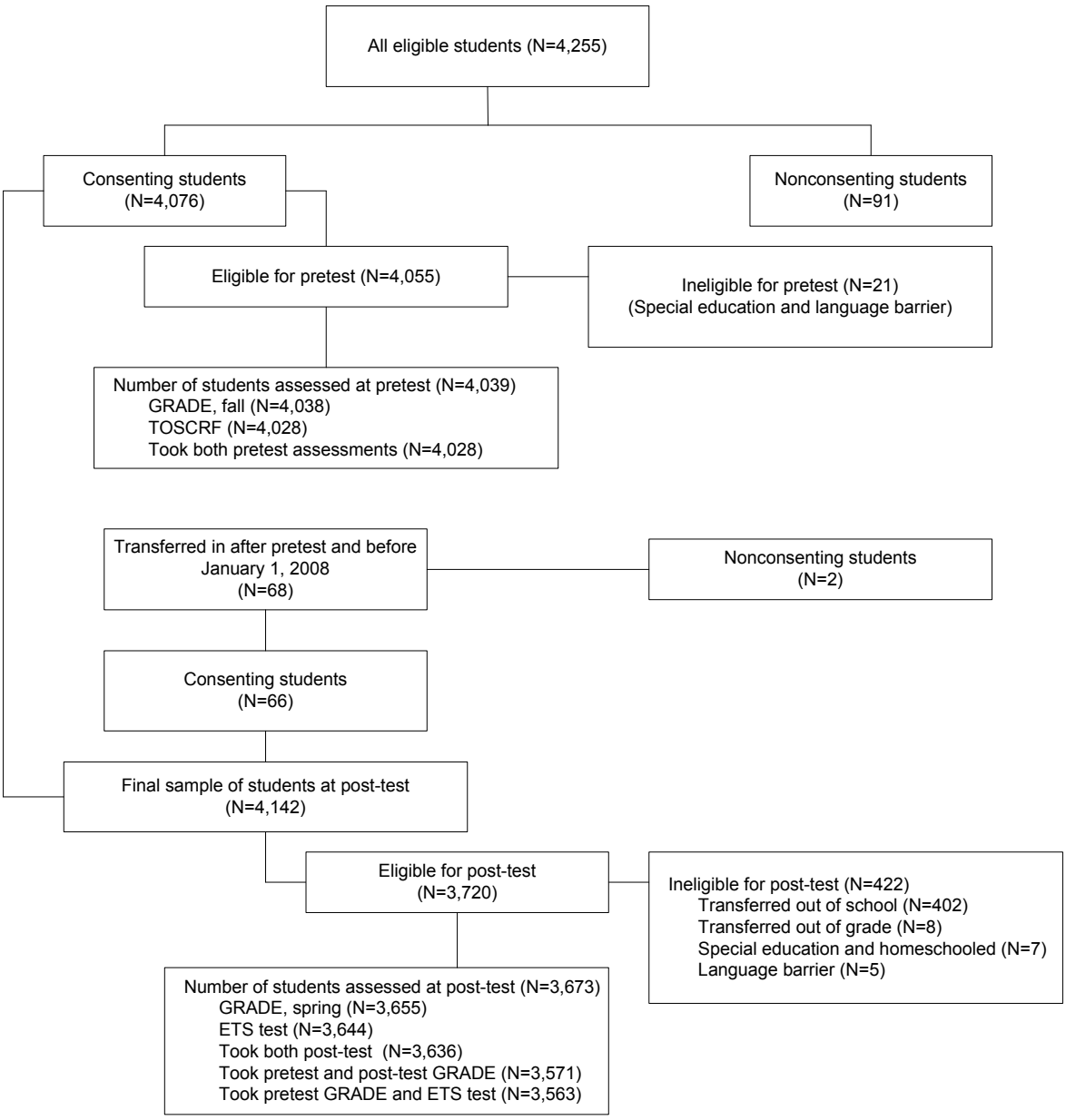


TABLE B.2
FLOW OF COHORT 1 STUDENTS THROUGH THE STUDY
Year 1



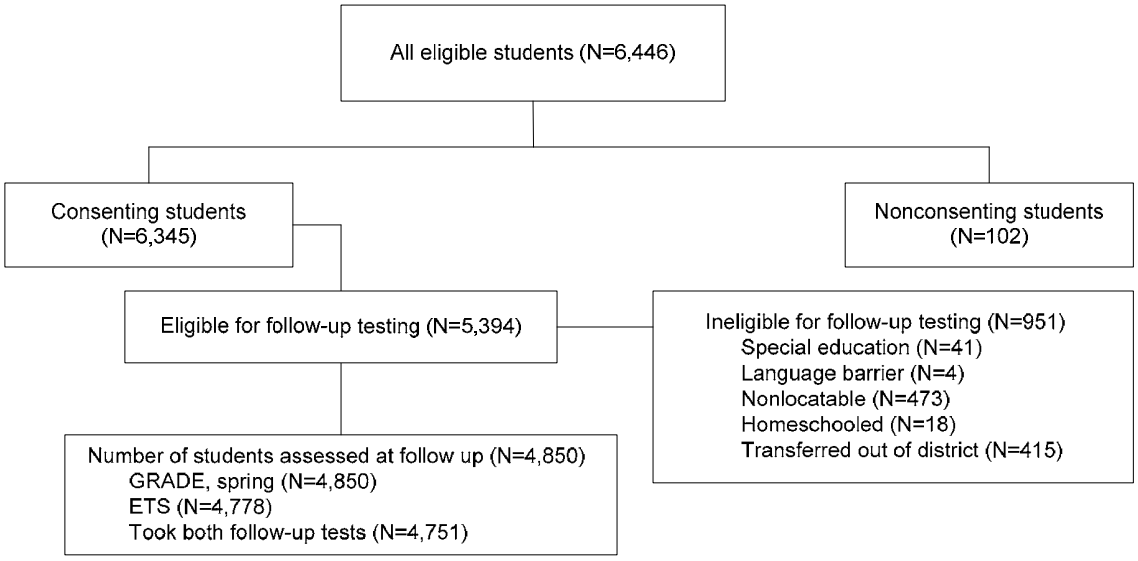
ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE B.2a
FLOW OF COHORT 2 STUDENTS THROUGH STUDY
Year 2



ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE B.2b
FLOW OF COHORT 1 STUDENTS THROUGH STUDY
Year 2



ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation.

APPENDIX C

OBTAINING PARENT CONSENT

This page is intentionally left blank.

A. YEAR 1

At the beginning of the 2006-2007 school year, the study team began the process of obtaining consent from parents of fifth-grade students attending study schools. We collected lists of all fifth-grade students in each study school (by classroom) and then sent letters to these students' parents requesting consent for their children to participate in the study. At the start of the spring semester, we again collected lists of fifth-grade students and sent consent letters to parents of students who had entered study classrooms after the baseline tests were administered but before January 1, 2007.

The letters sent home with students (which were translated into Spanish and Louisiana Creole for schools that requested it) explained the purpose of the study and all data collection activities involving students. The letters specified that students would be tested three times: at the beginning of the 2006-2007 school year, at the end of that year, and at the end of the 2007-2008 school year. A brochure with answers to frequently asked questions was also included in the mailing.

In most districts and with most students, passive consent procedures were implemented. Of the 6,446 students on teachers' fall or spring semester classroom lists, 937 attended schools in one district requiring active consent and 5,509 attended schools in the nine remaining districts requiring passive consent (Table C.1).

Parent consent was obtained for nearly all students (98 percent). We obtained consent for 93 percent of the students in the active consent district, and for 99 percent of the students in the passive consent districts.

There was no difference in consent rates by treatment or control status. Consent was obtained for 98 to 99 percent of students in each treatment and control condition (Table C.2).

B. YEAR 2

For students participating in the sixth-grade component of the second year of the study, no additional consent letters were distributed since the letters sent out in the 2006-2007 school year obtained consent for the Year 2 data collection. The parents of five students in the Cohort 1 sample withdrew their consent to participate in the second year of the study (three in the Reading for Knowledge intervention group and two in the control group), but the consent rate remained at 98 percent for the first cohort of students at the time of the follow-up survey in Year 2 (Table C.1).

For students in schools participating in the fifth-grade component of the second year of the study, we implemented the same consent procedures as were used in Year 1 of the study. The one key difference was that the consent letters indicated that students would be tested two times: at the beginning and end of the 2007-2008 school year. We used the same procedures for passive and active districts, as described above. Of the 4,255 students on teachers' fall or spring semester classroom lists, 660 attended schools in one district requiring active consent and 3,595 attended schools in the nine remaining districts requiring passive consent (Table C.1).

TABLE C.1

CONSENT RATES, BY TYPE OF CONSENT

All Eligible Students			Eligible Students in Passive Consent Districts (N = 9)			Eligible Students in Active Consent District (N = 1)		
	With Consent			With Consent			With Consent	
Total	Number	Percentage	Total	Number	Percentage	Total	Number	Percentage
Year 1								
Cohort 1 as of Post-Test								
6,446	6,350	98	5,509	5,478	99	937	872	93
Year 2								
Cohort 1 as of Follow Up								
6,446	6,345	98	5,509	5,474	99	937	871	93
Cohort 2 as of Post-Test								
4,255	4,142	97	3,595	3,584	100	660	558	85

TABLE C.2

CONSENT RATES, BY INTERVENTION

Intervention	All Eligible Students		
		With Consent	
	All	Number	Percentage
Year 1			
Cohort 1 as of Post-Test			
Total	6,446	6,350	98
Combined Treatment Group	5,055	4,983	99
Project CRISS	1,324	1,319	99
ReadAbout	1,256	1,246	99
Reading for Knowledge	1,220	1,191	98
Read for Real	1,255	1,227	98
Control Group	1,391	1,367	98
Year 2			
Cohort 1 as of Follow Up			
Total	6,446	6,345	98
Combined Treatment Group	5,055	4,978	99
Project CRISS	1,324	1,319	99
ReadAbout	1,256	1,246	99
Reading for Knowledge	1,220	1,188	97
Read for Real	1,255	1,227	98
Control Group	1,391	1,365	98
Cohort 2 as of Post-Test			
Total	4,255	4,142	97
Combined Treatment Group	3,033	2,948	98
Project CRISS	1,222	1,201	98
ReadAbout	1,123	1,108	99
Read for Real	688	639	93
Control Group	1,222	1,194	98

Parent consent was again obtained for nearly all students (97 percent). We obtained consent for 85 percent of the students in the active consent district, and for 100 percent of the students in the passive consent districts.

There was no difference in consent rates by treatment or control status. Consent was obtained for 93 to 98 percent of students in each treatment and control condition (Table C.2).

APPENDIX D

IMPLEMENTATION TIMELINE

This page is intentionally left blank.

TABLE D.1

IMPLEMENTATION SCHEDULE FOR INTERVENTIONS: NUMBER OF SCHOOL DAYS FROM START OF SCHOOL, BY DISTRICT

District Number	1	2	3	4	5	6[a]		7	8	9	10
School Calendar Type: Traditional (T) or Year-Round (Y)	T	T	T	T	T	T	Y	Y	T	T	T
Year 1, Cohort 1											
Days to Initial Scheduled Training											
Read for Real	–12	–9	–15	10	–7	10	–4	n.a.	–15	–10	–8
Project CRISS	–11	10	–13	23	22	2	n.a.	57	–15	–19	20
ReadAbout	–9	–12	–17	4	–8	11	40	–3; 6	–8	–9	–10
Reading for Knowledge	–11	–8	–15	33	–9	5	n.a.	–8	–7	–8	–11
Days Until Technology Was:[b]											
Ordered	19	–12	16	3	0	0	30	–10	4	–5	–3
Received	23	11	19	17	13	5	35	4	15	7	11
Ready for Use—First Set	38	16	32	33	21	18	48	8	24	9	14
Ready for Use—Second Set	n.a.	26	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	31	n.a.	n.a.
Year 2, Cohort 2											
Days to Initial Scheduled Training											
Read for Real[c]	20	n.a.	14	*	12	*	*	n.a.	23	n.a.	10
Project CRISS	–14	11	–13	–8	–20	–5	16	–14	–14	14	–19
ReadAbout	–14	16	–19	–5	–7	5	25	5	26	n.a.	–13
Days Until Technology Was:[d]											
Ordered	24	n.a.	15	–5	n.a.	n.a.	n.a.	n.a.	25	n.a.	31
Received	37	n.a.	36	48	n.a.	n.a.	n.a.	n.a.	78	n.a.	55
Students begin using program	55	14	45	6	0	49	69	8	38	n.a.	37

NOTE: A negative number in this table indicates that the training took place *before* the start of the school year. For example, the –12 days shown for District 1 for Read for Real indicates that the Read for Real training in District 1 took place 12 days before the start of the school year. Similarly, a positive number indicates that the training took place *after* the start of the school year.

[a]One participating district included schools following both year-round and traditional calendars.

[b]Technology installation applies only to the ReadAbout program. Technology refers to the computers, software, and other equipment needed to implement the program. The developer reported to Mathematica when the technology was ready for use.

[c]In Year 2, Read for Real provided training to new teachers only. An asterisk (*) in the table indicates that there were no new teachers to train in a particular district.

[d]In Year 2, about half of the schools did not require new equipment, and those that did were adding to the equipment they received in Year 1. Therefore, days between the start of school and the ordering and receiving of equipment are not applicable in some cases; also, students may have begun using the program before the installation of the new equipment.

n.a. = not applicable.

This page is intentionally left blank.

APPENDIX E

SAMPLE SIZES AND RESPONSE RATES

This page is intentionally left blank.

In Year 1 all fifth-grade teachers in study schools were considered eligible for the study, but individual teachers could decline to participate. Teachers who taught combined fourth-/fifth- or fifth-/sixth-grade classes were ineligible, as were teachers who taught self-contained special education classes. Table E.1a shows the final teacher sample, by treatment group, and the percentage of teachers who responded to the teacher survey in Year 1. The response rates shown reflect that, in Year 2, seven fifth-grade teachers who had not responded to the teacher survey in Year 1 completed the survey.

In Year 2, the study team administered surveys to sixth-grade teachers of students from the first cohort. Table E.1b shows the percentage of sixth-grade teachers who responded to the teacher survey in Year 2.[80]

Students enrolled in fifth-grade classes in study schools as of January 1, 2007 were eligible for the study's first cohort of students. Students in combined fourth-/fifth- or fifth-/sixth-grade classes were excluded, as were those in self-contained special education classes. Eligible students were considered in the study sample if parent consent was obtained (Table E.2). The same eligibility guidelines (and a sample cut-off date of January 1, 2008) were used in the second year of the study for the second cohort of fifth-grade students (see Table E.2).

Baseline tests were administered during regular class periods to in-sample students at the start of the students' first school year in the study (fall 2006 for Cohort 1 and fall 2007 for Cohort 2). The only in-sample students who were not eligible for testing were those whose limited English language skills precluded them from taking a test written in English. Most students who were absent on the initial test day were tested at subsequent make-up test sessions. Ninety-five percent of Cohort 1 students completed the baseline GRADE test, and 94 percent completed the baseline TOSCRF test; over 99 percent of students who took the baseline GRADE also took the baseline TOSCRF, and vice versa (Table E.3a). Ninety-seven percent of Cohort 2 students completed the baseline GRADE test, and 97 percent completed the baseline TOSCRF test; over 99 percent of students who took the baseline GRADE also took the baseline TOSCRF, and vice versa.

Post-tests were administered to in-sample students who had not transferred out of the school district at the time of testing. As was done at baseline, students whose limited English language skills at follow up precluded them from taking a test written in English were not included in post-testing. The post-tests were administered at the end of the students' first school year in the study, on two consecutive days, with make-up sessions scheduled for absent students (spring 2007 for Cohort 1 and spring 2008 for Cohort 2). Of the total sample of Cohort 1 students (including those who could not be tested because they were not geographically accessible), 88 percent completed the GRADE post-test and 87 percent completed the ETS post-test (Table E.3b). Eighty-eight percent of Cohort 2 students completed the GRADE post-test and 88 percent completed the ETS post-test. In addition, more than 98 percent of Cohort 1 students who took the GRADE post-test also took the ETS post-test, and more than 99 percent of those who took the ETS post-test also took the GRADE post-test. These numbers were similar for Cohort 2 students.

[80]Response rates by treatment/control group are not shown as this would result in teachers being counted more than once (as students from multiple treatment groups and the control group can have the same sixth-grade teacher).

TABLE E.1a

TEACHER SURVEY SAMPLE AND RESPONSE RATES, GRADE FIVE TEACHERS

	Teachers		
	Total	Number Completing Survey	Response Rate (Percentage)
Teachers of Cohort 1 Students in Year 1[a]			
Total	268	256	96
Combined Treatment Group	209	199	96
Project CRISS	52	52	100
ReadAbout	50	48	96
Reading for Knowledge	53	50	94
Read for Real	54	49	91
Control Group	59	57	97
Teachers of Cohort 2 Students in Year 2[b]			
Total	184	n.a.	n.a.
Combined Treatment Group	130	n.a.	n.a.
Project CRISS	53	n.a.	n.a.
ReadAbout	46	n.a.	n.a.
Read for Real	31	n.a.	n.a.
Control Group	54	n.a.	n.a.

[a]Response rates shown for Cohort 1 reflect additional efforts by the study team to administer teacher surveys in the second study year to the seven teachers who had not responded in the first year. All seven teachers contacted in the second year returned a completed survey.

[b]Teacher Surveys were not administered to fifth-grade teachers in Year 2, with one exception. The seven teachers who had not responded to the Teacher Survey in the first year of the study were asked to complete a Teacher Survey in the second year.

TABLE E.1b

TEACHER SURVEY SAMPLE AND RESPONSE RATES, GRADE SIX TEACHERS

Teachers of Cohort 1 Students in Year 2[a]			
	Teachers		
	Total	Number Completing Survey	Response Rate (Percentage)
Total	907	486	54
District 1	137	96	70
District 2	25	14	56
District 3	43	33	77
District 4	193	54	28
District 5	128	74	58
District 6	113	83	73
District 7	18	10	56
District 8	80	75	94
District 9	59	23	39
District 10	111	24	22

[a]Students were asked to provide the last name of their Language Arts/Reading, Social Studies, and Science teachers. The teachers were then given a survey to complete. Teacher information was obtained for 4,509 students of 6,350 in the Cohort 1 sample. Response rates by treatment/control group are not shown, as this would result in teachers being counted more than once (as students from multiple treatment/control groups can have the same sixth-grade teacher).

TABLE E.2

STUDENT SAMPLE

Year 1, Cohort 1			
	Pretest Sample	Transferred in before January 1, 2007	Total Sample[a]
Total	6,085	265	6,350
Combined Treatment Group	4,761	222	4,983
Project CRISS	1,241	78	1,319
ReadAbout	1,205	41	1,246
Reading for Knowledge	1,157	34	1,191
Read for Real	1,158	69	1,227
Control Group	1,324	43	1,367
Year 2, Cohort 2			
	Pretest Sample	Transferred in before January 1, 2008	Total Sample[a]
Total	4,076	68	4,142
Combined Treatment Group	2,900	48	2,948
Project CRISS	1,172	30	1,201
ReadAbout	1,101	10	1,108
Read for Real	627	8	639
Control Group	1,176	20	1,194

[a]The total number of students in the study sample includes students from the first and second columns. In particular, the sample includes students in study schools as of January 1 and for whom parental consent was obtained. Students who transferred out of their school district after January but before follow-up testing remained part of the sample (about 450 in Cohort 1, and about 400 in Cohort 2).

TABLE E.3a

STUDENT TEST SAMPLE AND RESPONSE RATES, PRETEST

	Total	Number Tested	Response Rate[a] (Percentage)	Percentage Who Took the Listed Test Who Also Took the Other Pretest[b]
Year 1				
Cohort 1, Fall 2006				
GRADE				
Total	6,349	6,010	95	99.6
Combined Treatment Group	4,987	4,708	94	99.6
Project CRISS	1,319	1,233	93	99.4
ReadAbout	1,245	1,186	95	99.7
Reading for Knowledge	1,195	1,138	96	99.7
Read for Real	1,228	1,151	94	99.6
Control Group	1,362	1,302	95	99.7
TOSCRF				
Total	6,349	5,994	94	99.9
Combined Treatment Group	4,987	4,696	94	99.8
Project CRISS	1,319	1,226	93	99.9
ReadAbout	1,245	1,186	95	99.7
Reading for Knowledge	1,195	1,137	95	99.8
Read for Real	1,228	1,147	93	99.9
Control Group	1,362	1,298	95	100.0
Year 2				
Cohort 2, Fall 2007				
GRADE				
Total	4,142	4,036	97	99.8
Combined Treatment Group	2,948	2,871	97	99.7
Project CRISS	1,201	1,158	96	99.7
ReadAbout	1,108	1,090	98	99.9
Read for Real	639	623	98	99.5
Control Group	1,194	1,165	97	99.8
TOSCRF				
Total	4,142	4,026	97	99.8
Combined Treatment Group	2,948	2,863	97	99.7
Project CRISS	1,201	1,154	96	99.7
ReadAbout	1,108	1,089	98	99.9
Read for Real	639	620	97	99.5
Control Group	1,194	1,163	97	99.8

Table E.3a *(continued)*

[a]The percentage of Cohort 1 students tested at pretest is based on the total sample, although about 265 students included in the sample transferred into participating schools after the pretest was completed. Of the students in the sample at the time of the pretest, 99 percent completed the GRADE and the TOSCRF.

[b]The GRADE and the TOSCRF were administered on the same day, so nearly all students who completed one pretest also completed the other pretest. However, a small number of students completed only one test: of those who completed the pretest GRADE, 99.6 percent also completed the TOSCRF; of those who completed the TOSCRF, 99.9 percent also completed the pretest GRADE.

[c]The percentage of Cohort 2 students tested at pretest is based on the total sample, although about 68 students included in the sample transferred into participating schools after the pretest was completed. Of the students in the sample at the pretest testing, 99 percent completed the GRADE and the TOSCRF.

[d]The GRADE and the TOSCRF were administered on the same day, so nearly all students who completed one pretest test also completed the other pretest. However, a small number of students completed only one test: of those who completed the pretest GRADE, 99.8 percent also completed the TOSCRF; of those who completed the TOSCRF, 99.8 percent also completed the pretest GRADE.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE E.3b

STUDENT TEST SAMPLE AND RESPONSE RATES, POST-TEST

Year 1

Cohort 1, Spring 2007

	Total	Number Tested	Response Rate (Percentage)[a]	Percentage Who Took the Listed Test Who Also Took the Other Post-Test[b]	Percentage Who Took the Listed Test Who Also Took the Pretest GRADE[c]
GRADE					
Total	6,349	5,573	88	98.5	84
Combined Treatment Group	4,987	4,394	88	98.4	84
Project CRISS	1,319	1,154	87	98.4	83
ReadAbout	1,245	1,095	88	99.1	85
Reading for Knowledge	1,195	1,067	90	98.0	87
Read for Real	1,228	1,078	88	98.0	84
Control Group	1,362	1,179	86	98.8	83
ETS					
Total	6,349	5,512	87	99.6	83
Combined Treatment Group	4,987	4,344	87	99.5	83
Project CRISS	1,319	1,139	86	99.7	82
ReadAbout	1,245	1,089	87	99.6	84
Reading for Knowledge	1,195	1,051	88	99.5	85
Read for Real	1,228	1,065	87	99.2	82
Control Group	1,362	1,168	85	99.7	83

Year 2

Cohort 2, Spring 2008

	Total	Number Tested	Response Rate (Percentage)[d]	Percentage Who Took the Listed Test Who Also Took the Other Post-Test[e]	Percentage Who Took the Listed Test Who Also Took the Pretest GRADE[f]
GRADE					
Total	4,142	3,665	88	99.2	98
Combined Treatment Group	2,948	2,604	88	99.0	98
Project CRISS	1,201	1,056	88	98.6	97
ReadAbout	1,108	994	89	99.3	98
Read for Real	639	554	87	99.5	98
Control Group	1,194	1,061	89	99.6	98
ETS					
Total	4,142	3,644	88	99.8	98
Combined Treatment Group	2,948	2,587	88	99.7	98
Project CRISS	1,201	1,045	87	99.6	97
ReadAbout	1,108	989	89	99.8	98
Read for Real	639	553	87	99.6	98
Control Group	1,194	1,057	88	100.0	99

Table E.3b (continued)

[a]The percentage of Cohort 1 students tested at follow up is based on the total sample, although about 450 of those students had transferred out of their school district before the follow-up tests. Of the students who had not transferred out of their district, about 94 percent completed the post-tests.

[b]The follow-up GRADE and ETS tests were administered on consecutive days to Cohort 1 students. Nearly all students who completed one test also completed the other test. However, a small number of students completed only one test: of those who completed the post-test GRADE, 98.5 percent also completed the ETS test; of those who completed the ETS test, 99.6 percent also completed the post-test GRADE.

[c]Some Cohort 1 students transferred into study schools after the pretest was completed, and some in-sample students transferred out of study schools before the follow-up test was administered. Eighty-four percent of the students completed both the pretest and post-test GRADE, and 83 percent completed both the pretest GRADE and the ETS test.

[d]The percentage of Cohort 2 students tested at follow up is based on the total sample, although about 400 of those students had transferred out of their school district before the follow-up tests. Of the students eligible for testing, about 99 percent completed the post-test.

[e]The follow-up GRADE and ETS tests were administered on consecutive days to Cohort 2 students. Nearly all students who completed one test also completed the other test. However, a small number of students completed only one test: of those who completed the post-test GRADE, 99.2 percent also completed the ETS test; of those who completed the ETS test, 99.8 percent also completed the post-test GRADE.

[f]Some Cohort 2 students transferred into study schools after the pretest was completed, and some in-sample students transferred out of study schools before the follow-up test was administered. Ninety-eight percent of the students completed both the pretest and post-test GRADE, and 98 percent completed both the pretest GRADE and the ETS test.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation.

At follow up for Cohort 1 students (spring 2008), 76 percent of students completed the GRADE follow-up test and 75 percent of students completed the ETS follow-up test (Table E.3c). Ninety-eight percent of students who completed the GRADE follow-up test also completed the ETS follow-up test, and 99 percent of students who completed the ETS follow-up test also completed the GRADE follow-up test.

All students who completed follow-up tests were included in the impact analysis. The proportion of students in each experimental condition with follow-up test scores is reported in Table G.2.

Table E.4 shows the classroom observation sample and response rates, and Table E.5 shows the treatment classrooms in the fidelity observation sample and response rates. Table E.6 shows response rates on the teacher surveys that collected data on students' use of informational text and teachers' allocation of time in the school day.

TABLE E.3c

STUDENT TEST SAMPLE AND RESPONSE RATES, FOLLOW UP

Year 2				
Cohort 1, Follow Up, Spring 2008				
	Total	Number Tested	Response Rate (Percentage)[a]	Percentage Who Took the Listed Test Who Also Took the Other Follow-Up Test
GRADE				
Total	6,349	4,850	76	98
Combined Treatment Group	4,987	3,828	77	98
Project CRISS	1,319	1,060	80	98
ReadAbout	1,245	967	78	98
Reading for Knowledge	1,195	899	75	98
Read for Real	1,228	902	74	99
Control Group	1,362	1,022	75	98
ETS				
Total	6,349	4,778	75	99
Combined Treatment Group	4,987	3,774	76	99
Project CRISS	1,319	1,037	79	99
ReadAbout	1,245	955	77	99
Reading for Knowledge	1,195	885	74	99
Read for Real	1,228	897	73	99
Control Group	1,362	1,044	76	96

[a]The percentage of students tested at the second follow up is based on the total Year 1 sample, although about 950 of those students were ineligible for follow-up testing, primarily because they had either transferred out of the district (415) or their enrollment status was unknown by the school district (473). Of the students who were located and eligible for testing, about 90 percent completed the follow-up tests.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation.

TABLE E.4

CLASSROOM OBSERVATION SAMPLE AND RESPONSE RATES

	Classrooms		
	Total	Number Observed	Response Rate (Percentage)
Year 1, Cohort 1			
Total[a]	270	264	98
Combined Treatment Group	213	207	97
Project CRISS	56	52	93
ReadAbout	50	49	98
Reading for Knowledge	53	52	98
Read for Real	54	54	100
Control Group	57	57	100
Year 2, Cohort 2			
Total[a]	190	175	92
Combined Treatment Group	135	126	93
Project CRISS	56	51	91
ReadAbout	46	46	100
Read for Real	33	29	88
Control Group	55	49	89

[a]The number of classrooms shown in this table differs from the number of teachers shown in Table E.1a because some teachers taught more than one class.

TABLE E.5

FIDELITY OBSERVATION SAMPLE AND RESPONSE RATES

	Teachers		
	Total[a]	Number Observed	Response Rate (Percentage)
Year 1, Cohort 1			
Combined Treatment Group[b]	218	209	96
Project CRISS	54	54	100
ReadAbout	53	53	100
Reading for Knowledge	54	45	83
Read for Real	57	57	100
Year 2, Cohort 2			
Combined Treatment Group	130	114	88
Project CRISS	53	46	87
ReadAbout	46	43	93
Read for Real	31	25	81

[a]The number of teachers shown in this table differs from the number shown in Table E.4 because this table focuses on number of *teachers,* while Table E.4 focuses on number of *classrooms.*

[b]One fidelity observation was conducted for each study teacher. The number of teachers shown in this table differs from the number shown in Table E.1a because the Teacher Survey was conducted at the start of the 2006–2007 school year, while the fidelity observations were conducted later in the year (after some teacher changes had occurred).

TABLE E.6

RESPONSE RATES FOR YEAR 2 TEACHER FORMS, GRADE FIVE TEACHERS

	Teachers		
	Total	Number Completing Form	Response Rate (Percentage)
Students' Use of Informational Text			
Total	184	156	85
Combined Treatment Group	130	109	84
Project CRISS	53	46	87
ReadAbout	46	41	89
Read for Real	31	22	71
Control Group	54	47	87
Teachers' Allocation of Time to Students' Daily Schedules[a]			
Total	209	187	89
Combined Treatment Group	151	135	89
Project CRISS	56	50	89
ReadAbout	52	46	88
Read for Real	43	39	91
Control Group	58	52	90

[a]The number of teachers shown in this pane of the table differs from the number shown in the top pane because this form was administered to all grade five teachers who were using, or had used, CRISS, Read for Real, or ReadAbout during Year 2 of the study, or who used one of these treatments during Year 1 and were still teaching at a school where we were testing students in spring 2008. The numbers in the top pane represent only those teachers who were teaching fifth grade in Year 2 of the study.

This page is intentionally left blank.

APPENDIX F

CREATION AND RELIABILITY OF CLASSROOM OBSERVATION AND TEACHER SURVEY MEASURES

This page is intentionally left blank.

A. ASSESSING INTER-RATER RELIABILITY

An important part of the analysis of data collected from classroom observations is an assessment of the reliability of the observation data across the staff conducting the observations. In the study's second year, data from 30 percent of classrooms are available for these calculations. Twenty percent of observations were randomly chosen to be reliability observations, which means that a second observer was randomly chosen to observe simultaneously with the observer assigned to that observation. The remaining 10 percent of the observations come from pairings of a master trainer with each observer at least once during the first two weeks of observation. This allows for a comparison of the data collected by the two observers during these observations.

In total, the study team had data from 70 pairs of observations that could be used to assess reliability of the observation data. Of these, 50 were pairs of regular field observation staff. An additional 20 were pairs in which a regular field observer did one observation and an expert observer acting in a quality control role did the second.

The inter-rater reliability of all of the scales in the study's second year was over 0.94 (0.94 to 0.98). Pearson correlations of the scale scores based on the two observers' tallies were calculated for the three study scales. The inter-rater reliability of the Traditional Interaction scale, the Reading Strategy Guidance scale, and the Classroom Management scale was 0.98, 0.97, and 0.94, respectively.

Inter-rater reliability for individual items from the classroom observation form was also analyzed. We calculated reliability by item by measuring the exact match percent agreement between observers in both types of pairs (reliability and quality control, during each interval). This method involves calculating agreements and disagreements tally by tally, to determine the exact match. That is, if observer one had six tallies and observer two had four tallies in the same cell, we counted four agreements and two disagreements. This measure of agreement thus takes into account the degree of variation between observers' tallies.

The calculation of inter-rater reliability was conducted in a way designed to avoid inflating reliability scores simply because the target behaviors were unobserved. Because there were many zeros, representing the "absence" of the indicated instructional behaviors, there was a possibility that reliability could be exaggerated by inclusion of zeros in reliability calculations, because reliability would be 100 percent if neither observer recorded a tally. To address this issue, we removed any intervals that had no tallies from the reliability calculations.

The inter-rater reliability (as measured by percent agreement between observers) for individual items from the classroom observation form in the study's second year ranged from 85 to 100 percent. The total percent agreement across all items was 92 percent (see Table F.1).[81]

[81] Appendix I shows key descriptive statistics (including means and standard deviations) for the full set of items from the classroom observation and fidelity instruments.

TABLE F.1

PERCENT AGREEMENT RELIABILITY FOR ACTIVE INTERVALS, BY ITEM

Item		Agreements of Observed Items		Agreements of Unobserved Items		Disagreements		Percent Agreement[a]	
		Year 1	Year 2	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Comprehension Items									
Modeling and Thinking Aloud									
1A	Background knowledge	3	0	408	277	1	0	99.76	100.00
2A	Text structure	1	10	411	271	0	1	100.00	99.67
3A	Various comprehension strategies	0	4	408	271	4	4	99.03	98.66
4A	Generating questions	1	2	410	274	1	0	99.76	100.00
5A	Text features	1	0	410	276	1	0	99.76	100.00
Total		**6**	**16**	**2,047**	**1,369**	**7**	**5**	**99.66**	**99.66**
Explaining/Reviewing									
1B	Background knowledge	160	162	354	233	47	21	91.62	95.17
2B	Text structure	111	75	355	242	44	21	91.37	94.12
3B	Various comprehension strategies	443	286	321	174	126	63	85.84	88.38
4B	Generating questions	96	73	326	223	45	21	90.36	93.75
5B	Text features	78	85	344	228	34	39	92.54	89.49
Total		**888**	**681**	**1,700**	**1,100**	**296**	**165**	**89.74**	**91.92**
Comprehension Student Practice									
1C	Background knowledge	301	294	348	227	38	21	94.47	96.26
2C	Text structure	169	126	356	235	49	25	91.46	93.83
3C	Various comprehension strategies	614	459	246	155	134	91	86.52	87.43
4C	Generating questions	161	138	287	206	78	34	85.17	91.44
5C	Text features	90	137	349	222	39	30	91.84	92.65
Total		**1,335**	**1,154**	**1,586**	**1,045**	**338**	**201**	**89.63**	**91.94**
Vocabulary Items									
Interactive Teaching									
6	Justifying responses	76	69	336	236	60	25	87.29	92.84
7	Higher-order questioning	388	278	228	176	171	83	78.27	85.07
8	Elaborating/clarifying the text	533	394	188	140	190	78	79.14	87.64

Table F.1 *(continued)*

Item		Agreements of Observed Items		Agreements of Unobserved Items		Disagreements		Percent Agreement[a]	
		Year 1	Year 2	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Total		**997**	**741**	**752**	**552**	**421**	**186**	**80.60**	**87.89**
Teaching Vocabulary									
V1	Providing definitions	288	173	227	173	122	46	80.85	88.81
V2	Providing examples/elaborations	488	285	213	160	131	63	84.25	88.05
V3	Providing visuals	136	83	324	234	64	26	87.79	92.82
V4	Teaching context clues	38	25	376	255	18	10	95.83	96.76
Total		**950**	**566**	**1,140**	**822**	**335**	**145**	**86.19**	**90.99**
Vocabulary Student Practice									
V5	Using knowledge of words	757	504	190	129	190	82	83.29	88.83
V6	Using context clues	30	33	390	251	16	11	96.33	96.50
Total		**787**	**537**	**580**	**380**	**206**	**93**	**86.90**	**91.13**
Items in Each Area									
Comprehension		3,226	2,592	6,085	4,066	1,062	557	89.76	92.63
Vocabulary		1,737	1,103	1,720	1,202	541	238	86.47	91.04
Total		**4,963**	**3,695**	**7,805**	**5,268**	**1,603**	**795**	**88.85**	**92.22**
Items Contained in the Classroom Observation Scales									
Traditional Interaction		3,159	2,277	3,778	2,633	1,158	548	85.69	90.39
Reading Strategy Guidance		1,876	1,477	3,704	2,447	631	363	89.84	90.93

SOURCE: Classroom observations.

NOTE: Inter-rater reliability calculations were based only on active intervals, which are those intervals during which the teacher and students were working on informational text and at least one teaching practice on the ERC form was observed by either member of the observer pair. If a teacher taught a lesson on informational text but was not observed to be using any of the teaching practices on the observation measure, that interval was not included.

[a]Reliability by item was calculated by measuring the exact match percent agreement between reliability (and quality control) observation pairs during each interval. This method involves calculating agreements and disagreements tally by tally, to determine the exact match. That is, if Observer 1 had six tallies and Observer 2 had four tallies in the same cell, this was counted as four agreements and two disagreements.

ERC = Expository Reading Comprehension.

B. ASSESSING CRITERION VALIDITY

Another important part of the analysis of classroom observation data is an examination of the criterion validity of the study's classroom observation scales. Criterion validity was measured by the extent to which these scales, measuring the incidence of teacher behaviors, are correlated with students' scores on reading comprehension tests. Achieving a high degree of validity for a scale suggests that affecting that scale has the potential to improve student achievement.

To examine this issue, we measured the extent to which the classroom observation scales are related to the study's key student test score outcomes in the study's second year. We conducted this analysis using classroom observation scales based on averages of activities across the observation intervals. We accounted for clustering of students within schools in calculating *p*-values, but we did not account for multiple comparisons because this is a purely exploratory analysis.

We found that one of the scales is positively and statistically significantly related to three of the four student test scores included in the study. The Classroom Management scale is statistically significantly related to the composite test scores (correlation = 0.19, *p*-value = 0.01); the GRADE scores (correlation = 0.17, *p*-value = 0.03); and the science reading comprehension assessment (correlation = 0.20, *p*-value = 0.01). We found no statistically significant relationship between any of the study's test scores and the Traditional Interaction or Reading Strategy scales.[82]

C. CREATION AND RELIABILITY OF CLASSROOM OBSERVATION MEASURES

The ERC observation form allowed the study team to collect consistent data from fifth-grade classrooms that make it possible to describe and compare teachers' instructional practices in different treatment and control groups and across cohorts. In the ERC observation form, observers recorded the number of times treatment and control group teachers engaged in specific teaching behaviors. There were up to 294 opportunities to record observed teaching practices (28 practices assessed in each of up to 10 intervals, plus a set of 14 items assessed once during an observation). Therefore, the classroom observation data needed to be condensed into a manageable number of variables for analysis so that we can present a coherent, summary picture of teachers' behavior. This appendix describes the process the study team used to obtain this more manageable number of variables.

We developed summary scales for groupings of specific items for Parts I and II of the ERC instrument. Part I of the instrument focused on interactive teaching practices, vocabulary instruction, and comprehension strategy instruction; Part II focused on classroom management and student engagement. The development of scales was done by implementing preliminary exploratory factor analysis, conducting a review of item content, and implementing item

[82]Results presented in Chapter V differ from those presented here because Chapter V analyses combined the cohort 1 and 2 samples whereas the results presented here are only for cohort 2.

response theory (IRT) scaling (Nunnally and Bernstein 1994; Lord 1980; Wright and Stone 1979; and Lord and Novick 1968).

The goal of the factor analysis was to identify preliminary groupings of items for Part I of the ERC instrument that appeared to represent key underlying dimensions. Any of the Part I items that were weakly related to the identified underlying dimensions were dropped from further psychometric analyses. This process ultimately resulted in three groupings of items for Part I.

A review of item content was used to identify groupings of items for Part II of the ERC instrument, due to the smaller number of items and more distinct content groupings of items in this section of the instrument. Two groupings of items for Part II were specified based on the thematic similarities of content shared between the items for each of the two groups. In total, across Parts I and II of the ERC, five groupings of items were identified.[83]

The goal of the IRT scaling was to estimate reliable and valid scores for teachers on scales that represent the underlying dimensions for the respective item groupings in Parts I and II of the ERC instrument. The data preparation, IRT scaling process, evaluation of IRT model fit, evaluation of reliability and validity of scores, and information on how to interpret the scores are described in detail below.

Data Preparation. To support the most-valid IRT item calibration and score estimation, we conducted additional data processing of the items in each of the five groupings. The tallies for items of Part I for each interval were averaged across the 10-minute intervals for each classroom within a single day. We then evaluated the frequency distributions of each item and created meaningful categories representing the extent to which behaviors were observed (such as low, medium, and high).[84] The category boundaries were determined based on investigation of the frequency distributions for each item.

Because the items of Part II of the ERC instrument have their own specified rating scales, there was no need to create categories for those items. Therefore, data for the items of Part II were analyzed according to these existing rating scales.

IRT Scaling Process. For each of these five groups of items, IRT scaling was used to develop variables measuring the underlying latent dimensions. The IRT model features a multivariate logistic regression of the probability for the demonstration (or level of response) on each item in a grouping (such as low, medium, or high) on the latent dimension as an underlying continuous variable, which was estimated by way of an iterative numerical process. The joint

[83]During the IRT scaling process, another dimension was specified in order to account for two items within Part II of the ERC that shared a common question stem. The additional dimension was specified to avoid estimation bias (it was *not* specified for use in the study's examination of the relationship between impacts and teacher practices).

[84]To permit sensitivity testing of the scales used in the analysis, we also created these categories based on sums of observed tallies across the 10-minute intervals for the day's observations. IRT scaling was done for data based on sums of tallies for items across the intervals, as well as averages of tallies for items across the intervals.

probabilities for the levels of demonstrations across the full set of items within a grouping, conditional on the underlying continuous variable used to represent the latent dimension, are used to estimate scores as proficiency estimates on the scale for the respective latent dimension. These scores quantify the levels of estimated proficiency for demonstrating the underlying skill for each latent dimension.

Scores for the five scales (that is, one scale for each of the five groupings of items) were estimated for all classrooms using a specific IRT technique. IRT item calibration and score estimation was done using the Multidimensional Random Coefficients Multinomial Logit Model (Adams et al. 1997).[85] This model was used to specify a multidimensional generalization of the Partial Credit Model (Masters and Wright 1997; Masters 1982), and is the core model of the software ACER ConQuest (Wu et al. 2007).

This modeling approach permitted us to properly address the ways in which the ERC items were interrelated because, as Adams et al. (1997) explain, the Multidimensional Random Coefficients Multinomial Logit Model can address two kinds of multidimensionality of assessment data: between-item multidimensionality and within-item multidimensionality. Between-item multidimensionality occurs when particular items load only on a single scale, but there are multiple scales due to the presence of multiple underlying dimensions. Within-item multidimensionality occurs when particular items load on more than one scale due to cross-loadings. The ERC data on this study exhibit both between-item and within-item multidimensionality.

Items in the scales had two to four categories for the levels of demonstration, which affected how they were treated during IRT scaling. Items with only two categories (low and high) were treated as dichotomous items for IRT item calibration, while items with more than two categories (low, medium, and high, for example) were treated as polychotomous items. Data for dichotomous and polychotomous items for scales were analyzed together during the IRT analysis; this was possible because the IRT software used permits analysis for scales that have mixtures of item types, even when the numbers of categories for items differ.

The model used to calculate scale scores assumes independence of the data points that contribute to the estimation of its parameters. Since 124 fifth-grade teachers were observed in both the first and second years of the study, we were concerned about item response correlation (i.e., for a given teacher participating in Years 1 and 2 of the study, the level of response for a particular item [or items] in Year 1 might be similar to the level of response for the same item [or items] in Year 2). To address that concern and to make the scale scores comparable across classrooms observed in Years 1 and 2, the study team fixed the IRT model parameters (item difficulty, variances for the latent distributions, and covariances between latent dimensions) to

[85]Using the Multidimensional Random Coefficients Multinomial Logit Model permitted (1) explicit modeling of the multidimensionality of the item data during analysis, facilitating proper estimation for the statistical characteristics of items, even as they contribute to multiple domains; (2) proper model specification when different items share common stems, necessitating additional dimensions to control for residual correlations between such items in order to avoid estimation bias; and (3) Bayesian estimators for both item and score parameter estimates, and an IRT-based reliability estimate for each scale overall and for the score of each classroom.

the levels estimated using the data from the observations conducted in the first year only. Based on this model, scale scores were then calculated for classrooms in the study's second year. Since the scores in Year 2 are estimated on the same calibrated scales as the scores in Year 1, we can make comparisons between the sets of scores from the two years.

Evaluation of IRT Model Fit. Overall, the IRT model fit the data well. Based on the guideline of 0.5 to 1.7 for reasonable infit and outfit mean square values for items of a clinical observation instrument (Wright and Linacre 1994), the scaling process resulted in acceptable overall model fit for each item contained in three of the scales we constructed (the Traditional Interaction scale, the Reading Strategy Guidance scale, and the Classroom Management scale) (see Table F.2).[86] Two additional scales that were created in this process were not used in the study's analyses in either year of the study due to concerns over their reliability or inter-rater reliability based on the IRT model parameter estimations conducted in the first study year. For one of these scales, reliability was the concern (with values of .43 for the version of the scale based on averages of teacher practice tallies and .58 for the version of the scale based on sums of tallies). For the other scale, inter-rater reliability was the concern (with values of .69 for the version of the scale based on averages of tallies and .73 for the version based on sums of tallies).

Additional statistical tests provide support for the use of the three reliable scales in the analysis. Based on data from classrooms observed in the first year of the study, the separation reliability estimate for item parameter estimation is 0.99, indicating a high level of reliability for the estimation of item parameters, given that a value of 1.0 represents, in theory, the maximum possible value for this parameter. In addition, the Chi-square test of item parameter equality based on the Year 1 observation data is statistically significant ($\chi^2 = 5233.70$, df = 34, $p < .05$). For the scale scores computed for the classrooms observed in the study's second year, separation reliability estimation and the Chi-square test of item parameter equality cannot be calculated because the parameters of the IRT model were fixed to the levels estimated using the data from the observations conducted in the first year only. Taken together, and since the same parameter estimates were used in constructing scale scores in Years 1 and 2, these statistics indicate that the IRT model provides a good fit for the observation data from both years of the study. The statistics also indicate that items function well enough to ensure acceptable levels of measurement precision at various points along the scales.

Reliability and Validity of Scores. The reliability of the scales based on classrooms observed in the study's second year is 0.65 for Traditional Interaction, 0.63 for Reading Strategy Guidance, and 0.86 for Classroom Management (Table F.3).

There is also evidence supporting the validity of the scales. First, the content of the items in Part I was based on experimental research from small-scale studies that investigated sound practices for reading comprehension and vocabulary instruction, and the content of items in Part II was based on a theoretical framework that identified some of the most-essential practices for classroom instruction in general, and the quality of classroom management in particular. Second, the content of the items in each scale is generally homogenous. Third, the empirical findings

[86]Fit at the level of each category for all items for the three scales was also examined. In general, results from this examination showed acceptable IRT model fit for the categories of all the items.

TABLE F.2

ITEM RESPONSE MODEL DIFFICULTY PARAMETERS, STANDARD ERRORS, OUTFIT AND INFIT STATISTICS, AND CORRECTED ITEM-TOTAL CORRELATIONS FOR ITEMS OF EACH SCALE

	Item Difficulty[a]		Standard Error[b]		Outfit Mean Square[c]		Infit Mean Square[d]		Corrected Item-Total Correlation[e]	
Item	Year 1	Year 2[f]	Year 1	Year 2[f]	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Traditional Interaction										
• Comprehension Item 4	505.34	505.34	0.49	n.a.	1.01	1.01	1.03	1.05	0.31	0.37
Comprehension Item 4C	502.05	502.05	0.42	n.a.	1.14	1.07	1.13	1.06	0.24	0.39
Comprehension Item 5B	506.64	506.64	0.53	n.a.	0.90	0.90	0.97	0.96	0.30	0.38
Comprehension Item 5C	506.16	506.16	0.52	n.a.	0.98	1.02	1.03	1.06	0.22	0.34
Comprehension Item 6C	503.79	503.79	0.43	n.a.	1.29	1.48	1.18	1.23	0.14	0.08
Comprehension Item 7C	503.70	503.70	0.48	n.a.	1.06	1.24	1.09	1.21	0.41	0.33
Comprehension Item 8	506.79	506.79	0.89	n.a.	1.06	1.09	1.05	1.09	0.37	0.28
Vocabulary Item 1	503.53	503.53	0.85	n.a.	1.26	1.18	1.17	1.13	0.25	0.17
Vocabulary Item 2	512.29	512.29	1.03	n.a.	0.86	0.93	0.87	0.91	0.38	0.43
Vocabulary Item 3	511.31	511.31	1.03	n.a.	0.89	1.03	0.87	0.98	0.43	0.28
Vocabulary Item 4	511.56	511.56	0.67	n.a.	1.02	1.06	1.05	1.08	0.25	0.21
Vocabulary Item 5	507.40	507.40	0.93	n.a.	0.86	0.98	0.89	0.99	0.31	0.26
Vocabulary Item 6	519.29	519.29	1.29	n.a.	1.24	1.34	1.15	1.23	0.17	0.21
Reading Strategy Guidance										
Comprehension Item 2B	516.85	516.85	1.18	n.a.	0.92	1.15	1.01	1.17	0.32	0.21
Comprehension Item 2C	514.19	514.19	1.09	n.a.	1.10	1.33	1.14	1.33	0.24	0.13
• Comprehension Item 3A	529.36	529.36	2.51	n.a.	1.22	1.10	1.07	1.03	0.14	0.14
Comprehension Item 3B	510.62	510.62	0.99	n.a.	0.82	0.86	0.91	0.91	0.44	0.16
Comprehension Item 3C	505.89	505.89	0.91	n.a.	0.97	1.03	1.00	1.04	0.37	0.06
Comprehension Item 4B	505.34	505.34	0.49	n.a.	1.01	1.01	1.03	1.05	0.35	0.36
Comprehension Item 4C	502.05	502.05	0.42	n.a.	1.14	1.07	1.13	1.06	0.26	0.34
Comprehension Item 5B	506.64	506.64	0.53	n.a.	0.90	0.90	0.97	0.96	0.43	0.28
Comprehension Item 5C	506.16	506.16	0.52	n.a.	0.98	1.02	1.03	1.06	0.38	0.24
Comprehension Item 6C	503.79	503.79	0.43	n.a.	1.29	1.48	1.18	1.23	0.23	0.09
Vocabulary Item 4	511.56	511.56	0.67	n.a.	1.02	1.06	1.05	1.08	0.14	0.11
Classroom Management										
Part 2, Item 10	471.92	471.92	1.36	n.a.	0.97	1.01	1.00	1.01	0.76	0.76
Part 2, Item 11	465.29	465.29	1.44	n.a.	0.90	0.95	1.11	1.15	0.76	0.80
Part 2, Item 13	473.41	473.41	1.05	n.a.	0.93	0.92	1.16	1.04	0.74	0.78
Part 2, Item 14	477.85	477.85	1.00	n.a.	0.71	1.25	0.98	1.03	0.78	0.80

SOURCE: Classroom observations.

[a]Item difficulty provides a sense of the extent to which different behaviors will be observed in classrooms. Classroom scores and item difficulty parameter estimates are expressed together on the same scale, so that teachers (classrooms) that are more likely to exhibit behaviors for particular items will score above the respective difficulty levels for those items, and teachers (classrooms) that are less likely to exhibit behaviors for the items will score below the difficulty levels for the items.

[b]The standard error is the estimation error of the item difficulty parameter.

[c]Outfit mean square is the average of the standardized residual variance for the item without any weighting (thus, it is sensitive to outliers). The expected value is 1.0, with values less than .5 and greater than 1.7 considered to indicate problematic items for a clinical observation measure (Wright and Linacre 1994).

Table F.2 *(continued)*

[d]Infit mean square is the average of the standardized residual variance after weighting for each individual residual variance, so that unexpected responses close to the item's difficulty are given greater weight. The expected value is 1.0, with values less than .5 and greater than 1.7 considered to indicate problematic items for a clinical observation measure (Wright and Linacre 1994).

[e]Corrected item-total correlation is the correlation between responses on an item and the total raw score that is calculated using the remaining set of items for the scale in order to correct for spuriousness.

[f]The item difficulty parameter estimate in Year 2 was constrained to its corresponding value from Year 1; therefore, it has no associated standard error.

n.a. = not applicable.

TABLE F.3

DESCRIPTIVE STATISTICS OF TEACHER INSTRUCTIONAL PRACTICES SCALE SCORES

Scale	Number of Classrooms	Reliability	Mean	Standard Deviation	Minimum	Maximum
Year 1						
Traditional Interaction	261	.70	500.00	6.53	486.37	517.38
Reading Strategy Guidance	261	.72	500.09	7.42	483.37	518.18
Classroom Management	261	.83	500.46	31.05	404.87	562.40
Year 2						
Traditional Interaction	173	.65	499.16	5.82	484.03	512.78
Reading Strategy Guidance	173	.63	501.10	6.12	479.59	513.48
Classroom Management	173	.86	499.23	32.30	408.40	550.55

SOURCE: Classroom observations.

demonstrate an acceptable level of IRT model fit for the items in each scale. Finally, the multidimensional IRT model specification posits that there are multiple latent dimensions that explain the statistical relationships between all possible pairs of items for the respective scales, and the extent to which the model fits the data (as indicated by the item fit statistics) provides supporting evidence of the presence of these latent dimensions/components.

Interpreting the Scale Scores. Figures F.1a through F.3 provide a way to interpret the levels of the scale scores for the classrooms observed in the study's second year. In particular, they provide a way to link a particular scale score to the ordinal categories that summarize the frequency with which teachers engaged in the practices underlying the three scales. For example, for the Traditional Interaction scale, Figure F.1a shows how 6 of the 13 items contained in the scale link to the levels of the scale scores. (Figure F.1b shows how the remaining 7 items in the scale link to the scale scores.) For example, a scale score of 560 corresponds to teachers explaining how to generate questions .56 to 2 times on average during each 10-minute interval (first bar) while that same score corresponds to teachers asking questions that go beyond a literal level 1.4 to 6.22 times during a 10-minute interval (last bar). It is important to note that teachers' actual scale score values do not vary as widely as the 400 to 600 range implied by the figures (as shown in the maximum and minimum values in Table F.3), because the actual scale scores reflect *multiple* teacher practices while each bar in Figures F.1a through F.3 represents just *one* teacher practice and the scale score that is possible based on that one practice. For example, in theory, a teacher could have scored as high as 600 (or as low as 400) on the Traditional Interaction scale, but none did so due to the levels of observed behaviors on all of the practices comprising that scale.

D. CREATION OF TEACHER EFFICACY AND SCHOOL PROFESSIONAL CULTURE SCALES

We used data from the Teacher Survey administered to fifth-grade teachers in the study's first year to construct a Teacher Efficacy scale and a School Professional Culture scale.

Teacher Efficacy Scale

Twelve items from the Teacher Survey were used to construct this scale (items borrowed with permission from Hoy and Woolfolk, 1993). These items are on a 0 to 5 Likert scale and correspond to teacher self-reports on attitudes and beliefs on student engagement (4 items), instructional strategies (4 items), and classroom management (4 items). To create the teacher efficacy scale, we averaged the responses to the 12 items for each teacher, so the original scale of 0 to 5 was preserved. A higher score on the scale represents more-positive teacher perceptions of their efficacy.

The reliability of the Teacher Efficacy scales exceeded 0.79 (0.79 to 0.90). The alpha for the overall Teacher Efficacy scale was 0.90, and the reliability of the Teacher Efficacy subscales was 0.83, 0.79, and 0.85, for efficacy in student engagement, efficacy in instructional strategies, and efficacy in classroom management, respectively (Table F.4).

FIGURE F.1a

LINK BETWEEN AVERAGE NUMBER OF TIMES PRACTICES WERE OBSERVED AND TRADITIONAL INTERACTION SCALE SCORES, YEAR 2

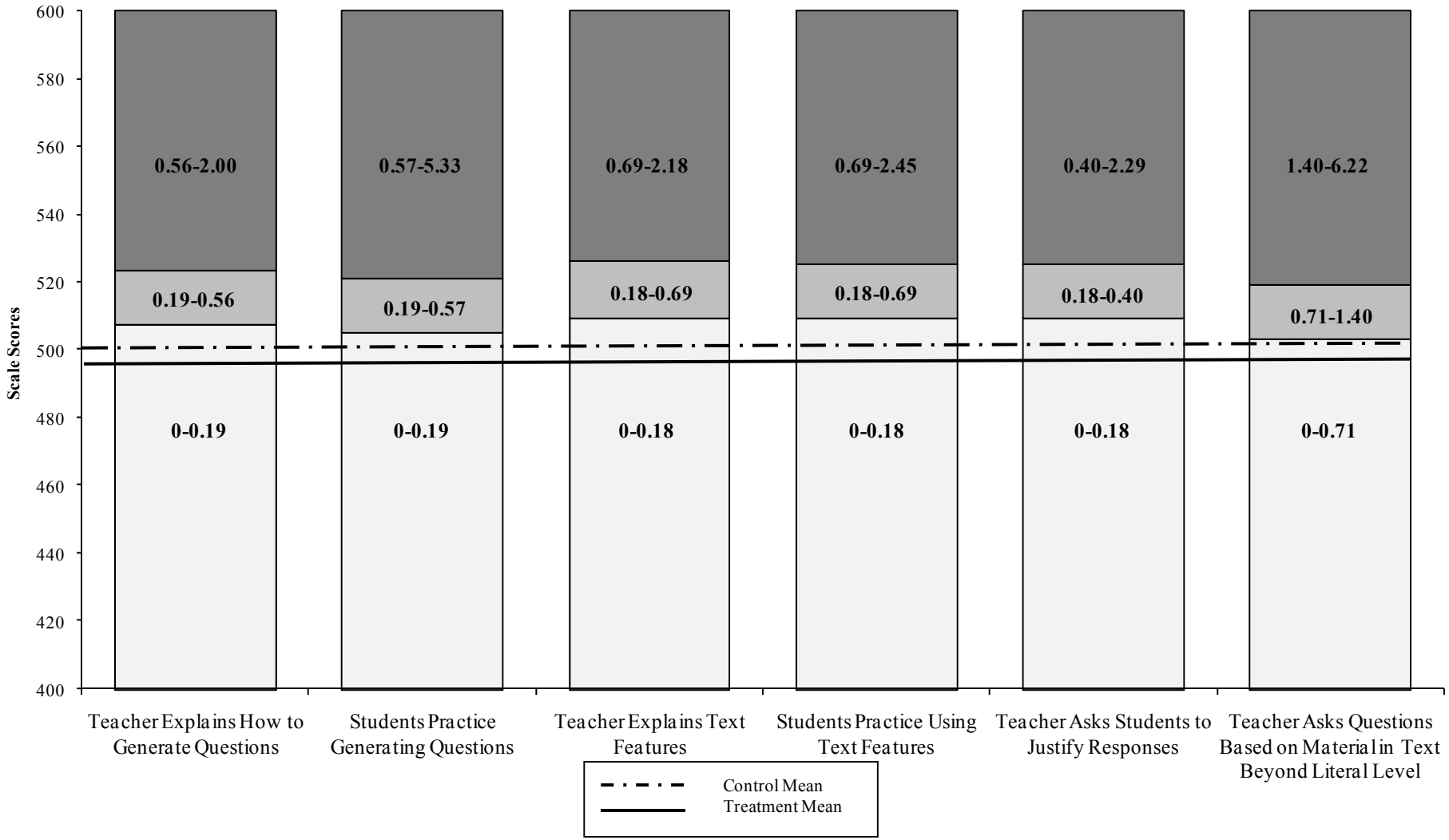


FIGURE F.1b

LINK BETWEEN AVERAGE NUMBER OF TIMES PRACTICES WERE OBSERVED AND TRADITIONAL INTERACTION SCALE SCORES, YEAR 2

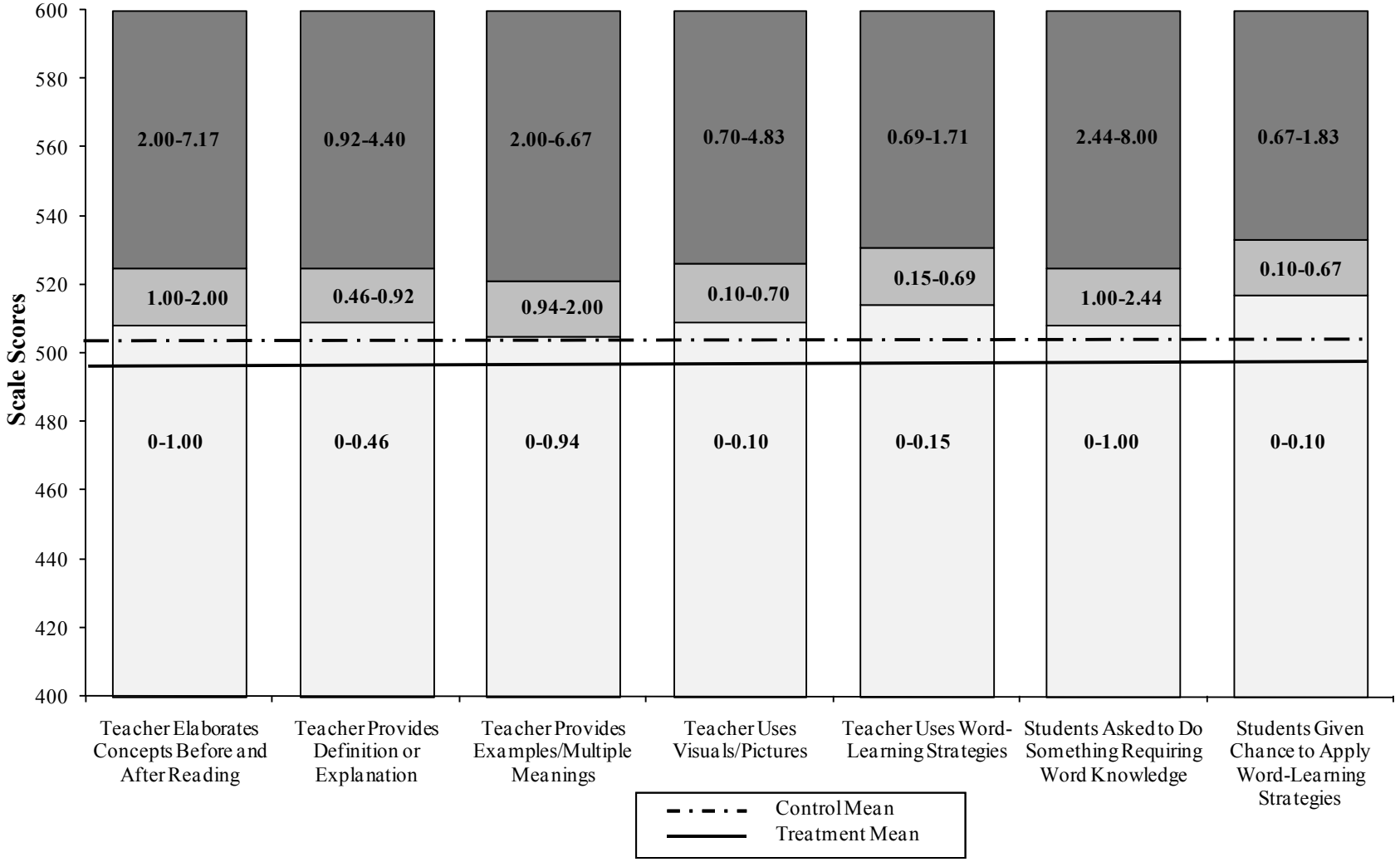


FIGURE F.2

LINK BETWEEN AVERAGE NUMBER OF TIMES PRACTICES WERE OBSERVED AND READING STRATEGY GUIDANCE SCALE SCORES, YEAR 2

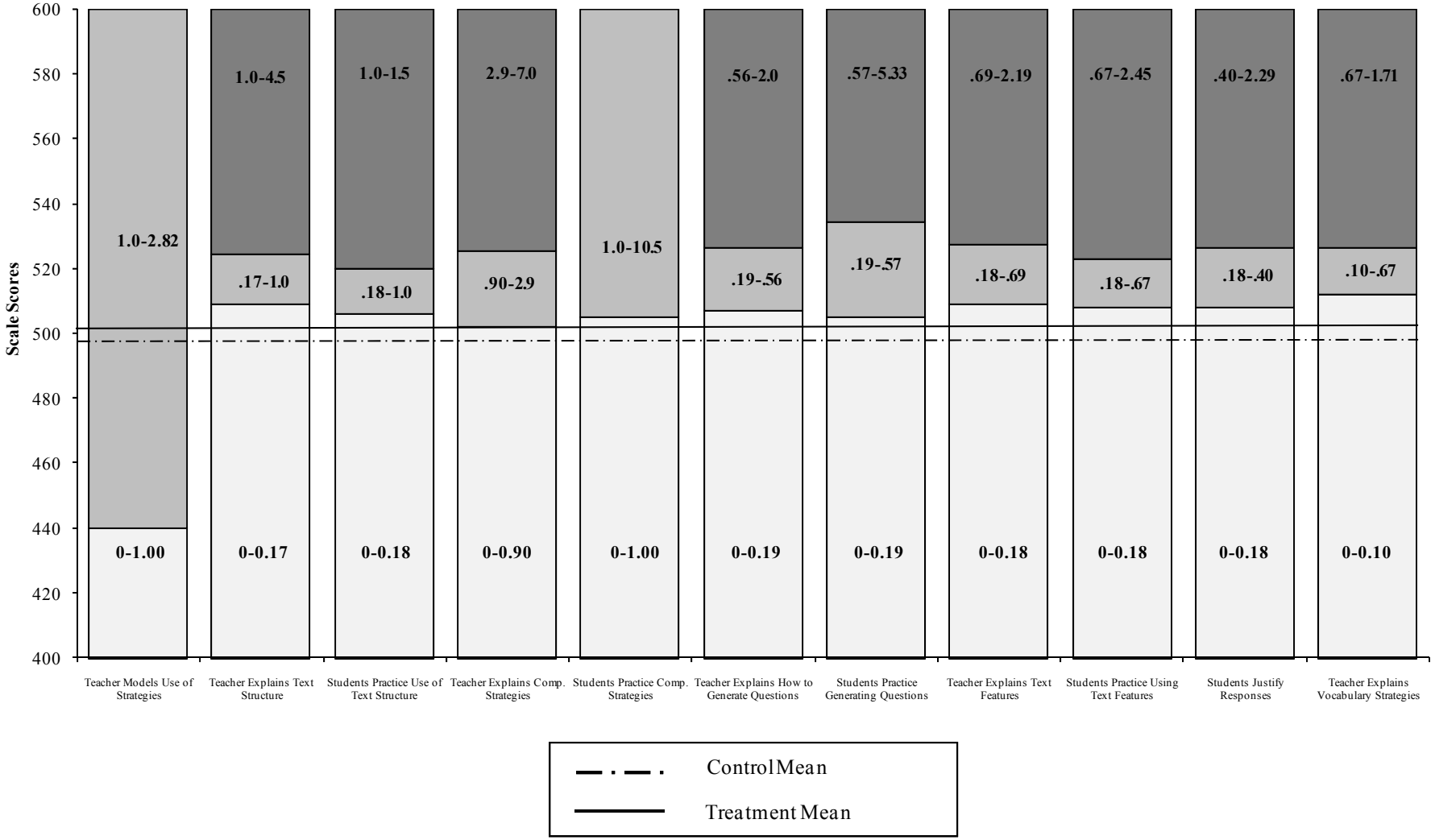


FIGURE F.3

LINK BETWEEN AVERAGE LIKERT-SCALE ITEM RATINGS AND SCALE SCORES FOR CLASSROOM MANAGEMENT, YEAR 2

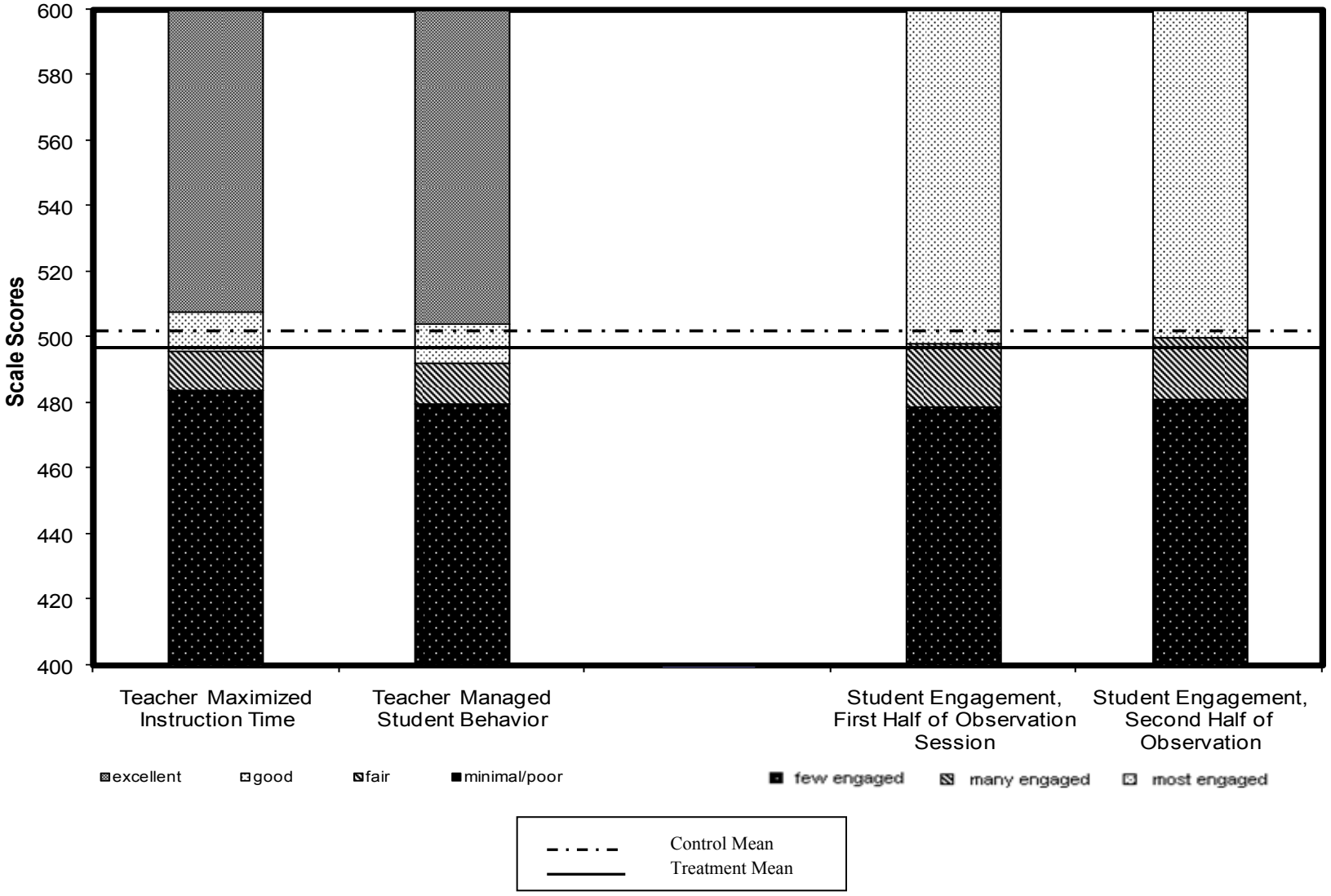


TABLE F.4

RELIABILITY OF THE TEACHER EFFICACY OVERALL SCALE AND SUBSCALES

Scale	Number of Items	Coefficient Alpha	Mean	Standard Deviation	Minimum	Maximum
Overall Teacher Efficacy	12	0.90	4.19	0.49	2.83	5.0
Efficacy in Student Engagement	4	0.83	4.07	0.62	2.25	5.0
Efficacy in Instructional Strategies	4	0.79	4.14	0.54	2.50	5.0
Efficacy in Classroom Management	4	0.85	4.34	0.56	2.25	5.0

SOURCE: Teacher Survey administered to fifth-grade teachers in Year 1 of the study.

School Professional Culture Scale

Thirty-five items from the Teacher Survey were used to construct this scale. The items correspond to teacher self-reports on attitudes and beliefs on reflective dialogue, perceptions about relationships among peers, access to new ideas, experience with changes being implemented in school, professional development opportunities, and leadership and support. The range of this scale is 0 to 10, and a higher score on the scale indicates more-positive teacher perceptions of the professional culture in their school.

This scale was constructed using a Rasch rating-scale model in Winsteps (Linacre 2006). In the Rasch rating-scale model, scale scores were constructed by estimating the probability of a specified response as a function of (1) each teacher's ability level for the construct being measured and (2) item difficulty. In IRT analyses, ability corresponds to the level of the attitude or belief being measured, and item difficulty corresponds to the prevalence of or likelihood of endorsing the attitude, belief, or behavior represented by each item in a scale. Most-prevalent beliefs, attitudes, or behaviors are least difficult to endorse, while less-prevalent ones are more difficult to endorse.

In the rating scale model, the scores are usually rescaled to correspond to the original scale on the items in order to ease interpretation. For the School Professional Culture scale, the scores were rescaled to a 0 to 10 scale. The rescaled scores were used in the statistical analyses presented in this report. Item difficulties were also rescaled with the least difficult items having low values on the scale. The item difficulties and teacher scores are thus placed on a common scale and the items are expected to be ordered hierarchically along the difficulty continuum. Therefore, the way to interpret these scales is that teachers are more likely to endorse items with difficulty below their scale score and less likely to endorse items with difficulty above their scale score. Given that scores estimated on a limited number of responses are less reliable than scores with more ratings, if 50 percent or more of the items in a scale were missing, the score for that teacher was set to missing.[87]

Several statistical tests indicate that this scale and its six subscales (corresponding to the six categories of attitudes and beliefs described above) are reliable and valid measures. Person separation reliability, infit mean square, and item difficulty were produced to evaluate the reliability and validity of the scales. Person separation reliability, which is equivalent to Cronbach's alpha and measures internal consistency of the scale, ranged from 0.66 to 0.87 for the overall scale and subscales (Table F.5). The infit mean square values for most of the items, which indicate whether the response items are consistent with the hierarchical ordering of the items, were close to 1, which suggests that most response patterns align with the hierarchical ordering of the items in the six subscales (Table F.6). Finally, the items in the six subscales were spread along the difficulty hierarchy, with item difficulty statistics ranging from 2.97 to 6.27.

[87]This occurred for only two teachers in the sample.

TABLE F.5

DESCRIPTIVE STATISTICS AND PERSON SEPARATION RELIABILITIES FOR THE OVERALL SCHOOL CULTURE SCALE AND SUBSCALES

Scale	Number of Items	Person Separation Reliability	Sample Size	Mean	Standard Deviation	Minimum	Maximum
Overall School Culture	35	.87	258	5.69	.47	4.53	7.86
Reflective Dialogue	4	.78	253	5.62	2.00	0	10
Perceptions About Relationships Among Peers	6	.82	258	8.17	1.95	2.26	9.99
Access to New Ideas	6	.75	258	5.04	1.30	2.21	10
Experience of Change	3	.66	256	5.97	1.85	1.21	9.99
Professional Development Opportunities	9	.86	257	5.74	1.46	2.55	10
Leadership and Support	7	.84	255	7.39	2.06	0	9.99

SOURCE: Teacher Survey administered to fifth-grade teachers in Year 1 of the study.

TABLE F.6

PSYCHOMETRIC STATISTICS FOR SCHOOL CULTURE SUBSCALES

Subscale/Item	Infit Mean Square[a]	Item Difficulty[b]
Reflective Dialogue		
During the past school year, how often have you had conversations with colleagues about …		
5a. The goals of this school?	.95	5.55
5b. Development of new curriculum?	1.06	6.02
5c. Managing classroom behavior?	1.25	4.11
5d. What helps students learn best?	.74	3.93
Perceptions About Relationships Among Peers		
How much do you disagree or agree with each of the following …		
6a. Teachers in this grade level trust each other.	.93	5.04
6b. It's OK in this grade level to discuss feelings, worries, and frustrations with other teachers.	.87	4.93
6c. Teachers respect other teachers who take the lead in grade-level improvement efforts.	.79	5.08
6d. Teachers in this grade level respect those colleagues who are expert at their craft.	.76	4.90
6e. To what extent do you feel respected by other teachers in this grade level?	1.42	4.30
6f. How many teachers in this grade level really care about each other?	1.06	4.76
Access to New Ideas		
How often have you …		
7a. Taken courses at a college or university relative to improving your school?	1.41	4.91
7b. Participated in a network with other teachers outside your school?	.86	4.53
7c. Discussed curriculum and instruction matters with an outside professional group or organization?	.85	4.74
7d. Attended professional development activities organized by your school (include meetings that focus on improving your teaching)?	1.10	2.97
7e. Attended workshops or courses sponsored by your school district (exclude required in-services)?	.85	3.71
7f. Attended professional development activities sponsored by the teachers' union?	.99	6.27
Experience of Change		
How much do you disagree or agree with each of the following …		
8a. Most changes introduced at this school involve only a few teachers; rarely does the whole faculty become involved (reverse-coded).	1.13	4.56
8b. We receive adequate professional development support for the changes we introduce at our school.	1.16	4.94
8c. Most changes introduced at this school gain little support among teachers (reverse-coded).	.68	4.64

Table F.6 (*continued)*

Subscale/Item	Infit Mean Square[a]	Item Difficulty[b]
Professional Development Opportunities		
Overall, my professional development experiences over the past school year …		
9a. Have included opportunities to work productively with teachers from other schools.	1.24	5.20
9b. Have included enough time to think carefully about, to try, and to evaluate new ideas.	.99	5.64
9c. Have deepened my understanding of subject matter.	.77	4.35
9d. Have helped me understand my students better.	.81	4.63
9e. Have been sustained and coherently focused, rather than being short term and unrelated.	.85	5.13
9f. Have included opportunities to work productively with colleagues in my school.	1.16	4.74
9g. Have led me to make changes in my teaching.	.71	3.99
9h. Have been closely connected to my school's improvement plan.	1.22	3.96
9i. Most of what I learn in professional development addresses the needs of the students in my classroom.	1.10	4.35
Leadership and Support		
How much do you disagree or agree with each of the following …		
10a. The principal at this school is strongly committed to shared decision making.	1.46	5.02
10b. The principal at this school works to create a sense of community in the school.	.80	4.46
10c. The principal at this school promotes parent and community involvement in the school.	.94	3.95
10d. The principal at this school supports and encourages teachers to take risks.	.91	5.12
10e. The principal at this school is willing to make changes.	.91	4.62
10f. Most changes introduced at this school receive strong support from the principal.	.80	4.99
10g. The principal at this school encourages teachers to try new methods of instruction.	1.11	4.48

SOURCE: Teacher Survey administered to fifth-grade teachers in Year 1 of the study.

[a]Infit mean square is the average of the standardized residual variance weighting for each individual residual variance so that unexpected responses close to the item's difficulty are given greater weight. The expected value is 1.0, with values less than .5 and greater than 1.7 generally considered poorly fitting items (Wright and Linacre 1994).

[b]Item difficulty is the relative likelihood that different opinions/perceptions of the professional culture in their schools will be endorsed by teachers. Items that are endorsed more frequently have lower values, and items that are endorsed less frequently have higher values. Teachers and items are placed on the same scale so that teachers who are highly likely to endorse the perceptions are below the item difficulty for their score, and teachers who are less likely to endorse the perceptions have difficulties above their score.

APPENDIX G

ESTIMATING IMPACTS

This page is intentionally left blank.

This appendix describes our approach to calculating impacts as part of our primary and secondary analyses. Our primary analyses focus on the central questions of whether any of the supplemental curricula individually, or as a group, improve a second cohort of students' scores on reading comprehension assessments after schools and teachers have a year of experience using the supplemental curriculum, and whether the interventions have an impact on the test scores of the first cohort of students one year after the end of their experience with the supplemental curricula. Our secondary analyses were designed to decompose overall impacts and thus improve our understanding of whether the supplemental curricula are particularly effective for certain subgroups, and to explore the pathways through which supplemental curricula affect student achievement.

A. BENCHMARK APPROACH TO CALCULATING PRIMARY IMPACTS

The benchmark approach to calculating impacts reflects decisions regarding methodological approaches determined most appropriate for this study. The approach also reflects input from the Department of Education (ED) and the study's Technical Work Group regarding suitable analytic approaches given the study's design and goals. Five key areas are addressed in our benchmark approach to estimating impacts: (1) regression adjustment, (2) clustering of students, (3) missing data, (4) multiple comparisons, and (5) weights.

1. Regression Adjustment

We calculated impacts using regression adjustment in order to increase the statistical precision of our impact estimates, which would enable us to detect smaller treatment effects. We also adjusted for any characteristics of schools, teachers, or students that differed significantly between treatment and control groups at baseline. Although random assignment ensures no systematic differences between the treatment and control groups in the characteristics of students, teachers, or schools, it is still possible that random differences will exist between the groups. By regression adjusting for these random differences, we can greatly improve the precision of our impact estimates. With regression adjustment, the minimum detectable effect size (MDES) of this study is 0.14 standard deviations for the impacts on post-test scores of the second cohort of students and 0.25 for the impacts on follow-up scores of the first cohort of students. Without regression adjustment, the MDES would have been 0.44 standard deviations for the impacts on post-test scores of the second cohort of students and 0.50 for the impacts on follow-up scores of the first cohort of students. The covariates in our impact models are baseline GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher sex, teacher age, teacher race, and district indicators.

We also included district fixed effects in our regression model (in the form of district indicator variables) to further increase statistical precision.[88] We treat district effects as fixed

[88]Alternatively, we could have included block indicator variables, which would have reduced the degrees of freedom for the impact regressions from 67 to 63. As a robustness check, we conducted statistical tests using 63

rather than random because (1) districts were not randomly sampled and (2) districts were not randomly assigned. Stated differently, if we were to repeat the study we would have the same districts represented in the study and in the treatment and control groups, meaning that districts do not vary and do not contribute to variation in impacts.

In equation form, the regression model we estimated when examining impacts for the first cohort of students in the second year is:

$$(1)\ y_{i,j} = \alpha + \delta_1 CRISS_j + \delta_2 RA_j + \delta_3 R4K_j + \delta_4 R4R_j + \beta X_{i,j} + \sum_{k=1}^{10} \gamma_k D_k + u_j + \varepsilon_{i,j}$$

where i and j index students and schools, respectively; *CRISS*, *RA*, *R4K*, and *R4R* are treatment group indicators (for Project CRISS, ReadAbout, Reading for Knowledge, and Read for Real, respectively); X represents covariates; D_1-D_{10} are district indicators; u is a school-level random intercept; and ε is a student-level random intercept. The impact of the interventions relative to the control group (the omitted category) is given by the coefficients on the treatment group indicators. For example, the impact of Project CRISS is given by δ_1. We estimated two versions of this model in the study's second year—one in which the outcome is the post-test (measured at the end of fifth grade for the first cohort) and one in which the outcome is the follow-up test (measured at the end of sixth grade for the first cohort). To test for differences in impacts between post-test and follow-up, we estimated a stacked regression model that allowed us to calculate cross-equation covariance terms. Below we describe how we account for the correlation between students within schools that is implied by the school-level random intercept.

The regression model we estimated to assess impacts of the interventions after *schools* have had a year of experience using them is:

$$(2)\ \begin{aligned} y_{i,j} = {} & \alpha + \delta_1 CRISS_j + \delta_2 RA_j + \delta_3 R4K_j + \delta_4 R4R_j + \lambda_1 C_{i,j} + \\ & \delta_5 C_{i,j} \cdot CRISS_j + \delta_6 C_{i,j} \cdot RA_j + \delta_7 C_{i,j} \cdot R4K_j + \delta_8 C_{i,j} \cdot R4R_j + \beta X_{i,j} + \sum_{k=1}^{10} \gamma_k D_k + u_j + \varepsilon_{i,j} \end{aligned}$$

where C is a cohort indicator variable and other variables are defined as above. Note that both first and second cohort students are included in this regression and the outcome variable is the students' post-test scores (from spring 2008 for Cohort 2 students and spring 2007 for Cohort 1 students). The coefficients δ_1 through δ_4 give the impact of the interventions for the first cohort and δ_5 through δ_8 give the change in impacts for the second cohort.

The regression model we estimated to assess impacts of the interventions after *teachers* have had a year of experience using them is:

(continued)

degrees of freedom instead of 67 and found that p-values increased by less than 0.001, which does not change the statistical significance of any of our findings.

$$
\begin{aligned}
(3)\quad & y_{i,j} = \alpha + \delta_1 CRISS_j + \delta_2 RA_j + \delta_3 R4R_j + \lambda_1 C_{i,j} + \lambda_2 TCH_{i,j} + \\
& \delta_4 C_{i,j} \cdot CRISS_j + \delta_5 C_{i,j} \cdot RA_j + \delta_6 C_{i,j} \cdot R4R_j + \\
& \delta_7 TCH_{i,j} \cdot CRISS_j + \delta_8 TCH_{i,j} \cdot RA_j + \delta_9 TCH_{i,j} \cdot R4R_j + \\
& \delta_{10} C_{i,j} \cdot TCH_{i,j} \cdot CRISS_j + \delta_{11} C_{i,j} \cdot TCH_{i,j} \cdot RA_j + \delta_{12} C_{i,j} \cdot TCH_{i,j} \cdot R4R_j + \\
& \beta X_{i,j} + \sum_{k=1}^{10} \gamma_k D_k + u_j + \varepsilon_{i,j}
\end{aligned}
$$

Where *TCH* is a variable that indicates whether teachers have been in the study schools for both study years and other variables are defined as above. Note that *TCH* can take on values of either 0 or 1 for both Cohorts 1 and 2. The coefficients δ_1 through δ_3 give the impact of the interventions for the first cohort and teachers who are in the study for just one year; δ_4 through δ_6 show how impacts change for the second cohort with teachers in the study for just one year; δ_7 through δ_9 show how impacts change for the first cohort with teachers in the study for two years; and δ_{10} through δ_{12} show how impacts change for second cohort students taught by teachers in the study for two years. For example, the impact for Project CRISS on second cohort students taught by teachers in the study for two years is $\delta_1 + \delta_4 + \delta_7 + \delta_{10}$.

2. Clustering of Students

To account for correlation in the error term between students within the same schools, we estimated standard errors using generalized estimating equations (GEE) in the software package R. This approach yields impact estimates that are the same as ordinary least squares (OLS) impact estimates, but adjusts the standard errors to account for clustering of students within schools. This approach also allows us to calculate cross-equation covariance terms which are used when adjusting p-values for multiple comparisons.

An alternative approach to account for clustering would be to estimate a mixed effects model using hierarchical linear modeling (HLM) or software such as SAS (using the *proc mixed* command). The difference between estimating our impact model using HLM instead of GEE is that HLM gives different weights to different schools in order to minimize the variance of the impact, whereas the GEE approach weights schools according to our random assignment probability weights. HLM yields impacts that are intended to generalize to a hypothetical super-population of schools and students whereas the GEE approach yields impacts that generalize to the schools and students selected for this study. We chose GEE instead of HLM for our benchmark approach because we believe that generalizing to a larger population is not consistent with the study design and because GEE allows for easy calculation of cross-equation covariances, which help to increase statistical power when adjusting for multiple comparisons.

3. Missing Data

We encounter missing data in two contexts. First, we encounter missing *covariate* data in our impact regressions. Second, we encounter missing *outcome* data when estimating impacts on

the GRADE and ETS follow-up tests. We discuss how each of these is addressed in the analysis below.

Missing Covariates[89]

We implemented an approach to account for missing covariates to maximize the number of observations that would contribute to the estimation of impacts of the curricula. We account for missing covariates by imputing the missing variable to the mean of the variable and including a missing value indicator in our regression equation. This approach results in unbiased impact estimates because treatment status is uncorrelated with the other covariates.[90] By using this approach, we ensure that the parameter estimate for each covariate is based only on nonmissing observations while allowing an observation that is missing data on one covariate to still contribute to estimating the effects of covariates for which that observation is not missing data. (In the context of this evaluation, the primary concern is ensuring that all observations with follow-up data contribute to the estimation of the coefficients on the treatment status indicators.) This approach may result in parameter estimates for covariates with missing data that do not fully represent the entire study sample. Because the purpose of including covariates is to increase the precision of the impact estimates, this issue has little practical significance in this context. Table G.1 shows the proportion of the sample missing each of the covariates included in our impact regressions.

Missing Follow-up Tests

Missing follow-up test score data have two potential implications. First, if students who have follow-up test score data in a treatment group are different from those who have follow-up test score data in the control group, then impacts could be biased. Evidence of this kind of bias would be either a differential rate of nonresponse between the treatment and control groups or different characteristics of respondents between treatment and control groups. Second, if students who are missing test score data are different from those who are not, then the impacts calculated for the analysis sample (that is, students who are not missing the outcome variable) might not be completely representative of students in the study sample.

Our analysis indicates that the impact estimates are unlikely to be biased due to differential nonresponse between the treatment and control groups. The proportion of students with a score on each test is between 85 and 90 percent for the post-test for Cohort 2, with no statistically

---

[89]This discussion applies only to missing covariates, such as baseline test score and race/ethnicity. It does not apply to the treatment indicator variables. The treatment indicator variables are never missing because we know the random assignment status of every school in the study.

[90]Jones (1996) derives the bias in the regression coefficient on a variable, $x_{1,}$ when missing values of a second variable, $x_{2,}$ are imputed to a constant value and a dummy variable is included in the regression. The bias is a function of the correlation between x_1 and x_2, such that if the correlation is zero, the bias is zero. In our study, x_1 is treatment status, x_2 is another covariate, and by random assignment the correlation between the two is zero (in expectation).

TABLE G.1

PROPORTION OF SAMPLE MISSING EACH COVARIATE, BY OUTCOME, YEAR 2 ANALYSES

	Composite Test Score	GRADE Score	Social Studies Reading Comprehension Score	Science Reading Comprehension Score
Cohort 2, Post-Test				
School Location[a]	0.0	0.0	0.0	0.0
Teacher Race[b]	38.0	38.1	37.7	38.1
Baseline GRADE	2.5	2.5	2.5	2.0
Baseline TOSCRF	2.6	2.6	2.6	2.2
Student ELL Status	1.9	1.9	1.9	1.7
Student Race[c]	30.2	30.2	29.9	30.3
Student Ethnicity[d]	0.6	0.6	0.6	0.6
Cohort 1, Follow Up				
School Location[a]	3.7	3.7	3.7	3.7
Teacher Race[b]	20.1	20.1	20.3	20.0
Baseline GRADE	3.7	3.7	3.6	3.8
Baseline TOSCRF	4.0	4.1	4.0	4.0
Student ELL Status	22.8	22.7	22.2	23.2
Student Race[c]	34.8	34.8	34.6	34.9
Student Ethnicity[d]	54.1	54.1	53.7	54.6

[a]School location includes indicators for "Urban," "Urban Fringe," and "Rural" locations.

[b]Teacher race includes indicators for "White," "Black," "Asian," and "Native American/Pacific Islander."

[c]Student race includes indicators for "White," "Black," "Asian," and "Native American/Pacific Islander."

[d]Student ethnicity includes an indicator for "Hispanic."

ELL = English language learner; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

significant differences between intervention and control groups (Table G.2). Among first cohort students, the proportion with a follow-up test score on each test is between 72 and 79 percent, with two statistically significant differences between the Project CRISS group and the control group (on the GRADE and ETS science comprehension assessments). The number of statistically significant differences in Table G.2 (two) is less than we would expect by random chance out of the 27 differences presented. In addition, as shown in Tables G.3-G.8, the average characteristics of students with follow-up test scores do not differ systematically among the treatment and control groups. Of the 405 comparisons made in these three tables, 14 are statistically significant (assuming independent tests, we would expect approximately 20 statistically significant differences by random chance). We conclude from these comparisons that the internal validity of the study is not threatened by missing follow-up test score data.

However, there is evidence that nonrespondents are lower achieving than respondents (Tables G.9 and G.10). Specifically, we see evidence that nonrespondents have lower baseline test scores, are more likely to be overage for grade, and have more absences from school then respondents. We also see evidence that nonrespondents differ from respondents in terms of race and gender.

We used nonresponse weights to account for these differences in baseline characteristics of students who do and do not have a follow-up test. These weights are described in detail in Section 5.

4. Multiple Comparisons

In this study, making clear distinctions between effects that are real and those that are due to chance is complicated by the issue of multiple comparisons. By comparing multiple intervention groups to a control group, for multiple outcomes, the probability that one of those differences will appear to be statistically significant is greater than the probability that a single difference will appear statistically significant. Intuitively, this is similar to the difference between the probability of a *single* toss of a coin yielding heads and the probability that *at least one of several* coin tosses will yield heads.

Our benchmark approach to adjusting *p*-values to account for multiple comparisons begins with the establishment of several different sets, or domains, of multiple tests. Each domain pertains to a separate research question. We then adjust *p*-values for tests within these domains so that we control the probability of drawing a false conclusion. The domains are described in Chapters III and IV.

Within domains we calculate *p*-values using a generalized version of the Dunnett (1955) adjustment. Dunnett's approach takes into account correlations between tests due to a shared control group, drawing critical values based on a multivariate t-distribution. Hothorn, Bretz, and Westfall (2008) implement a more generalized procedure that is also based on a multivariate t-distribution but adjusts *p*-values for multiple tests taking into account correlations that arise for *any* reason (not just a common control group). We use this approach to adjust for both multiple treatment groups and multiple outcomes. For the secondary analyses described below, we also adjust for multiple subgroups. This procedure requires covariance estimates between all impacts,

TABLE G.2

PROPORTION OF STUDENTS WITH TEST SCORES IN YEAR 2, BY EXPERIMENTAL CONDITION

Follow-Up Tests	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Cohort 2, Post-Test						
GRADE	89.0	88.0	90.0	87.0	n.a.	88.0
		(0.69)	(0.75)	(0.59)		(0.86)
ETS Social Studies Comprehension	89.9	88.0	87.0	85.0	n.a.	87.0
		(0.62)	(0.51)	(0.22)		(0.45)
ETS Science Comprehension	88.0	85.0	90.0	87.0	n.a.	88.0
		(0.40)	(0.48)	(0.70)		(0.84)
Cohort 1, Follow Up						
GRADE	73.0	79.0*	77.0	73.0	75.0	76.0
		(0.03)	(0.23)	(0.85)	(0.57)	(0.25)
ETS Social Studies Comprehension	72.0	78.0	75.0	74.0	74.0	75.0
		(0.09)	(0.41)	(0.71)	(0.73)	(0.33)
ETS Science Comprehension	72.0	78.0*	76.0	72.0	76.0	76.1
		(0.04)	(0.21)	(0.84)	(0.28)	(0.16)

SOURCE: Reading comprehension tests administered by study team.

NOTE: The *p-values* from t-tests of treatment and control group differences in means are presented in parentheses. These tests account for clustering of students within schools. Students in the study were randomly assigned to take *either* the ETS social studies reading comprehension assessment *or* the ETS science reading comprehension assessment. The GRADE was administered to all students in the study.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; n.a. = not applicable.

*Statistically different at the .05 level.

TABLE G.3

AVERAGE BASELINE CHARACTERISTICS OF COHORT 2 STUDENTS WITH GRADE POST-TEST SCORES, BY EXPERIMENTAL CONDITION

	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Percentage in Study Schools at Baseline	98.0	96.0 (0.10)	98.0 (0.97)	98.0 (0.62)	97.0 (0.34)
GRADE Score (Average)	100.4	102.0 (0.37)	100.7 (0.86)	100.7 (0.85)	101.1 (0.62)
TOSCRF Score (Average)	88.9	90.0 (0.47)	88.9 (1.00)	89.9 (0.59)	89.5 (0.67)
Female (Percentage)	50.0	50.0 (0.90)	52.0 (0.36)	49.0 (0.73)	51.0 (0.68)
Age (Average)	10.7	10.7 (0.31)	10.6 (0.85)	10.7 (0.11)	10.7 (0.73)
Overage (Percentage)[a]	18.0	22.0 (0.21)	17.0 (0.84)	24.0 (0.06)	19.0 (0.60)
Hispanic (Percentage)	26.0	25.0 (0.93)	32.0 (0.65)	15.0 (0.42)	30.0 (0.72)
Race (Percentage)					
White	31.0	41.0 (0.47)	32.0 (0.62)	33.0 (0.70)	37.0 (0.72)
Black	45.0	37.0 (0.47)	45.0 (0.62)	51.0 (0.70)	42.0 (0.72)
Asian	2.0	2.0 (0.47)	3.0 (0.62)	1.0 (0.70)	2.0 (0.72)
Native American	22.0	18.0 (0.47)	17.0 (0.62)	14.0 (0.70)	17.0 (0.72)
Number of Days Absent in Prior School Year (Average)	8.3	7.7 (0.57)	7.9 (0.73)	7.0 (0.22)	7.7 (0.56)
Eligible for Free or Reduced-Price Lunch (Percentage)	76.0	69.0 (0.33)	72.0 (0.60)	75.0 (0.94)	72.0 (0.57)
Classified as English Language Learner (Percentage)	17.0	17.0 (0.94)	16.0 (0.97)	4.0* (0.00)	16.0 (0.84)
Identified as Having a Disability (Percentage)[b]	11.0	11.0 (0.93)	10.0 (0.77)	16.0 (0.31)	12.0 (0.79)
Number of Students[c]	**1,194**	**1,201**	**1,108**	**639**	**2,948**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

Table G.3 *(continued)*

NOTE: Baseline for students in Cohort 2 was fall 2007. Post-test data for Cohort 2 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of treatment and control group differences in means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2007.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number with GRADE post-test scores. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

TABLE G.4

AVERAGE BASELINE CHARACTERISTICS OF COHORT 1 STUDENTS WITH FOLLOW-UP GRADE SCORES, BY EXPERIMENTAL CONDITION

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Percentage in Study Schools at Baseline	98.0	95.0* (0.01)	98.0 (0.60)	96.0 (0.05)	98.0 (0.83)	97.0 (0.06)
GRADE Score (Average)	100.5	101.9 (0.42)	100.0 (0.71)	99.5 (0.40)	101.2 (0.70)	100.6 (0.92)
TOSCRF Score (Average)	88.7	89.5 (0.50)	88.4 (0.66)	88.0 (0.41)	89.9 (0.34)	89.0 (0.79)
Female (Percentage)	49.0	52.0 (0.10)	53.0* (0.05)	52.0 (0.17)	49.0 (0.93)	52.0 (0.09)
Age (Average)	10.7	10.7 (0.39)	10.7 (0.72)	10.8 (0.19)	10.7 (0.77)	10.7 (0.39)
Overage (Percentage)[a]	20.0	22.0 (0.54)	21.0 (0.77)	24.0 (0.33)	21.0 (0.73)	22.0 (0.48)
Hispanic (Percentage)	74.0	73.0 (0.89)	77.0 (0.68)	76.0 (0.87)	72.0 (0.82)	75.0 (0.93)
Race (Percentage)						
White	37.0	43.0 (0.92)	38.0 (1.00)	42.0 (0.58)	44.0 (0.70)	42.0 (0.85)
Black	42.0	39.0 (0.92)	44.0 (1.00)	43.0 (0.58)	41.0 (0.70)	42.0 (0.85)
Asian	3.0	2.0 (0.92)	3.0 (1.00)	2.0 (0.58)	2.0 (0.70)	3.0 (0.85)
Native American	17.0	14.0 (0.92)	14.0 (1.00)	11.0 (0.58)	10.0 (0.70)	13.0 (0.85)
Number of Days Absent in Prior School Year (Average)	10.7	10.1 (0.83)	10.8 (1.00)	14.4 (0.48)	10.7 (0.99)	11.4 (0.82)
Eligible for Free or Reduced-Price Lunch (Percentage)	60.0	59.0 (0.92)	61.0 (0.84)	58.0 (0.82)	59.0 (0.91)	59.0 (0.96)
Classified as English Language Learner (Percentage)	26.0	26.0 (0.99)	31.0 (0.71)	34.0 (0.63)	25.0 (0.91)	29.0 (0.74)
Identified as Having a Disability (Percentage)[b]	10.0	9.0 (0.68)	11.0 (0.81)	11.0 (0.87)	11.0 (0.82)	10.0 (0.99)
Number of Students[c]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

Table G.4 *(continued)*

NOTE: Baseline for students in Cohort 1 was fall 2006. Follow-up data for Cohort 1 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of treatment and control group differences in means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2006.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number with follow-up GRADE scores. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

TABLE G.5

AVERAGE BASELINE CHARACTERISTICS OF COHORT 2 STUDENTS WITH SOCIAL STUDIES READING COMPREHENSION POST-TEST SCORES, BY EXPERIMENTAL CONDITION

	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Percentage in Study Schools at Baseline	97.0	96.0 (0.50)	98.0 (0.20)	98.0 (0.27)	97.0 (0.81)
GRADE Score (Average)	100.4	102.2 (0.32)	100.5 (0.95)	100.5 (0.96)	101.0 (0.63)
TOSCRF Score (Average)	88.6	90.1 (0.33)	89.1 (0.67)	90.0 (0.45)	89.6 (0.44)
Female (Percentage)	49.0	52.0 (0.36)	50.0 (0.80)	51.0 (0.68)	51.0 (0.37)
Age (Average)	10.7	10.7 (0.61)	10.6 (0.46)	10.7 (0.41)	10.7 (0.79)
Overage (Percentage)[a]	19.0	22.0 (0.43)	18.0 (0.63)	24.0 (0.16)	20.0 (0.93)
Hispanic (Percentage)	26.0	25.0 (0.98)	33.0 (0.61)	16.0 (0.49)	30.0 (0.68)
Race (Percentage)					
White	29.0	39.0 (0.60)	31.0 (0.91)	31.0 (0.61)	36.0 (0.86)
Black	45.0	37.0 (0.60)	47.0 (0.91)	54.0 (0.61)	43.0 (0.86)
Asian	3.0	2.0 (0.60)	3.0 (0.91)	2.0 (0.61)	3.0 (0.86)
Native American	23.0	19.0 (0.60)	16.0 (0.91)	13.0 (0.61)	16.0 (0.86)
Number of Days Absent in Prior School Year (Average)	8.5	8.1 (0.68)	7.8 (0.61)	6.8 (0.15)	7.8 (0.51)
Eligible for Free or Reduced-Price Lunch (Percentage)	76.0	68.0 (0.26)	74.0 (0.79)	77.0 (0.88)	73.0 (0.63)
Classified as English Language Learner (Percentage)	18.0	17.0 (0.86)	18.0 (0.96)	4.0* (0.00)	16.0 (0.79)
Identified as Having a Disability (Percentage)[b]	10.0	11.0 (0.81)	11.0 (0.92)	17.0 (0.23)	12.0 (0.56)
Number of Students[c]	**1,194**	**1,201**	**1,108**	**639**	**2,948**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

Table G.5 *(continued)*

NOTE: Baseline for students in Cohort 2 was fall 2007. Post-test data for Cohort 2 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of treatment and control group differences in means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2007.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number with social studies reading comprehension post-test scores. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

TABLE G.6

AVERAGE BASELINE CHARACTERISTICS OF COHORT 1 STUDENTS WITH FOLLOW-UP SOCIAL STUDIES READING COMPREHENSION SCORES, BY EXPERIMENTAL CONDITION

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Percentage in Study Schools at Baseline	98.0	96.0* (0.02)	98.0 (0.59)	96.0* (0.01)	98.0 (0.61)	97.0* (0.05)
GRADE Score (Average)	100.8	101.9 (0.50)	100.0 (0.61)	99.6 (0.33)	101.0 (0.87)	100.6 (0.89)
TOSCRF Score (Average)	88.7	89.5 (0.54)	88.9 (0.79)	88.4 (0.80)	89.9 (0.23)	89.2 (0.55)
Female (Percentage)	47.0	54.0* (0.05)	55.0* (0.01)	52.0 (0.17)	48.0 (0.83)	52.0 (0.06)
Age (Average)	10.7	10.7 (0.52)	10.7 (0.75)	10.8 (0.12)	10.7 (0.93)	10.7 (0.46)
Overage (Percentage)[a]	21.0	22.0 (0.71)	21.0 (0.85)	25.0 (0.34)	21.0 (0.87)	22.0 (0.59)
Hispanic (Percentage)	74.0	74.0 (0.94)	78.0 (0.65)	75.0 (0.94)	72.0 (0.79)	75.0 (0.93)
Race (Percentage)						
White	37.0	42.0 (0.68)	36.0 (0.98)	42.0 (0.21)	45.0 (0.38)	41.0 (0.52)
Black	41.0	39.0 (0.68)	45.0 (0.98)	44.0 (0.21)	43.0 (0.38)	42.0 (0.52)
Asian	4.0	2.0 (0.68)	3.0 (0.98)	2.0 (0.21)	2.0 (0.38)	2.0 (0.52)
Native American	17.0	15.0 (0.68)	15.0 (0.98)	11.0 (0.21)	10.0 (0.38)	13.0 (0.52)
Number of Days Absent in Prior School Year (Average)	10.4	10.0 (0.91)	9.8 (0.87)	13.9 (0.51)	10.7 (0.92)	11.0 (0.82)
Eligible for Free or Reduced-Price Lunch (Percentage)	60.0	59.0 (0.82)	59.0 (0.87)	59.0 (0.92)	58.0 (0.71)	59.0 (0.81)
Classified as English Language Learner (Percentage)	25.0	27.0 (0.86)	30.0 (0.63)	33.0 (0.63)	25.0 (0.99)	29.0 (0.66)
Identified as Having a Disability (Percentage)[b]	10.0	9.0 (0.66)	12.0 (0.71)	12.0 (0.68)	11.0 (0.97)	11.0 (0.92)
Number of Students[c]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

Table G.6 *(continued)*

NOTE: Baseline for students in Cohort 1 was fall 2006. Follow-up data for Cohort 1 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of treatment and control group differences in means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2006.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number with follow-up social studies reading comprehension scores. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

TABLE G.7

AVERAGE BASELINE CHARACTERISTICS OF COHORT 2 STUDENTS WITH SCIENCE READING COMPREHENSION POST-TEST SCORES, BY EXPERIMENTAL CONDITION

	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Percentage in Study Schools at Baseline	99.0	96.0* (0.02)	97.0 (0.16)	99.0 (0.97)	0.97* (0.04)
GRADE Score (Average)	100.5	101.7 (0.52)	100.9 (0.79)	101.0 (0.77)	101.1 (0.66)
TOSCRF Score (Average)	89.3	89.8 (0.75)	88.8 (0.70)	89.9 (0.77)	89.4 (0.97)
Female (Percentage)	51.0	50.0 (0.71)	54.0 (0.38)	47.0 (0.36)	50.0 (0.84)
Age (Average)	10.6	10.7 (0.16)	10.6 (0.79)	10.8* (0.02)	10.7 (0.38)
Overage (Percentage)[a]	16.0	23.0 (0.12)	16.0 (0.97)	24.0* (0.04)	19.0 (0.39)
Hispanic (Percentage)	27.0	25.0 (0.84)	32.0 (0.71)	14.0 (0.36)	30.0 (0.79)
Race (Percentage)					
White	33.0	42.0 (0.58)	33.0 (0.93)	36.0 (.)	38.0 (0.91)
Black	44.0	37.0 (0.58)	44.0 (0.93)	48.0 (.)	41.0 (0.91)
Asian	2.0	2.0 (0.58)	3.0 (0.93)	0.0 (.)	2.0 (0.91)
Native American	21.0	17.0 (0.58)	18.0 (0.93)	15.0 (.)	17.0 (0.91)
Number of Days Absent in Prior School Year (Average)	8.2	7.1 (0.33)	7.8 (0.81)	7.2 (0.39)	7.5 (0.56)
Eligible for Free or Reduced-Price Lunch (Percentage)	75.0	70.0 (0.52)	70.0 (0.47)	74.0 (0.83)	72.0 (0.59)
Classified as English Language Learner (Percentage)	16.0	18.0 (0.74)	15.0 (0.85)	4.0* 0.00	15.0 (0.87)
Identified as Having a Disability (Percentage)[b]	11.0	10.0 (0.75)	10.0 (0.60)	16.0 (0.46)	11.0 (1.00)
Number of Students[c]	**1,194**	**1,201**	**1,108**	**639**	**2,948**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

Table G.7 *(continued)*

NOTE: Baseline for students in Cohort 2 was fall 2007. Post-test data for Cohort 2 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of treatment and control group differences in means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2007.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number with science reading comprehension post-test scores. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

TABLE G.8

AVERAGE BASELINE CHARACTERISTICS OF COHORT 1 STUDENTS WITH FOLLOW-UP SCIENCE READING COMPREHENSION SCORES, BY EXPERIMENTAL CONDITION

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Percentage in Study Schools at Baseline	98.0	95.0 (0.05)	98.0 (0.87)	96.0 (0.29)	98.0 (0.93)	97.0 (0.28)
GRADE Score (Average)	100.5	101.9 (0.43)	100.1 (0.78)	99.4 (0.39)	101.6 (0.55)	100.7 (0.83)
TOSCRF Score (Average)	88.7	89.5 (0.44)	87.8 (0.34)	87.6 (0.20)	89.9 (0.39)	88.7 (0.94)
Female (Percentage)	50.0	51.0 (0.72)	49.0 (0.99)	53.0 (0.38)	50.0 (0.93)	51.0 (0.61)
Age (Average)	10.7	10.7 (0.45)	10.7 (0.90)	10.7 (0.37)	10.7 (0.57)	10.7 (0.48)
Overage (Percentage)[a]	20.0	21.0 (0.65)	20.0 (0.87)	24.0 (0.40)	21.0 (0.76)	21.0 (0.60)
Hispanic (Percentage)	74.0	72.0 (0.81)	77.0 (0.71)	77.0 (0.78)	72.0 (0.79)	75.0 (0.95)
Race (Percentage)						
White	37.0	43.0 (0.97)	38.0 (0.89)	42.0 (0.90)	43.0 (0.84)	42.0 (0.89)
Black	42.0	38.0 (0.97)	43.0 (0.89)	42.0 (0.90)	41.0 (0.84)	41.0 (0.89)
Asian	2.0	3.0 (0.97)	4.0 (0.89)	2.0 (0.90)	3.0 (0.84)	3.0 (0.89)
Native American	17.0	14.0 (0.97)	14.0 (0.89)	12.0 (0.90)	10.0 (0.84)	13.0 (0.89)
Number of Days Absent in Prior School Year (Average)	10.9	9.9 (0.75)	11.5 (0.85)	14.8 (0.44)	10.6 (0.90)	11.5 (0.79)
Eligible for Free or Reduced-Price Lunch (Percentage)	58.0	59.0 (0.88)	63.0 (0.46)	57.0 (0.83)	59.0 (0.89)	60.0 (0.77)
Classified as English Language Learner (Percentage)	27.0	26.0 (0.89)	31.0 (0.75)	35.0 (0.59)	25.0 (0.84)	29.0 (0.78)
Identified as Having a Disability (Percentage)[b]	11.0	9.0 (0.48)	11.0 (0.91)	10.0 (0.76)	11.0 (0.79)	10.0 (0.70)
Number of Students[c]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

Table G.8 *(continued)*

NOTE: Baseline for students in Cohort 1 was fall 2006. Follow-up data for Cohort 1 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of treatment and control group differences in means are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2006.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number with follow-up science reading comprehension scores. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE G.9

BASELINE CHARACTERISTICS OF COHORT 2 STUDENTS WITH AND WITHOUT POST-TEST SCORES

	GRADE		Social Studies Reading Comprehension		Science Reading Comprehension	
	Students with a Score	Students Without a Score	Students with a Score	Students Without a Score	Students with a Score	Students Without a Score
Percentage in Study Schools at Baseline	97.0	99.0* (0.02)	97.0	98.0 (0.39)	97.0	97.0 (1.00)
GRADE Score (Average)	101.0	98.7* (0.00)	101.0	100.5 (0.11)	101.1	100.5* (0.02)
TOSCRF Score (Average)	89.4	87.3* (0.00)	89.4	89.0 (0.11)	89.4	89.0 (0.09)
Female (Percentage)	50.0	46.0* (0.03)	50.0	49.0 (0.43)	51.0	49.0 (0.38)
Age (Average)	10.7	10.9* (0.00)	10.7	10.7* (0.01)	10.7	10.7* (0.00)
Overage (Percentage)[a]	20.0	34.0* (0.00)	20.0	22.0* (0.04)	19.0	23.0* (0.00)
Hispanic (Percentage)	26.0	28.0 (0.75)	26.0	26.0 (0.68)	26.0	26.0 (0.92)
Race (Percentage)						
White	35.0	39.0* (0.00)	33.0	37.0* (0.01)	36.0	34.0* (0.01)
Black	44.0	50.0* (0.00)	44.0	44.0* (0.01)	43.0	46.0* (0.01)
Asian	2.0	2.0* (0.00)	3.0	2.0* (0.01)	2.0	2.0* (0.01)
Native American	18.0	8.0* (0.00)	18.0	16.0* (0.01)	18.0	16.0* (0.01)
Number of Days Absent in Prior School Year (Average)	7.8	9.7* (0.01)	7.9	8.1 (0.39)	7.6	8.3* (0.01)
Eligible for Free or Reduced-Price Lunch (Percentage)	72.0	75.0 (0.31)	73.0	73.0 (0.63)	72.0	73.0 (0.32)
Classified as English Language Learner (Percentage)	15.0	17.0 (0.47)	15.0	15.0 (0.64)	14.0	15.0 (0.25)
Identified as Having a Disability (Percentage)[b]	12.0	10.0 (0.48)	12.0	11.0 (0.54)	11.0	12.0 (0.62)
Number of Students[c]	**3,664**	**478**	**1,825**	**2,317**	**1,816**	**2,326**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

NOTE: Baseline for students in Cohort 2 was fall 2007. Post-test data for Cohort 2 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of differences in means between students with and without test scores are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2007.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number participating in the study. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

TABLE G.10

BASELINE CHARACTERISTICS OF COHORT 1 STUDENTS WITH AND WITHOUT FOLLOW-UP TEST SCORES

	GRADE		Social Studies Reading Comprehension		Science Reading Comprehension	
	Students with a Score	Students Without a Score	Students with a Score	Students Without a Score	Students with a Score	Students Without a Score
Percentage in Study Schools at Baseline	97.0	92.0* (0.00)	97.0	95.0* (0.00)	97.0	95.0* (0.00)
GRADE Score (Average)	100.6	98.4* (0.00)	100.7	99.8* (0.00)	100.7	99.7* (0.00)
TOSCRF Score (Average)	88.9	87.2* (0.00)	89.1	88.1* (0.00)	88.7	88.4 (0.21)
Female (Percentage)	51.0	45.0* (0.00)	51.0	49.0* (0.02)	50.0	49.0 (0.31)
Age (Average)	10.7	10.8* (0.00)	10.7	10.7 (0.07)	10.7	10.7* (0.02)
Overage (Percentage)[a]	22.0	26.0* (0.00)	22.0	24.0 (0.16)	21.0	24.0* (0.00)
Hispanic (Percentage)	74.0	71.0 (0.08)	75.0	73.0 (0.08)	75.0	73.0 (0.14)
Race (Percentage)						
White	41.0	42.0 (0.62)	40.0	41.0 (0.28)	41.0	41.0 (0.23)
Black	42.0	42.0 (0.62)	42.0	42.0 (0.28)	41.0	42.0 (0.23)
Asian	3.0	2.0 (0.62)	2.0	3.0 (0.28)	3.0	2.0 (0.23)
Native American	13.0	12.0 (0.62)	14.0	13.0 (0.28)	14.0	13.0 (0.23)
Number of Days Absent in Prior School Year (Average)	11.3	13.6* (0.03)	11.0*	12.4* (0.03)	11.5	12.1 (0.21)
Eligible for Free or Reduced-Price Lunch (Percentage)	59.0	59.0 (0.90)	59.0	59.0 (0.77)	60.0	59.0 (0.67)
Classified as English Language Learner (Percentage)	29.0	26.0 (0.27)	28.0	28.0 (0.84)	29.0	27.0 (0.20)
Identified as Having a Disability (Percentage)[b]	11.0	12.0 (0.17)	11.0	11.0 (0.76)	10.0	11.0 (0.25)
Number of Students[c]	**5,572**	**777**	**2,759**	**3,590**	**2,746**	**3,603**

SOURCE: Student Records Form; baseline GRADE and TOSCRF tests administered by study team.

NOTE: Baseline for students in Cohort 1 was fall 2006. Follow-up data for Cohort 1 students were collected in spring 2008. Baseline characteristics are reported only for students who were present in study schools at baseline. The *p-values* from tests of differences in means between students with and without test scores are presented in parentheses. These tests account for clustering of students within schools.

[a]We considered a fifth grader to be overage for grade if he or she was 11 or older as of September 1, 2006.

[b]A student was identified as having a disability if any of the following categories were indicated on the Student Records Form: autism, deaf-blindness, developmental delay, emotional disturbance, hearing impairment, learning disability, mental retardation, orthopedic impairment, other health impairment, speech or language impairment, traumatic brain injury, visual impairment, and other disability not included in this list.

[c]The number of students presented in this row is the number participating in the study. Response rates vary across items.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level.

which is why cross-equation covariances are needed (and is one reason we chose GEE instead of HLM for our benchmark model).

5. Weights

Accounting for nonresponse and random assignment probabilities in our benchmark models required the use of weights with two components. The overall weight used in the analysis is the product of these two components.[91]

The first component involves weighting by the inverse of random assignment probabilities. In districts where the number of schools is evenly divisible by five, every school has an equal chance of being assigned to one of the five experimental conditions (four treatment groups and one control group). However, in districts where the number of schools is not evenly divisible by five, we conducted random assignment such that the probability of being assigned to the control group is higher than the probability of being assigned to any given treatment group.[92] We take these assignment probabilities into account in our analysis so that all five experimental groups are balanced in terms of their representation of school districts. For the fifth-grade component of the second year of the study, we calculate the weights as if there were three intervention groups (since the Reading for Knowledge intervention was not included in that component).[93]

The second component of the weight involves accounting for nonresponse to adjust for differences in baseline characteristics of students who do and do not have a post-test (or follow-up) test (as described above in Section 3). For each post-test (or follow-up) test score, we estimated a propensity regression model where the outcome is a binary variable that equals one if a student has a post-test (or follow-up) test score and zero otherwise. We calculated the expected probability of having a post-test (or follow-up) test score for every student using baseline

[91]In all, eight weights were created for each of the study's second year components. Weights were created for each of the study's four test scores (ETS science comprehension, ETS social studies comprehension, GRADE, and the composite). Weights for each of the four test scores were created in two ways, corresponding to the two types of comparisons being made: (1) the pooled treatment group versus the control group and (2) all pairwise comparisons (both between treatment groups and between each treatment group and the control group).

[92]If all schools in the control group within a district left the study, we would lose the ability to calculate any impacts in that district. To reduce the chance of this happening, we chose to assign "extra" schools in a district to the control group.

[93]For convenience, we use the same final weight in all analyses. However in some analyses, for example impact regressions that include school district dummy variables, the component of the final weight that reflects variation in random assignment probabilities is not needed and has no effect.

data.[94,95] We then created a weight that is inversely proportional to the probability of having a post-test (or follow-up) test score, meaning that students with a lower probability of having a post-test (or follow-up) test score are weighted more heavily in our analysis.

B. BENCHMARK APPROACH TO CALCULATING SECONDARY IMPACTS

The secondary analyses examine how impacts vary by student and teacher characteristics, school conditions, and teacher practices. Each of these analyses is implemented by interacting the treatment dummy variables in equations 1, 2, and 3 with subgroup dummy variables. However, the interpretation of these impacts differs depending on whether the subgroup is defined at baseline or could itself be affected by the interventions. Subgroups defined by student characteristics (such as baseline test scores), teacher characteristics (such as years of experience), and school conditions (such as concentration of ELL students in the school) cannot be affected by the intervention. Impacts for these subgroups can be interpreted as causal. Subgroups defined by teacher practices, self-reported past professional development, teaching efficacy, and school professional culture, however, could be affected by the interventions, which complicates interpretation because the treatment and control groups are no longer equivalent within those subgroups. Impacts for these subgroups cannot be interpreted as causal.

The benchmark approach for the secondary analysis is the same as for the primary analysis in all ways but one. The secondary analysis uses the same approach for regression adjustment, clustering, missing data, and weights. The only difference in the benchmark approach between the secondary analysis and the primary analysis is how we deal with multiple comparisons. For the secondary analysis, we do not adjust for multiple comparisons across all subgroups. We adjust only for multiple comparisons within each subgroup analysis. This is described in Chapters III and IV.

---

[94]The baseline data used in the propensity score models included students' demographic characteristics (age, gender, race, ethnicity, whether the student is disabled, and whether the student received any reading services), students' baseline scores on the GRADE and TOSCRF assessments, characteristics of each student's teacher (degree and experience), and characteristics of each student's school (percentage of students eligible for free or reduced-price lunch and percentage of students classified as English language learners). Only those characteristics that were statistically significant were kept in the final model for each of the eight weights.

[95]Because of the extent to which baseline test scores are associated with nonresponse (see Tables G.9 and G.10), separate nonresponse models were estimated for students *without* baseline test score data. Because of the small number of students that fell into this category, a weighting class approach was used to develop nonresponse weights for these students. In this method, students are assigned to cells based on their characteristics and then the respondents in each cell are essentially weighted up to represent the nonrespondents in that cell. The same set of characteristics listed above (with the exception of baseline test scores) was used in this approach.

This page is intentionally left blank.

APPENDIX H

ASSESSING ROBUSTNESS OF THE IMPACTS

This page is intentionally left blank.

This appendix describes the robustness of our impact estimates to variations in the benchmark model described in Appendix G and to additional issues that might influence our findings.

A. ROBUSTNESS OF THE BENCHMARK APPROACH

The benchmark approach reflects the methodological choices we made to calculate impacts. While we think these are the best methodological choices for this study, there are valid alternatives to many of these choices that could potentially alter our findings. In this section we assess the sensitivity of our findings to variations in our benchmark model. Specifically, we assess sensitivity to:

1. **Inclusion of Covariates.** The statistical significance of the findings shown in Chapters III and IV are not sensitive to the inclusion of covariates (Table H.1).

2. **Use of Nonresponse Weights.** The statistical significance of the findings shown in Chapters III and IV are not sensitive to the use of nonresponse weights (Table H.1).

3. **Approach to Adjusting for Clustering.** Our findings are not sensitive to the method used to account for clustering. A comparison of the estimates generated by generalized estimating equations (GEE, our benchmark approach) and hierarchical linear modeling (HLM) shows that our findings would not have been different if we had used HLM instead of GEE (Table H.2). In particular, using HLM does not affect the statistical significance of the findings reported in Chapters III and IV. When examining the statistical significance of the findings in Table H.2, we used the Bonferroni adjustment to account for multiple comparisons, because it was not possible to obtain the cross-equation variances from HLM that were needed for the multiple comparison procedures used in the benchmark models presented in Chapters III and IV.

B. SENSITIVITY TO ADDITIONAL ISSUES

After completing our descriptive and impact analyses, we identified several additional issues to investigate through sensitivity analysis. Below we list these issues and the results of our sensitivity analyses.

Students with Only Baseline and Follow-up Tests

Restricting the analysis sample to only students with both baseline and follow-up tests does not change the statistical significance of the study's findings. The positive effect of ReadAbout on the social studies post-test for Cohort 2 students whose teachers had one year of experience with the study curricula remains statistically significant (Table H.3). All of the other findings remain statistically insignificant.

TABLE H.1

SENSITIVITY OF IMPACT ESTIMATES TO ALTERNATIVE SPECIFICATIONS

	Difference in Spring Test Scores Between Each of the Following and the Control Group:				
	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Composite Test Score[a]					
Benchmark model[b]					
Cohort 2, Post-Test, School Experience	0.00	0.05	-0.02	n.a.	0.02
Cohort 2, Post-Test, Teacher Experience	0.02	0.09	0.03	n.a.	0.05
Cohort 1, Follow Up	-0.01	0.00	0.06	0.05	0.02
Model with no covariates					
Cohort 2, Post-Test, School Experience	-0.06	0.10	-0.06	n.a.	0.01
Cohort 2, Post-Test, Teacher Experience	-0.01	0.11	0.00	n.a.	0.04
Cohort 1, Follow Up	0.04	-0.00	0.05	0.09	0.04
Model with weights that adjust for random assignment probability but *not* nonresponse					
Cohort 2, Post-Test, School Experience	0.01	0.05	0.00	n.a.	0.02
Cohort 2, Post-Test, Teacher Experience	0.05	0.08	0.04	n.a.	0.06
Cohort 1, Follow Up	0.01	0.01	0.05	0.05	0.03
GRADE Score					
Benchmark model[b]					
Cohort 2, Post-Test, School Experience	-0.28	-0.08	-0.56	n.a.	-0.26
Cohort 2, Post-Test, Teacher Experience	0.16	0.24	-0.21	n.a.	0.08
Cohort 1, Follow Up	-0.75	-0.14	0.52	0.31	-0.04
Model with no covariates					
Cohort 2, Post-Test, School Experience	-0.95	0.78	-1.07	n.a.	-0.32
Cohort 2, Post-Test, Teacher Experience	-0.47	0.90	-0.58	n.a.	0.11
Cohort 1, Follow Up	0.25	-0.13	0.44	0.86	0.36
Model with weights that adjust for random assignment probability but *not* nonresponse					
Cohort 2, Post-Test, School Experience	-0.52	-0.13	-0.50	n.a.	-0.32
Cohort 2, Post-Test, Teacher Experience	-0.10	0.12	-0.12	n.a.	0.01
Cohort 1, Follow Up	-0.79	-0.17	0.19	0.22	-0.14
Social Studies Reading Comprehension Assessment Score					
Benchmark model[b]					
Cohort 2, Post-Test, School Experience	0.09	4.63	0.47	n.a.	2.21
Cohort 2, Post-Test, Teacher Experience	0.27	6.43*	3.03	n.a.	3.25
Cohort 1, Follow Up	1.42	-0.65	1.70	3.22	1.08
Model with no covariates					
Cohort 2, Post-Test, School Experience	-1.37	5.85	-1.02	n.a.	1.72
Cohort 2, Post-Test, Teacher Experience	-0.35	8.01*	1.71	n.a.	3.46
Cohort 1, Follow Up	0.22	-0.96	0.71	3.43	0.87

Table H.1 *(continued)*

	Difference in Spring Test Scores Between Each of the Following and the Control Group:				
	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Model with weights that adjust for random assignment probability but *not* nonresponse					
Cohort 2, Post-Test, School Experience	0.57	4.80	0.78	n.a.	2.43
Cohort 2, Post-Test, Teacher Experience	1.01	6.48*	3.18	n.a.	3.49
Cohort 1, Follow Up	1.78	-0.67	1.17	2.97	0.95
Science Reading Comprehension Assessment Score					
Benchmark model[b]					
Cohort 2, Post-Test, School Experience	0.58	1.66	-0.31	n.a.	0.83
Cohort 2, Post-Test, Teacher Experience	0.08	0.10	0.07	n.a.	0.08
Cohort 1, Follow Up	1.37	1.92	3.18	1.35	2.23
Model with no covariates					
Cohort 2, Post-Test, School Experience	-2.11	3.03	-1.94	n.a.	0.01
Cohort 2, Post-Test, Teacher Experience	0.39	3.55	0.38	n.a.	1.75
Cohort 1, Follow Up	1.96	1.32	2.50	2.84	2.14
Model with weights that adjust for random assignment probability but *not* nonresponse					
Cohort 2, Post-Test, School Experience	0.83	1.46	-0.35	n.a.	0.71
Cohort 2, Post-Test, Teacher Experience	2.14	2.38	1.68	n.a.	1.99
Cohort 1, Follow Up	3.10	2.95	4.62	1.90	3.33

SOURCE: Reading comprehension tests administered by study team.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]The "benchmark" model includes weights that adjust for nonresponse and random assignment probability and the following covariates: pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher gender, teacher age, teacher race, and district indicators.

GRADE = Group Reading Assessment and Diagnostic Evaluation; n.a. = not applicable; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE H.2

COMPARISON OF BENCHMARK AND HLM MODELS

	Impact and Standard Error for Each of the Following:									
	Project CRISS		ReadAbout		Read for Real		Reading for Knowledge		Combined Treatment Group	
	Impact	Std. Error	Impact	Std. Error	Impact	Std. Error	Impact	Std. Error	Impact	Std. Error
Composite Test Score[a]										
Cohort 2, Post-Test, School Experience										
Benchmark	-0.01	0.05	0.04	0.04	-0.02	0.05	n.a.	n.a.	0.02	0.04
HLM	0.03	0.04	0.09	0.04	0.01	0.05	n.a.	n.a.	0.05	0.03
Cohort 1, Follow Up										
Benchmark	-0.02	0.04	0.00	0.04	0.05	0.04	0.05	0.04	0.02	0.03
HLM	0.01	0.05	0.01	0.04	0.07	0.05	0.06	0.04	0.03	0.03
GRADE Score										
Cohort 2, Post-Test, School Experience										
Benchmark	-0.34	0.68	-0.17	0.61	-0.58	0.71	n.a.	n.a.	-0.28	0.50
HLM	0.09	0.58	0.41	0.58	-0.32	0.64	n.a.	n.a.	0.06	0.39
Cohort 1, Follow Up										
Benchmark	-0.91	0.58	-0.27	0.61	0.32	0.66	0.20	0.51	-0.15	0.41
HLM	-0.84	0.67	-0.27	0.63	0.27	0.66	0.26	0.64	-0.01	0.42
Social Studies Reading Comprehension Assessment Score										
Cohort 2, Post-Test, School Experience										
Benchmark	-0.11	1.95	4.38	1.47	0.32	2.14	n.a.	n.a.	2.07	1.35
HLM	0.94	1.85	6.45	1.84	1.50	2.08	n.a.	n.a.	3.70	1.51
Cohort 1, Follow Up										
Benchmark	1.30	1.85	-0.65	1.63	1.44	1.74	3.17	2.19	1.19	1.22
HLM	2.03	2.10	-0.44	2.03	1.92	2.09	3.03	2.02	1.30	1.18
Science Reading Comprehension Assessment Score										
Cohort 2, Post-Test, School Experience										
Benchmark	0.35	2.07	1.57	2.29	-0.40	2.22	n.a.	n.a.	0.77	1.70
HLM	-0.55	0.73	0.32	0.71	-0.43	0.81	n.a.	n.a.	-0.25	0.50
Cohort 1, Follow Up										
Benchmark	0.04	1.91	0.06	1.76	0.10	1.97	0.04	1.96	0.06	1.42
HLM	1.33	1.88	1.94	1.82	3.28	1.89	1.39	1.83	2.29	1.29

SOURCE: Reading comprehension tests administered by study team.

NOTE: Impacts and standard errors are reported using the benchmark approach, which uses generalized estimating equations, and HLM. The social studies and science reading comprehension assessments were developed by ETS. Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include baseline GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher gender, teacher age, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; n.a. = not applicable; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE H.3

DIFFERENCES IN SPRING TEST SCORES BETWEEN TREATMENT AND CONTROL GROUPS, FOR STUDENTS WITH PRETEST AND POST-TEST OR FOLLOW-UP SCORES

	Difference Between Each of the Following and the Control Group:				
	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Composite Test Score[a]					
Cohort 2, Post-Test, School Experience	-0.02	0.04	-0.01	n.a.	0.02
Number of Students	**2,118**	**2,041**	**1,575**	**n.a.**	**5,734**
Cohort 2, Post-Test, Teacher Experience	0.03	0.08	0.05	n.a.	0.06
Number of Students	**1,157**	**1,089**	**623**	**n.a.**	**2,869**
Cohort 1, Follow Up	0.01	0.01	0.06	0.06	0.03
Number of Students	**994**	**935**	**855**	**868**	**3,652**
GRADE Score					
Cohort 2, Post-Test, School Experience	-0.61	-0.27	-0.42	n.a.	-0.38
Number of Students	**2,112**	**2,035**	**1,564**	**n.a.**	**5,711**
Cohort 2, Post-Test, Teacher Experience	0.04	0.18	0.12	n.a.	0.12
Number of Students	**1,020**	**976**	**541**	**n.a.**	**2,537**
Cohort 1, Follow Up	-0.61	-0.06	0.41	0.44	0.09
Number of Students	**991**	**933**	**852**	**862**	**3,638**
Social Studies Reading Comprehension Assessment Score					
Cohort 2, Post-Test, School Experience	-0.36	4.43	0.29	n.a.	1.95
Number of Students[b]	**1,057**	**990**	**778**	**n.a.**	**2,825**
Cohort 2, Post-Test, Teacher Experience	0.43	6.49*	3.34	n.a.	3.29
Number of Students[b]	**512**	**474**	**272**	**n.a.**	**1,258**
Cohort 1, Follow Up	2.13	-0.34	1.88	3.20	1.47
Number of Students[b]	**484**	**452**	**427**	**424**	**1,787**
Science Reading Comprehension Assessment Score					
Cohort 2, Post-Test, School Experience	0.63	1.47	-0.33	n.a.	0.84
Number of Students[b]	**1,029**	**1,034**	**781**	**n.a.**	**2,844**
Cohort 2, Post-Test, Teacher Experience	3.17	2.69	2.11	n.a.	2.67
Number of Students[b]	**499**	**498**	**270**	**n.a.**	**1,267**
Cohort 1, Follow Up	1.39	2.03	3.38	1.83	2.26
Number of Students[b]	**488**	**468**	**418**	**427**	**1,801**

SOURCE: Reading comprehension tests administered by study team.

NOTE: Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher gender, teacher age, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]These sample sizes are smaller than for the other tests because students were randomly assigned to take either the Social Studies or the Science Reading Comprehension Assessment, and no student took both.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

GRADE = Group Reading Assessment and Diagnostic Evaluation; n.a. = not applicable; TOSCRF = Test of Silent Contextual Reading Fluency.

Interacting Treatment Status with Continuous Measures of Prior Achievement

The use of continuous subgroup indicators changed one of the two achievement subgroup findings. Our benchmark subgroup analyses compared impacts for students with above-median prior achievement to impacts for students with below-median prior achievement. (As described in Chapter III, we also estimated several other variations based on different cutoffs to form the subgroups.) As an additional sensitivity test, we also estimated a model in which a continuous measure of prior achievement was interacted with treatment indicator variables.

We find no statistically significant interactions in these analyses (not shown in table). The one finding that differs from what was presented in Chapter III is for subgroups formed by students' baseline comprehension levels. In the benchmark models discussed in Chapter III, we found a statistically significantly greater impact for CRISS students with comprehension levels in the top third of the sample relative to those who scored in the middle third. In these sensitivity tests, none of the interactions between the treatment indicator and baseline GRADE scores were statistically significant.

Impacts for Novice Teachers

Teacher experience subgroup results were not sensitive to the subgroup cutoff used. We assessed the sensitivity of impacts to the way in which we defined the teacher experience subgroups. In one approach, we used 10 years of experience (the study's median) as the cutoff. In the other, we compared the effects of the interventions on test scores for students taught by teachers with less than five years of experience and students taught by teachers with five or more years of experience. In both sets of analyses (for both the fifth- and sixth-grade components of the second year of the study), we found no statistically significant differences in subgroup impacts (see Appendix L tables).

Sensitivity of Teacher Practice Scales

We assessed the sensitivity of the benchmark approach to the way in which we constructed the teacher practice scales. As noted in Chapter II, the benchmark approach to forming teacher practice scales used *averages* of behavior tallies across classroom observation intervals for each teacher and item. As a sensitivity test, we also constructed the scales using the same items for each of the scales, but using *sums* of behavior tallies across intervals. Findings based on sums (shown in Table H.4) were similar to those based on averages (shown in Table II.19), except the statistically significant, negative effect observed for Project CRISS on the Traditional Interaction scale based on averages was no longer statistically significant when the scale was based on sums.

As an additional sensitivity test, we considered a different set of teacher instructional practices scales. These scales were constructed by grouping all items pertaining to teaching comprehension to create a Teaching Comprehension scale, and all items regarding teaching vocabulary to create a Teaching Vocabulary scale. These scales were also created in two ways: using sums and using averages of tallies from the classroom observations. There were no statistically significant differences between treatment and control group teachers' scores on any of these scales (Table H.5).

TABLE H.4

DIFFERENCE IN SPRING CLASSROOM PRACTICES BETWEEN TREATMENT AND CONTROL GROUP COHORT 2 TEACHERS, FOR SCALES BASED ON SUMS OF TALLIES ACROSS OBSERVATION INTERVALS

		Difference Between Each of the Following and the Control Group:			
	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Traditional Interaction Scale	502.00	-2.94	-1.70	3.00	-1.11
Reading Strategy Guidance Scale	499.49	1.51	0.97	0.66	1.08
Classroom Management Scale	502.70	30.46	45.12	53.57	40.89
Number of Teachers in Cohort 2[a]	**54**	**53**	**46**	**31**	**130**

SOURCE: Classroom observations.

NOTE: The scales presented in this table were constructed to capture the frequency of the behaviors in each instructional practice domain shown above, using sums of tallies across observation intervals for each teacher and item. For each scale, the numbers reported in the column labeled "Control Group Mean" are the average predicted outcomes for all students as if they were in the control group. Regression-adjusted impacts were calculated taking into account the clustering of teachers within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher gender, teacher age, teacher race, and district indicators. Smaller scale values represent lower levels of behaviors in the instructional practice domain, while larger values represent higher values of the behaviors.

[a]The number of teachers shown in this row is the number of teachers participating in the study.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

TABLE H.5

DIFFERENCES IN SPRING CLASSROOM PRACTICES BETWEEN TREATMENT AND CONTROL GROUP COHORT 2 TEACHERS, FOR TEACHING COMPREHENSION AND TEACHING VOCABULARY SCALES

		Difference Between Each of the Following and the Control Group:			
	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Teaching Comprehension Scale, Based on Averages of Tallies	500.43	-1.13	-1.55	0.78	-0.81
Teaching Comprehension Scale, Based on Sums of Tallies	500.59	-0.55	-0.11	2.21	0.31
Teaching Vocabulary Scale, Based on Averages of Tallies	500.72	-3.81	-2.07	4.42	-1.13
Teaching Vocabulary Scale, Based on Sums of Tallies	501.93	-3.27	-0.22	4.95	-0.11
Number of Teachers in Cohort 2[a]	**54**	**53**	**46**	**31**	**130**

SOURCE: Classroom observations.

NOTE: The scales presented in this table were constructed to capture the frequency of the behaviors in each instructional practice domain shown above. For each scale, the numbers reported in the column labeled "Control Group Mean" are the average predicted outcomes for all students as if they were in the control group. Regression-adjusted impacts were calculated taking into account the clustering of teachers within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher gender, teacher age, teacher race, and district indicators. Smaller scale values represent lower levels of behaviors in the instructional practice domain, while larger values represent higher values of the behaviors.

[a]The number of teachers shown in this row is the number of teachers participating in the study.

GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

Year 1 Impacts for Schools That Remained in Study for Both Years

One additional sensitivity test conducted relates to the impacts estimated in the first year of the study. In the first year, all 89 schools that agreed to participate in the study were included in the impact estimates. In the second year, 61 of those 89 schools agreed to continue participating in the study. To examine whether the nature of the Year 1 impacts might have differed if only the 61 schools participating in the second year had participated in the first year, we estimated impacts for (1) students in schools that participated in the study for two years and (2) students in schools that participated in only the first year of the study. In these analyses, we found that had we estimated impacts in the first year using only the 61 schools that continued participating in the second year, we would have observed one additional statistically significant negative impact in the first year of the study. In particular, those analyses showed a statistically significant negative impact of Read for Real on post-test scores from the first year of the study (effect size: -0.16) (not shown in table).

This page is intentionally left blank.

APPENDIX I

KEY DESCRIPTIVE STATISTICS FOR CLASSROOM OBSERVATION AND FIDELITY DATA

This page is intentionally left blank.

TABLE I.1

DESCRIPTIVE STATISTICS FOR EXPOSITORY READING COMPREHENSION CLASSROOM OBSERVATION INSTRUMENT ITEMS, BASED ON THE AVERAGE NUMBER OF TIMES EACH PRACTICE WAS OBSERVED DURING THE 10-MINUTE OBSERVATION INTERVALS

	Year 1				Year 2			
	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies
Part I, Comprehension								
Activates prior knowledge and/or previews text before reading								
Teacher models	0.01	0.04	.949	.925	0.00	0.02	n.a.[b]	n.a.[b]
Teacher explains, reviews, provides examples and elaborations	0.61	0.72	.937	.896	0.71	1.02	.991	.986
Students practice	1.07	1.24	.982	.963	1.06	1.35	.995	.992
Explicit comprehension instruction that teaches students about text structure								
Teacher models	0.00	0.03	1.00[a]	n.a.[b]	0.00	0.04	.996	1.00
Teacher explains, reviews, provides examples and elaborations	0.24	0.54	.974	.964	0.32	0.67	.981	.970
Students practice	0.34	0.78	.978	.967	0.50	1.02	.990	.979
Explicit comprehension instruction that teaches students how to use comprehension strategies								
Teacher models	0.01	0.04	.021	.973	0.01	0.11	.692	.532
Teacher explains, reviews, provides examples and elaborations	1.22	1.59	.978	.970	1.03	1.28	.973	.961
Students practice	1.75	2.09	.981	.974	1.78	1.99	.978	.974
Explicit comprehension instruction that teaches students how to generate questions								
Teacher models	0.00	0.04	.798	1.00	0.00	0.03	1.00	1.00
Teacher explains, reviews, provides examples and elaborations	0.24	0.41	.790	.677	0.27	0.40	.846	.950
Students practice	0.43	0.62	.916	.893	0.47	0.71	.975	.964
Explicit comprehension instruction that teaches text features to interpret text								
Teacher models	0.00	0.02	.778	1.00	0.00	0.02	n.a.[b]	n.a.[b]
Teacher explains, reviews, provides examples and elaborations	0.19	0.33	.943	.914	0.17	0.37	.919	.886
Students practice	0.24	0.46	.870	.806	0.22	0.45	.987	.983
Teacher asks students to justify their responses	0.24	0.38	.656	.504	0.27	0.41	.975	.969

Table I.1 *(continued)*

	Year 1				Year 2			
	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies
Teacher asks questions based on material in the text that are beyond the literal level	0.96	1.19	.941	.922	0.90	1.14	.967	.960
Teacher elaborates, clarifies, or links concepts during and after text reading	1.29	1.34	.941	.929	1.17	1.28	.986	.984
Part I, Vocabulary								
Teacher provides an explanation and/or a definition or asks a student to read a definition	0.71	0.72	.905	.879	0.54	0.62	.955	.940
Teacher provides examples, contrasting examples, multiple meanings, immediate elaborations to students' responses	0.87	0.99	.971	.961	0.80	0.91	.960	.952
Teacher uses visuals/pictures, gestures related to word meaning, facial expressions, or demonstrations to discuss/demonstrate word meanings	0.23	0.54	.922	.881	0.21	0.46	.986	.992
Teacher teaches word-learning strategies using context clues, word parts, root meaning	0.09	0.21	.970	.969	0.09	0.21	.888	.858
Students do or are asked to do something that requires knowledge of words	1.39	1.46	.967	.963	1.47	1.42	.982	.980
Students are given an opportunity to apply word-learning strategies using context clues, word parts, and root meaning	0.12	0.52	.938	.918	0.10	0.27	.969	.946
Part I, Grouping Arrangements and Text Reading								
Teacher is working with:								
Whole class (≥75% of class)	0.82	0.26	.924	n.a.	0.85	0.22	.972	n.a.
Large group (> 6 students, < 75% of class)	0.02	0.12	.962	n.a.	0.02	0.10	.979	n.a.
Small groups (3-6 students)	0.21	0.29	.919	n.a.	0.16	0.24	.960	n.a.
Pairs	0.09	0.19	.852	n.a.	0.08	0.15	.911	n.a.
An individual	0.04	0.10	.924	n.a.	0.05	0.15	.980	n.a.
No direct student contact	0.01	0.06	.528	n.a.	0.01	0.06	1.00	n.a.
Text Reading (applies to reading connected text)								
Supported oral reading (includes choral and round-robin reading)	0.39	0.36	.908	n.a.	0.46	0.37	.976	n.a.
Independent silent reading	0.25	0.32	.956	n.a.	0.22	0.28	.979	n.a.
Independent or buddy oral reading	0.32	0.35	.929	n.a.	0.21	0.30	.989	n.a.
Teacher reads aloud	0.17	0.27	.737	n.a.	0.12	0.24	.563	n.a.

Table I.1 *(continued)*

	Year 1				Year 2			
	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies
Teacher reads aloud with students following along silently	0.16	0.26	.865	n.a.	0.24	0.31	.850	n.a.
Text not present	0.05	0.15	.814	n.a.	0.07	0.15	1.00	n.a.
Text present but not being read	0.23	0.25	.788	n.a.	0.25	0.24	.976	n.a.
Part II, Overall Effectiveness of Instruction								
Gave inaccurate and/or confusing explanations or feedback	0.03	0.13	.334	n.a.	0.08	0.21	.926	n.a.
Missed opportunity to correct or address error	0.05	0.20	1.00	n.a.	0.09	0.24	.916	n.a.
Provided opportunities for most students to participate actively during teacher-led instruction	0.87	0.30	.844	n.a.	0.82	0.34	1.00	n.a.
Paced instruction so that the length of the comprehension or vocabulary activities was appropriate for this age group	0.88	0.28	.813	n.a.	0.82	0.34	.760	n.a.
Taught using outlining and/or note taking	0.32	0.41	.797	n.a.	0.26	0.39	.904	n.a.
Used graphic organizers	0.33	0.43	.888	n.a.	0.29	0.38	1.00	n.a.
Kept students thinking for two or more seconds before calling on a student to respond to a complex question	0.62	0.45	.711	n.a.	0.49	0.46	.819	n.a.
Gave independent/pairs/small-group practice in answering comprehension questions or applying comprehension strategy(ies) with expected written product	0.56	0.45	.769	n.a.	0.47	0.44	.926	n.a.
Used writing activities in response to reading (does not include fill-in-the-blank or one-word answers)	0.39	0.45	.874	n.a.	0.34	0.41	.827	n.a.
Part II, Overall Management/Responsiveness to Students								
Teacher maximized the amount of time available for instruction	3.25	0.82	.861	n.a.	3.26	0.78	.916	n.a.
Teacher managed student behavior effectively in order to avoid disruptions and provide productive learning environments	3.40	0.74	.863	n.a.	3.39	0.80	.932	n.a.
Teacher redirected discussion if a student response was leading the group off topic/focus	3.30	0.73	.602	n.a.	3.12	0.92	.847	n.a.

Table I.1 *(continued)*

	Year 1				Year 2			
	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies	Mean	Standard Deviation	Reliability, All Observation Pairs	Reliability, Excluding Observation Pairs with Zero Tallies
Part II, Overall Student Engagement During Observation								
Student engagement during the first half of the observation session	2.64	0.55	.842	n.a.	2.72	0.51	.895	n.a.
Student engagement during the remainder of the observation session	2.58	0.59	.873	n.a.	2.61	0.58	.895	n.a.

SOURCE: Classroom observations.

NOTE: Reliability was calculated using Pearson correlation coefficients. In the Year 1 and Year 2 table panes above, the first reliability column includes all nonmissing paired observations, while the second column removes from the calculations observer pairs that reported zero tallies on that specific item (note that the second reliability column is relevant only for the vocabulary and comprehension sections of Part I where observers recorded tallies of the number of times teachers engaged in each behavior, so n.a. [not applicable] is shown for all of the other items). For Part I vocabulary and comprehension items, the means, standard deviations, and reliability estimates shown are for the average of the classroom tallies across all the observed 10-minute intervals (up to 10 intervals per teacher).

[a]This reliability estimate of 1.0 seems to be inconsistent with the reported standard deviation, which is greater than zero. This occurs because only a *subset* of observations can be used for the reliability estimates, while the full set of observations is used in calculating the means and standard deviations. For this item, all of the observations used for the reliability calculations had zero tallies, which corresponds to a reliability estimate equal to 1.0.

[b]Inter-rater reliability could not be calculated as there were no remaining observer pairs after dropping the pairs with zero tallies.

n.a. = not applicable.

TABLE I.2

DESCRIPTIVE STATISTICS FOR PROJECT CRISS FIDELITY OBSERVATION ITEMS

	Year 1		Year 2	
	Percentage	Standard Deviation	Percentage	Standard Deviation
Teachers Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Provide instruction or lead activities to generate background knowledge about a topic or concept before students read about it	67.31	46.91	67.35	46.89
Help students set goals and determine a purpose before beginning to read	63.46	46.87	61.22	48.72
Have students read a written text	84.62	36.08	91.84	27.38
Lead students during and/or after reading in transforming information activities (for example, graphic organizer, guided discussion)	82.69	37.83	85.71	34.99
Include informal or formal writing in the transforming activities (including note taking)	76.92	40.34	79.59	40.30
Use the transforming activities to teach the content of the lesson	76.92	40.34	75.51	43.00
Discuss or reflect on students' metacognitive processes during the transforming activities	46.15	48.95	42.86	49.49
Lead the whole class in a reflection discussion at the end of the lesson using questions such as: A. Metacognition: How did you evaluate your comprehension? B. Background knowledge: Did I assist you in thinking about what you already knew? C. Purpose setting: Did you have clear purposes? D. Active involvement: How were you actively engaged? E. Discussion: How did discussion clarify your thinking? F. Writing: How did you use writing to help you learn? G. Transformation: What were the different ways you transformed information? How did this help you? H. Teacher modeling: Did I do enough modeling?	—[b]	—[b]	16.33	36.96
Number of Teachers[c]	**54**		**53**	

SOURCE: Classroom observations.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers are included in these calculations. The percentage of teachers who reported using Project CRISS was 94.23 percent in Year 1 and 94.09 percent in Year 2. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]Value suppressed to protect teacher confidentiality.

[c]The number of teachers presented in this row is the number participating in the study.

TABLE I.3

DESCRIPTIVE STATISTICS FOR READ FOR REAL FIDELITY OBSERVATION ITEMS

	Learn Observation Days				Practice Observation Days			
	Year 1		Year 2		Year 1		Year 2	
	Percentage	Standard Deviation	Percentage	Standard Deviation	Percentage	Standard Deviation	Percentage	Standard Deviation
Teachers Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]								
Before Reading								
Reads or asks a student to read the explanation of the Before Reading focus strategy	55.00	49.75	68.75	46.35	54.55	49.79	35.71	47.92
Discusses the strategy with students	45.00	49.75	75.00	43.30	54.55	49.79	35.71	47.92
Reads or asks a student to read the information in the My Thinking box	55.00	49.75	50.00	50.00	n.a.	n.a.	n.a.	n.a.
Asks students to apply the strategy	45.00	49.75	43.75	49.61	57.58	49.42	42.86	49.49
Discusses students' comments	n.a.	n.a.	n.a.	n.a.	48.48	49.98	42.86	49.49
During Reading								
Reads or asks a student to read the explanation of the During Reading focus strategy	60.00	48.99	81.25	39.03	48.48	49.98	50.00	50.00
Discusses the strategy with the students	65.00	47.70	62.50	48.41	n.a.	n.a.	n.a.	n.a.
Reads or asks a student to read the information in the My Thinking box (notes from the reading partner)	60.00	48.99	68.75	46.35	42.42	49.42	42.86	49.49
Asks students to share their thinking about the strategy	60.00	48.99	50.00	50.00	n.a.	n.a.	n.a.	n.a.
Reminds students to write notes about the strategy	n.a.	n.a.	n.a.	n.a.	36.36	48.10	57.14	49.49
Stops and addresses the My Thinking notes at the "red strategy buttons"	65.00	47.70	62.50	48.41	69.70	45.96	50.00	50.00

Table I.3 *(continued)*

	Learn Observation Days				Practice Observation Days			
	Year 1		Year 2		Year 1		Year 2	
	Percentage	Standard Deviation	Percentage	Standard Deviation	Percentage	Standard Deviation	Percentage	Standard Deviation
After Reading[b]								
Reads and/or asks students to read the selection	70.00	45.83	56.25	49.61	69.70	45.96	64.29	47.92
Reads or asks a student to read the After Reading focus strategy	35.00	47.70	31.25	46.35	24.24	42.85	28.57	45.18
Discusses or asks questions about the strategy	25.00	43.30	31.25	46.35	21.21	40.88	21.43	41.03
Reads or asks a student to read the information in the My Thinking box	20.00	40.00	31.25	46.35	n.a.	n.a.	n.a.	n.a.
Gives a written assignment highlighting the After Reading focus strategy	n.a.	n.a.	n.a.	n.a.	15.15	35.86	21.43	41.03
Calls on students to implement the After Reading focus strategy	15.00	35.71	31.25	46.35	n.a.	n.a.	n.a.	n.a.
Comprehension								
Administers the open book comprehension test	—[c]	—[c]	—[c]	—[c]	9.09	28.75	—[c]	—[c]
Corrects tests with the class	—[c]	—[c]	—[c]	—[c]	—[c]	—[c]	0.00	0.00
Discusses responses	—[c]	—[c]	—[c]	—[c]	—[c]	—[c]	0.00	0.00
Organizing Information								
Reads or asks a student to read the information from the reading partner	20.00	40.00	18.75	39.03	n.a.	n.a.	n.a.	n.a.
Discusses the graphic organizer	30.00	45.83	—[c]	—[c]	n.a.	n.a.	n.a.	n.a.
Asks students to complete the graphic organizer	n.a.	n.a.	n.a.	n.a.	12.12	32.64	28.57	45.18

Table I.3 *(continued)*

	Learn Observation Days				Practice Observation Days			
	Year 1		Year 2		Year 1		Year 2	
	Percentage	Standard Deviation	Percentage	Standard Deviation	Percentage	Standard Deviation	Percentage	Standard Deviation
Writing for Comprehension								
Reads or asks a student to read the information from the reading partner	15.00	35.71	—[c]	—[c]	n.a.	n.a.	n.a.	n.a.
Reads or asks a student to read the summary	20.00	40.00	—[c]	—[c]	n.a.	n.a.	n.a.	n.a.
Asks students to write a summary based on their completed graphic organizer	n.a.	n.a.	n.a.	n.a.	—[c]	—[c]	—[c]	—[c]
Identifies how the paragraphs and sentences in the summary correspond to the information on the graphic organizer	15.00	35.71	—[c]	—[c]	n.a.	n.a.	n.a.	n.a.
Discusses the Three Parts of a Summary								
Introduction	20.00	40.00	—[c]	—[c]	n.a.	n.a.	n.a.	n.a.
Body	20.00	40.00	—[c]	—[c]	n.a.	n.a.	n.a.	n.a.
Conclusion	20.00	40.00	—[c]	—[c]	n.a.	n.a.	n.a.	n.a.
Sample Size:[d] First Year = 57; Second Year = 31								

SOURCE: Classroom observations.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers are included in these calculations. The percentage of teachers who reported using Read for Real was 86.79 percent in Year 1 and 83.33 percent in Year 2. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]The vocabulary and fluency items are not included in the table because developers noted they were not essential for implementation of the Read for Real intervention.

[c]Value suppressed to protect teacher confidentiality.

[d]The number of teachers presented in this row is the number participating in the study.

n.a. = not applicable.

TABLE I.4

DESCRIPTIVE STATISTICS FOR READABOUT FIDELITY OBSERVATION ITEMS

	Year 1		Year 2	
	Percentage	Standard Deviation	Percentage	Standard Deviation
Teachers Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]				
Used the ReadAbout materials	91.30	28.18	95.56	20.61
Computer workstation used	89.13	31.13	68.89	46.29
Independent workstation used	58.70	49.24	55.56	49.69
Provided direction instruction (explain and/or model) on the comprehension or vocabulary strategy or skill	76.09	42.66	77.78	41.57
Provided opportunities for students to apply the comprehension or vocabulary skill (guided practice)	80.43	39.67	80.00	40.00
Provided students instruction on the selected 6+1 Writing Trait	0.00	0.00	—[b]	—[b]
Provided opportunities to apply the 6+1 Writing Trait Model	0.00	0.00	—[b]	—[b]
Sample Size[c]	**53**		**46**	

SOURCE: Classroom observations.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers are included in these calculations. The percentage of teachers who reported using ReadAbout was 100 percent in Year 1 and 95.71 percent in Year 2. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]Value suppressed to protect teacher confidentiality.

[c]The number of teachers presented in this row is the number participating in the study.

TABLE I.5

DESCRIPTIVE STATISTICS FOR FIDELITY OBSERVATION ITEMS FOR READING FOR KNOWLEDGE DIRECT INSTRUCTION OBSERVATION DAYS

	Year 1	
	Percentage	Standard Deviation
Teachers Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]		
Post the reading goal	38.09	50.32
Present the reading goal	57.14	50.32
Present the cooperative learning goal	38.09	50.32
Ask students to review vocabulary or provide practice and instruction (Exception: This is not done on the first day of a new unit.)	—[b]	—[b]
Build background knowledge about the topic of text or about a skill/strategy	66.67	49.24
Explain a skill/strategy or remind students of a skill/strategy, recently learned	71.42	47.67
Read the text aloud and (1) think aloud or model a skill/strategy, or (2) ask the students to apply a skill/strategy	52.38	51.18
Follow the recommended pacing for the lesson	57.14	50.96
Award cooperation and/or improvement points during the lesson	52.38	51.18
Sample Size[c]	**54**	

SOURCE: Classroom observations.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers are included in these calculations. The percentage of teachers who reported using Reading for Knowledge in Year 1 is 83.33 percent. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[b]Value suppressed to protect teacher confidentiality.

[c]The number of teachers presented in this row is the number participating in the study.

TABLE I.6

DESCRIPTIVE STATISTICS FOR FIDELITY OBSERVATION ITEMS FOR READING FOR KNOWLEDGE COOPERATIVE GROUPS OBSERVATION DAYS

	Year 1	
	Percentage	Standard Deviation
Teachers Observed to Have Done the Following During the Time When Their Classes Were Observed:[a]		
Post the reading goal	60.61	49.90
Present the reading goal	87.88	33.60
Present the cooperative learning goal	66.67	48.26
Ask students to review vocabulary or provide practice and instruction (Exception: This is not done on the first day of a new unit.)	54.55	50.40
Use a whole-group or partner activity to discuss key points about the day's skill/strategy	81.82	39.66
Provide feedback and prompts to partner pairs during partner reading	81.82	39.66
Chart individual students' progress on the setting goals and charting progress forms during partner reading	27.27	45.68
Review routines for Team Talk discussion	51.52	50.70
Read Team Talk questions aloud	60.61	49.90
Circulate within the classroom and monitor team discussions and provide prompts	78.79	42.00
Ask team members to share with the class their response and reasoning to Team Talk questions	75.76	43.99
Follow the recommended pacing for the lesson	54.55	50.40
Award cooperation and/or improvement points during the lesson	60.61	49.19
Sample Size[c]	**54**	

SOURCE: Classroom observations.

[a]Fidelity observations were conducted only for teachers implementing the assigned curricula; however, all teachers are included in these calculations. The percentage of teachers who reported using Reading for Knowledge in Year 1 is 83.33 percent. We assumed that teachers who were not implementing the curricula did not engage in the activities listed in this table.

[c]The number of teachers presented in this row is the number participating in the study.

This page is intentionally left blank.

APPENDIX J

UNADJUSTED MEANS

This page is intentionally left blank.

TABLE J.1

UNADJUSTED MEANS FOR TREATMENT AND CONTROL GROUPS

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Cohort 1						
Pretest (Fall 2006) Test Scores						
TOSCRF	88.24	89.08	87.84	87.84	89.7	88.62
GRADE	99.83	100.86	99.58	99.25	101.13	100.21
Post-Test (Spring 2007) Test Scores						
Composite	0.02	0.06	-0.04	-0.07	0.02	-0.01
GRADE	100.80	100.70	99.83	100.09	101.32	100.75
Social Studies Reading Comprehension Assessment	501.79	501.15	499.81	497.37	501.05	499.87
Science Reading Comprehension Assessment	501.51	502.53	499.99	498.17	499.39	500.06
Follow Up (Spring 2008) Test Scores						
Composite[a]	-0.03	0.05	-0.07	-0.05	0.08	0.00
GRADE	96.44	97.35	95.69	95.75	97.69	96.64
Social Studies Reading Comprehension Assessment	500.08	501.78	497.78	498.04	503.47	500.30
Science Reading Comprehension Assessment	497.66	502.03	498.49	500.04	501.84	500.62
Number of Students[b]	**1,362**	**1,319**	**1,245**	**1,228**	**1,195**	**4,987**
Cohort 2						
Pretest (Fall 2007) Test Scores						
TOSCRF	89.26	89.56	88.45	89.52	n.a.	89.13
GRADE	100.54	101.46	100.24	100.23	n.a.	100.73
Post-Test (Spring 2008) Test Scores						
Composite[a]	0.00	0.01	-0.01	-0.02	n.a.	0.00
GRADE	101.34	101.48	100.42	100.91	n.a.	100.96
Social Studies Reading Comprehension Assessment	500.81	502.09	503.48	500.57	n.a.	502.28
Science Reading Comprehension Assessment	503.10	502.87	502.94	502.28	n.a.	502.77
Number of Students[c]	**1,194**	**1,201**	**1,108**	**639**	**n.a.**	**2,948**

SOURCE: Reading comprehension tests administered by study team.

Table J.1 *(continued)*

NOTE: The social studies and science reading comprehension assessments were developed by ETS.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]The number of students presented in this row is the number of Cohort 1 students participating in the study.

[c]The number of students presented in this row is the number of Cohort 2 students participating in the study.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; n.a. = not applicable; TOSCRF = Test of Silent Contextual Reading Fluency.

APPENDIX K

IMPACT TABLES INCLUDING P-VALUES THAT HAVE NOT BEEN ADJUSTED FOR MULTIPLE COMPARISONS

This page is intentionally left blank.

TABLE K.1

DIFFERENCES IN POST-TEST SCORES BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING FIFTH GRADE COHORTS 1 AND 2 WITHOUT ADJUSTMENTS FOR MULTIPLE COMPARISONS

	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Composite Test Score[a]					
Cohort 1 Students (Spring 2007)					
Impact	-0.01	-0.01	-0.04	-0.07	-0.04
Effect Size		-0.01	-0.04	-0.08	-0.05
p-value		0.87	0.40	0.09	0.16
Cohort 2 Students (Spring 2008)					
Impact	-0.04	0.00	0.05	-0.02	0.02
Effect Size		0.00	0.06	-0.02	0.02
p-value		0.99	0.28	0.76	0.61
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		0.01	0.09	0.05	0.06
Difference in Effect Size		0.01	0.10	0.06	0.08
p-value for the Difference		0.90	0.18	0.35	0.21
GRADE Score					
Cohort 1 Students (Spring 2007)					
Impact	100.55	-0.19	-0.64	-0.76	-0.60
Effect Size		-0.01	-0.05	-0.06	-0.04
p-value		0.77	0.32	0.19	0.19
Cohort 2 Students (Spring 2008)					
Impact	100.76	-0.28	-0.08	-0.56	-0.26
Effect Size		-0.02	-0.01	-0.04	-0.02
p-value		0.68	0.90	0.42	0.60
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		-0.09	0.56	0.20	0.34
Difference in Effect Size		-0.01	0.04	0.01	0.02
p-value for the Difference		0.92	0.54	0.81	0.64
Social Studies Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	500.30	-1.36	-0.38	-2.28	-1.18
Effect Size		-0.05	-0.01	-0.08	-0.04
p-value		0.51	0.78	0.11	0.28
Cohort 2 Students (Spring 2008)					
Impact	499.83	0.09	4.63*	0.47	2.21
Effect Size		0.00	0.16	0.02	0.07
p-value		0.96	0.00	0.82	0.12
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		1.45	5.01*	2.75	3.39
Difference in Effect Size		0.05	0.17	0.09	0.11
p-value for the Difference		0.57	0.02	0.26	0.06

Table K.1 (*continued*)

	Control Group	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Science Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	500.60	0.31	-1.07	-2.71	-1.38
Effect Size		0.01	-0.04	-0.10	-0.05
p-value		0.84	0.54	0.15	0.31
Cohort 2 Students (Spring 2008)					
Impact	501.59	0.58	1.66	-0.31	0.83
Effect Size		0.02	0.06	-0.01	0.03
p-value		0.78	0.48	0.89	0.63
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		0.27	2.73	2.41	2.21
Difference in Effect Size		0.01	0.10	0.09	0.08
p-value for the Difference		0.91	0.30	0.42	0.31
Number of Students in Cohort 1[b]	**1,368**	**1,316**	**1,248**	**1,227**	**3,791**
Number of Students in Cohort 2[c]	**1,196**	**1,202**	**1,111**	**634**	**2,947**

SOURCE: Reading comprehension tests administered by study team.

NOTE: For each outcome, the numbers reported in the column labeled "Control Group" are the average predicted outcomes for all students as if they were in the control group. The numbers reported in the remaining columns are, by row: (1) the impact, (2) the effect size, and (3) the *p-value* of the impact. For each outcome, the differences between cohort impacts are also reported. The social studies and science reading comprehension assessments were developed by ETS. Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]The number of students presented in this row is the number of Cohort 1 students participating in the study. The proportion of students in each experimental condition with post-test scores is reported in Appendix G.

[c]The number of students presented in this row is the number of Cohort 2 students participating in the study. The proportion of students in each experimental condition with post-test scores is reported in Appendix G.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE K.2

DIFFERENCES IN POST-TEST SCORES BETWEEN TREATMENT AND CONTROL GROUPS, COMPARING FIFTH-GRADE COHORT 1 AND 2 STUDENTS WITH TEACHERS IN THE STUDY FOR TWO CONSECUTIVE YEARS WITHOUT ADJUSTMENTS FOR MULTIPLE COMPARISONS

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Composite Test Score[a]					
Cohort 1 Students (Spring 2007)					
Impact	0.06	-0.06	-0.06	-0.09	-0.08
Effect Size		-0.07	-0.08	-0.10	-0.09
p-value		0.34	0.22	0.09	0.08
Cohort 2 Students (Spring 2008)					
Impact	-0.04	0.02	0.09	0.03	0.05
Effect Size		0.03	0.10	0.03	0.06
p-value		0.66	0.10	0.61	0.23
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		0.08	0.15*	0.11	0.13*
Difference in Effect Size		0.10	0.18	0.14	0.15
p-value for the Difference		0.31	0.02	0.08	0.04
GRADE Score					
Cohort 1 Students (Spring 2007)					
Impact	101.4	-1.07	-0.94	-1.48	-1.14
Effect Size		-0.08	-0.07	-0.11	-0.08
p-value		0.18	0.21	0.07	0.05
Cohort 2 Students (Spring 2008)					
Impact	100.1	0.16	0.24	-0.21	0.08
Effect Size		0.01	0.02	-0.02	0.01
p-value		0.84	0.75	0.80	0.89
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		1.23	1.17	1.27	1.23
Difference in Effect Size		0.09	0.09	0.09	0.09
p-value for the Difference		0.26	0.20	0.15	0.13
Social Studies Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	502.6	-3.53	-1.78	-1.56	-2.09
Effect Size		-0.12	-0.06	-0.05	-0.07
p-value		0.18	0.25	0.45	0.16
Cohort 2 Students (Spring 2008)					
Impact	500.0	0.27	6.43*	3.03	3.25
Effect Size		0.01	0.22	0.10	0.11
p-value		0.93	0.99	0.17	0.07
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		3.80	8.21*	4.59	5.34
Difference in Effect Size		0.13	0.28	0.15	0.18
p-value for the Difference		0.31	0.00	0.15	0.03

Table K.2 (*continued*)

	Control Group Mean	Project CRISS	ReadAbout	Read for Real	Combined Treatment Group
Science Reading Comprehension Assessment Score					
Cohort 1 Students (Spring 2007)					
Impact	503.4	-0.07	-2.11	-3.04	-1.76
Effect Size		0.00	-0.08	-0.11	-0.06
p-value		0.97	0.38	0.11	0.32
Cohort 2 Students (Spring 2008)					
Impact	501.7	2.22	2.91	1.87	2.35
Effect Size		0.08	0.10	0.07	0.08
p-value		0.30	0.28	0.40	0.22
Difference Between Cohort 1 and Cohort 2					
Difference in Impact		2.29	5.02	4.91	4.10
Difference in Effect Size		0.08	0.18	0.18	0.15
p-value for the Difference		0.42	0.10	0.09	0.11
Number of Students with Teachers in Study for Two Years[b]					
Cohort 1	**933**	**845**	**902**	**487**	**2,234**
Cohort 2	**949**	**775**	**815**	**478**	**2,068**

SOURCE: Reading comprehension tests administered by study team.

NOTE: For each outcome, the numbers reported in the column labeled "Control Group Mean" are the average predicted outcomes for all students as if they were in the control group. The numbers reported in the remaining columns are, by row: (1) the impact, (2) the effect size, and (3) the *p-value* of the impact. For each outcome, the differences between cohort impacts are also reported. The social studies and science reading comprehension assessments were developed by ETS. Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, whether students were overage for grade, teacher sex, teacher age, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]Counts reflect the number of students with nonmissing teacher data.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE K.3

DIFFERENCES IN POST-TEST AND FOLLOW-UP TEST SCORES BETWEEN TREATMENT AND CONTROL GROUPS, COHORT 1 STUDENTS WITHOUT ADJUSTMENTS FOR MULTIPLE COMPARISONS

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Composite Test Score[a]						
Post-Test (Spring 2007)						
Impact	0.01	-0.01	-0.03	-0.06	-0.11*	-0.07*
Effect Size		-0.02	-0.04	-0.06	-0.13	-0.08
p-value		0.75	0.50	0.22	0.01	0.02
Follow Up (Spring 2008)						
Impact	-0.06	-0.01	0.00	0.06	0.05	0.02
Effect Size		-0.01	0.00	0.07	0.06	0.03
p-value		0.84	0.96	0.14	0.20	0.38
Difference Between Post-Test and Follow Up						
Difference in Impact		0.01	0.03	0.12*	0.17*	0.09*
Difference in Effect Size		0.01	0.04	0.13	0.18	0.10
p-value for the Difference		0.91	0.57	0.04	0.01	0.03
GRADE Score						
Post-Test (Spring 2007)						
Impact	100.96	-0.44	-0.68	-0.74	-1.45*	-1.01*
Effect Size		-0.03	-0.05	-0.05	-0.11	-0.07
p-value		0.48	0.35	0.22	0.02	0.02
Follow Up (Spring 2008)						
Impact	96.04	-0.75	-0.14	0.52	0.31	-0.04
Effect Size		-0.05	-0.01	0.04	0.02	0.00
p-value		0.19	0.82	0.44	0.55	0.92
Difference Between Post-Test and Follow Up						
Difference in Impact		-0.31	0.54	1.25	1.76	0.97
Difference in Effect Size		-0.02	0.04	0.09	0.13	0.07
p-value for the Difference		0.67	0.57	0.11	0.02	0.11
Social Studies Reading Comprehension Assessment Score						
Post-Test (Spring 2007)						
Impact	500.40	-0.67	-0.36	-1.38	-1.91	-1.36
Effect Size		-0.02	-0.01	-0.05	-0.06	-0.05
p-value		0.77	0.83	0.41	0.19	0.21
Follow Up (Spring 2008)						
Impact	498.15	1.42	-0.65	1.70	3.22	1.08
Effect Size		0.05	-0.02	0.06	0.11	0.04
p-value		0.44	0.69	0.34	0.15	0.4
Difference Between Post-Test and Follow Up						
Difference in Impact		2.09	-0.29	3.08	5.13	2.44
Difference in Effect Size		0.07	-0.01	0.10	0.17	0.08
p-value for the Difference		0.39	0.89	0.19	0.06	0.16

Table K.3 (*continued*)

	Control Group	Project CRISS	ReadAbout	Read for Real	Reading for Knowledge	Combined Treatment Group
Science Reading Comprehension Assessment Score						
Post-Test (Spring 2007)						
Impact	500.61	0.94	-0.42	-1.14	-5.43*	-1.92
Effect Size		0.03	-0.02	-0.04	-0.2	-0.07
p-value		0.54	0.79	0.61	0.00	0.14
Follow Up (Spring 2008)						
Impact	497.27	1.37	1.92	3.18	1.35	2.23
Effect Size		0.04	0.06	0.10	0.04	0.07
p-value		0.48	0.32	0.10	0.51	0.13
Difference Between Post-Test and Follow Up						
Difference in Impact		0.43	2.33	4.31	6.78*	4.15*
Difference in Effect Size		0.01	0.08	0.14	0.24	0.14
p-value for the Difference		0.87	0.31	0.16	0.03	0.05
Number of Cohort 1 Students in Year 1[b]	**1,368**	**1,316**	**1,248**	**1,227**	**1,191**	**4,982**
Number of Cohort 1 Students in Year 2[c]	**1,008**	**1,048**	**960**	**893**	**901**	**3,802**

SOURCE: Reading comprehension tests administered by study team.

NOTE: For each outcome, the numbers reported in the column labeled "Control Group" are the average predicted outcomes for all students as if they were in the control group. The numbers reported in the remaining columns are, by row: (1) the impact, (2) the effect size, and (3) the *p-value* of the impact. For each outcome, the differences between impacts for the post-test and follow up are also reported. The social studies and science reading comprehension assessments were developed by ETS. Regression-adjusted impacts were calculated taking into account the clustering of students within schools. Variables in this model include pretest GRADE and TOSCRF scores, student ethnicity and race, student English language learner status, school location, teacher race, and district indicators.

[a]The composite is based on the three tests presented in this table. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores.

[b]The number of students presented in this row is the number of Cohort 1 students participating in the study in Year 1. The proportion of students in each experimental condition with post-test scores is reported in Appendix G.

[c]The number of students presented in this row is the number of Cohort 1 students participating in the study in Year 2. The proportion of students in each experimental condition with follow-up test scores is reported in Appendix G.

ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Statistically different at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

APPENDIX L

SUBGROUP IMPACT TABLES

This page is intentionally left blank.

TABLE L.1

DIFFERENCES IN EFFECTS ON THE COMPOSITE POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard error	*p-value*	Estimate	Standard error	*p-value*	Estimate	Standard error	*p-value*	Estimate	Standard error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	0.04	0.09	0.96	0.03	0.08	0.94	-0.05	0.08	0.90	0.02	0.07	0.78
Pretest TOSCRF score, above sample median	-0.08	0.06	0.35	-0.06	0.06	0.57	0.00	0.06	1.00	-0.04	0.05	0.45
Pretest TOSCRF score, top third (vs. bottom)	-0.12	0.08	0.31	-0.04	0.07	0.89	0.03	0.08	0.95	-0.04	0.06	0.49
Pretest TOSCRF score, middle third (vs. bottom)	-0.09	0.06	0.29	-0.03	0.05	0.92	-0.08	0.08	0.62	-0.06	0.05	0.22
Pretest TOSCRF score, top third (vs. middle)	0.00	0.05	1.00	0.01	0.04	0.99	0.12	0.05	0.06	0.03	0.04	0.42
Pretest GRADE score, above national norm	-0.03	0.05	0.92	0.02	0.06	0.99	0.01	0.06	1.00	0.02	0.05	0.71
Pretest GRADE score, above sample median	-0.03	0.05	0.92	0.02	0.06	0.99	0.01	0.06	1.0.0	0.02	0.05	0.71
Pretest GRADE score, top third (vs. bottom)	0.07	0.08	0.72	0.05	0.09	0.91	0.02	0.07	0.99	0.07	0.07	0.34
Pretest GRADE score, middle third (vs. bottom)	-0.05	0.08	0.90	0.02	0.06	0.95	0.03	0.07	0.96	0.01	0.06	0.87
Pretest GRADE score, top third (vs. middle)	0.12*	0.05	0.02	-0.02	0.04	0.93	0.01	0.06	1.00	0.04	0.04	0.21
Classified as ELL	0.14	0.10	0.32	0.09	0.10	0.65	0.46*	0.13	0.00	0.15	0.08	0.07
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	0.03	0.10	0.98	-0.17	0.10	0.27	0.14	0.08	0.20	-0.02	0.07	0.76
More than 5 Years Teaching Experience	-0.16	0.10	0.27	-0.08	0.12	0.79	0.02	0.11	0.99	-0.06	0.08	0.45
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-0.22	0.11	0.11	-0.01	0.11	1.00	-0.22	0.10	0.09	-0.10	0.09	0.27

Table L.1 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard error	*p-value*	Estimate	Standard error	*p-value*	Estimate	Standard error	*p-value*	Estimate	Standard error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-0.13	0.08	0.26	-0.13	0.10	0.49	-0.13	0.08	0.25	-0.15*	0.06	0.03
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	-0.09	0.11	0.76	-0.12	0.13	0.69	-0.03	0.10	0.98	-0.06	0.08	0.48
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	-0.05	0.12	0.95	-0.08	0.08	0.61	0.11	0.12	0.69	-0.01	0.07	0.88
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	0.01	0.09	1.00	0.15	0.08	0.20	0.19	0.11	0.28	0.06	0.07	0.37
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	0.03	0.07	0.95	-0.13	0.08	0.30	0.11	0.08	0.45	-0.01	0.06	0.87
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-0.10	0.08	0.57	0.01	0.07	1.00	-0.03	0.09	0.99	-0.05	0.06	0.35
Above Sample Median Classroom Management Scale Score (499.9)	-0.07	0.07	0.69	-0.01	0.08	1.00	-0.10	0.11	0.73	-0.06	0.06	0.26

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table are adjusted for multiple-hypotheses testing. The composite is based on the GRADE and the social studies and science reading comprehension tests. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores. The social studies and science reading comprehension assessments were developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE L.2

DIFFERENCES IN EFFECTS ON THE GRADE POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	0.37	1.36	1.00	0.69	1.24	1.00	-0.30	1.32	1.00	0.57	1.06	0.92
Pretest TOSCRF score, above sample median	-1.15	0.88	0.77	-0.71	0.87	0.98	-0.40	1.00	1.00	-0.52	0.74	0.85
Pretest TOSCRF score, top third (vs. bottom)	-1.52	1.22	0.82	-0.93	1.17	0.98	-0.68	1.28	1.00	-0.83	0.99	0.77
Pretest TOSCRF score, middle third (vs. bottom)	-1.65	1.11	0.67	-1.50	0.89	0.53	-2.49	1.58	0.60	-1.70	0.84	0.13
Pretest TOSCRF score, top third (vs. middle)	0.08	0.86	1.00	0.55	0.71	0.99	1.70	0.96	0.46	0.64	0.74	0.75
Pretest GRADE score, above national norm	-0.97	0.72	0.75	0.76	0.82	0.95	-1.12	0.84	0.76	0.04	0.68	1.00
Pretest GRADE score, above sample median	-0.97	0.72	0.75	0.76	0.82	0.95	-1.12	0.84	0.76	0.04	0.68	1.00
Pretest GRADE score, top third (vs. bottom)	-0.28	1.01	1.00	0.86	1.09	0.98	-1.22	1.02	0.84	0.31	0.90	0.98
Pretest GRADE score, middle third (vs. bottom)	-1.38	1.10	0.82	0.24	0.91	1.00	-0.60	1.23	1.00	-0.22	0.90	0.99
Pretest GRADE score, top third (vs. middle)	0.82	0.69	0.86	-0.27	0.71	1.00	-0.09	1.06	1.00	0.23	0.59	0.96
Classified as ELL	2.97	1.07	0.06	0.40	1.30	1.00	6.55	2.52	0.09	1.79	1.07	0.25
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	1.33	1.21	0.88	-1.39	1.49	0.94	1.99	1.14	0.47	0.40	0.95	0.95
More than 5 Years Teaching Experience	-0.74	1.37	1.00	-1.03	1.48	0.99	-0.85	1.33	0.99	-0.85	1.09	0.80
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-2.26	1.53	0.63	-0.06	1.77	1.00	-2.71	1.57	0.45	-1.03	1.34	0.78

Table L.2 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-0.81	1.07	0.98	-1.43	1.55	0.95	-0.69	1.03	0.99	-1.26	0.94	0.43
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	-1.09	1.62	0.99	-1.86	2.11	0.95	0.06	1.54	1.00	-0.80	1.41	0.89
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	-0.92	1.47	0.99	-1.45	1.17	0.80	0.54	1.82	1.00	-0.63	0.98	0.86
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	0.92	1.14	0.98	2.57	1.20	0.25	3.51	1.51	0.18	1.58	1.00	0.30
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	1.77	0.89	0.31	-1.56	1.26	0.81	2.37	1.05	0.19	0.74	0.85	0.73
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-1.44	1.14	0.80	0.34	0.96	1.00	0.19	1.19	1.00	-0.71	0.78	0.71
Above Sample Median Classroom Management Scale Score (499.9)	-0.51	1.10	1.00	0.75	1.27	1.00	-0.14	1.27	1.00	-0.20	0.72	0.99

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table are adjusted for multiple-hypotheses testing. The composite is based on the GRADE and the social studies and science reading comprehension tests. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE L.3

DIFFERENCES IN EFFECTS ON THE ETS SOCIAL STUDIES POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	1.45	2.68	1.00	-2.23	3.68	1.00	-0.62	2.91	1.00	-0.20	2.49	1.00
Pretest TOSCRF score, above sample median	-0.22	2.76	1.00	-0.47	2.57	1.00	2.57	2.88	0.96	0.01	2.34	1.00
Pretest TOSCRF score, top third (vs. bottom)	0.86	3.27	1.00	2.70	3.06	0.97	7.60	3.54	0.24	2.69	2.74	0.68
Pretest TOSCRF score, middle third (vs. bottom)	1.24	3.79	1.00	3.95	2.79	0.72	6.08	3.32	0.43	3.48	2.67	0.47
Pretest TOSCRF score, top third (vs. middle)	3.11	3.14	0.94	2.76	2.37	0.87	3.08	2.20	0.72	2.69	2.10	0.47
Pretest GRADE score, above national norm	3.29	3.23	0.92	-0.98	2.92	1.00	4.09	3.19	0.79	1.85	2.57	0.84
Pretest GRADE score, above sample median	3.29	3.23	0.92	-0.98	2.92	1.00	4.09	3.19	0.79	1.85	2.57	0.84
Pretest GRADE score, top third (vs. bottom)	8.64	4.29	0.30	0.40	4.34	1.00	3.82	4.92	0.98	4.46	3.62	0.49
Pretest GRADE score, middle third (vs. bottom)	0.01	4.78	1.00	0.14	3.81	1.00	1.77	4.83	1.00	0.78	3.30	0.99
Pretest GRADE score, top third (vs. middle)	6.67	3.47	0.37	-1.05	2.16	1.00	0.90	3.08	1.00	1.74	2.38	0.82
Classified as ELL	4.07	5.20	0.98	1.45	5.36	1.00	12.87	5.62	0.18	3.59	4.73	0.81
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	1.40	4.79	1.00	-4.79	3.46	0.71	1.82	4.40	1.00	-0.91	3.14	0.98
More than 5 Years Teaching Experience	-6.62	4.18	0.55	-1.88	4.66	1.00	-5.10	4.08	0.79	-4.07	3.32	0.51
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-4.19	5.06	0.97	0.94	3.45	1.00	-2.64	4.70	1.00	-0.68	3.25	0.99

Table L.3 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-5.08	4.33	0.85	-4.44	3.46	0.78	0.28	4.54	1.00	-5.65	2.56	0.09
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	-6.43	4.70	0.71	-12.09*	4.13	0.04	-10.28	4.52	0.18	-8.34*	3.27	0.04
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	-3.07	5.26	1.00	-0.24	3.19	1.00	2.32	4.26	1.00	-0.91	3.09	0.98
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	1.40	4.03	1.00	1.76	3.24	1.00	0.68	4.56	1.00	1.86	2.56	0.83
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	-1.23	4.03	1.00	-5.60	3.49	0.56	4.24	4.49	0.94	-0.46	3.00	1.00
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-0.31	4.30	1.00	0.07	3.42	1.00	-2.00	4.40	1.00	-1.89	3.04	0.89
Above Sample Median Classroom Management Scale Score (499.9)	1.24	2.98	1.00	-1.62	3.57	1.00	-1.67	4.55	1.00	0.38	2.80	1.00

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table are adjusted for multiple-hypotheses testing. The social studies reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE L.4

DIFFERENCES IN EFFECTS ON THE ETS SCIENCE POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	-4.15	4.23	0.94	0.69	3.54	1.00	-1.92	3.41	1.00	-1.72	3.06	0.91
Pretest TOSCRF score, above sample median	-4.06	2.93	0.72	-3.76	4.06	0.95	1.51	3.61	1.00	-1.93	3.03	0.88
Pretest TOSCRF score, top third (vs. bottom)	-8.56	3.65	0.16	-3.64	3.96	0.96	0.51	4.43	1.00	-4.01	3.46	0.56
Pretest TOSCRF score, middle third (vs. bottom)	-5.02	3.20	0.61	1.04	3.21	1.00	-1.59	4.34	1.00	-1.46	2.97	0.94
Pretest TOSCRF score, top third (vs. middle)	-3.48	2.00	0.48	-3.94	2.21	0.45	2.05	2.61	0.98	-1.8	1.87	0.69
Pretest GRADE score, above national norm	0.57	3.16`	1.00	0.56	3.55	1.00	0.21	3.98	1.00	1.76	3.06	0.91
Pretest GRADE score, above sample median	0.57	3.16	1.00	0.56	3.55	1.00	0.21	3.98	1.00	1.76	3.06	0.91
Pretest GRADE score, top third (vs. bottom)	2.79	4.18	0.99	2.64	4.28	1.00	0.68	5.42	1.00	3.55	3.93	0.72
Pretest GRADE score, middle third (vs. bottom)	-2.82	4.62	1.00	-0.04	4.05	1.00	-0.28	5.23	1.00	-0.62	3.94	1.00
Pretest GRADE score, top third (vs. middle)	5.67	2.22	0.10	0.25	2.03	1.00	1.09	2.77	1.00	2.76	1.82	0.33
Classified as ELL	0.79	3.59	1.00	4.42	4.31	0.92	4.73	3.75	0.8	4.01	3.22	0.49
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	-0.49	4.10	1.00	-5.91	5.33	0.88	5.09	4.89	0.91	-1.46	3.71	0.96
More than 5 Years Teaching Experience	-3.21	5.95	1.00	1.13	7.19	1.00	7.51	8.96	0.97	2.03	5.82	0.98
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-5.58	5.18	0.88	0.17	5.69	1.00	-10.93	6.67	0.51	-2.77	5.11	0.91

Table L.4 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-3.76	4.39	0.97	-4.71	5.04	0.95	-15.09*	4.89	0.03	-7.05	4.02	0.21
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	3.29	4.35	0.98	3.49	5.10	0.99	1.17	5.26	1.00	3.46	3.66	0.66
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	2.42	4.23	1.00	-4.22	4.35	0.93	4.33	4.04	0.89	0.69	3.13	0.99
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	-2.17	4.16	1.00	7.81	5.47	0.69	4.25	4.99	0.97	1.76	3.84	0.95
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	1.13	3.92	1.00	-1.51	4.88	1.00	2.84	4.60	1.00	0.84	3.65	0.99
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-4.22	3.61	0.86	-0.85	4.24	1.00	-3.86	4.34	0.96	-3.51	3.20	0.60
Above Sample Median Classroom Management Scale Score (499.9)	-7.59	2.91	0.09	-0.70	3.91	1.00	-3.22	4.21	0.98	-4.12	2.87	0.37

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table are adjusted for multiple-hypotheses testing. The science reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE L.5

DIFFERENCES IN EFFECTS ON THE COMPOSITE FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-0.15	0.06	0.07	0.07	0.06	0.66	-0.13	0.11	0.64	-0.08	0.06	0.60	-0.06	0.04	0.14
Pretest TOSCRF score, above sample median	-0.16	0.07	0.10	0.05	0.06	0.83	-0.02	0.05	0.99	0.01	0.05	1.00	-0.01	0.03	0.67
Pretest TOSCRF score, top third (vs. bottom)	-0.13*	0.05	0.03	0.07	0.04	0.16	-0.07	0.05	0.46	-0.04	0.04	0.82	-0.03	0.03	0.22
Pretest TOSCRF score, middle third (vs. bottom)	-0.04	0.06	0.94	0.17*	0.05	0.01	0.07	0.04	0.37	0.10	0.06	0.30	0.05	0.03	0.12
Pretest TOSCRF score, top third (vs. middle)	-0.06	0.05	0.66	0.00	0.04	1.00	-0.04	0.06	0.9	-0.04	0.03	0.76	-0.03	0.03	0.32
Pretest GRADE score, above national norm	-0.05	0.06	0.81	-0.02	0.06	0.99	-0.06	0.05	0.68	-0.04	0.08	0.96	-0.08	0.05	0.07
Pretest GRADE score, above sample median	-0.05	0.06	0.81	-0.02	0.06	0.99	-0.06	0.05	0.70	-0.04	0.08	0.96	-0.09	0.05	0.07
Pretest GRADE score, top third (vs. bottom)	-0.02	0.04	0.98	0.00	0.04	1.00	0.02	0.05	0.99	-0.01	0.05	1.00	-0.03	0.04	0.39
Pretest GRADE score, middle third (vs. bottom)	-0.06	0.05	0.58	-0.03	0.05	0.93	-0.10	0.04	0.07	-0.03	0.06	0.98	-0.13*	0.04	0.00
Pretest GRADE score, top third (vs. middle)	0.01	0.04	1.00	0.01	0.03	0.99	0.09	0.06	0.32	0.01	0.05	1.00	0.05	0.03	0.06
Classified as ELL	-0.15	0.08	0.24	-0.15	0.08	0.22	0.01	0.10	1.00	-0.15	0.09	0.36	-0.11*	0.04	0.01
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	0.10	0.06	0.44	0.05	0.05	0.82	0.05	0.09	0.98	0.04	0.06	0.97	0.06	0.03	0.10
More than 5 Years Teaching Experience	0.07	0.04	0.34	-0.05	0.06	0.90	-0.06	0.11	0.96	-0.04	0.09	0.99	-0.01	0.04	0.81
Above Sample Median of Teacher Efficacy Scale Score (4.16)	0.05	0.05	0.79	-0.06	0.06	0.74	0.12	0.06	0.26	0.01	0.06	1.00	0.02	0.03	0.46

Table L.5 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	0.09	0.08	0.70	0.03	0.08	0.99	0.01	0.07	1.00	0.02	0.08	1.00	0.04	0.04	0.31
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-0.08	0.10	0.89	-0.06	0.09	0.93	-0.10	0.07	0.56	-0.15	0.11	0.57	-0.08	0.05	0.08
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	0.15	0.13	0.66	0.27	0.13	0.18	0.11	0.11	0.77	0.16	0.12	0.50	0.13	0.08	0.09
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	0.10	0.05	0.29	0.10	0.04	0.07	-0.04	0.08	0.97	-0.14	0.06	0.07	0.01	0.03	0.80
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-0.09	0.06	0.48	-0.15	0.06	0.08	0.03	0.08	0.99	0.02	0.07	0.99	-0.07	0.04	0.08
Above Sample Median Classroom Management Scale Score (502.7)	0.12	0.06	0.22	0.14*	0.05	0.03	0.09	0.07	0.66	0.05	0.07	0.93	0.09*	0.03	0.01

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The *p-values* presented in this table are adjusted for multiple-hypotheses testing. The composite is based on the GRADE and the social studies and science reading comprehension tests. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores. The social studies and science reading comprehension assessments were developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE L.6

DIFFERENCES IN EFFECTS ON THE GRADE FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-2.66*	0.72	0.01	0.14	0.86	1.00	-1.56	0.92	0.63	-1.45	0.85	0.63	-1.35*	0.50	0.03
Pretest TOSCRF score, above sample median	-2.10	1.22	0.63	0.64	0.93	1.00	0.08	0.80	1.00	-0.33	0.83	1.00	-0.46	0.59	0.81
Pretest TOSCRF score, top third (vs. bottom)	-2.12*	0.60	0.01	0.91	0.45	0.40	-0.33	0.54	1.00	-0.66	0.53	0.91	-0.64	0.39	0.27
Pretest TOSCRF score, middle third (vs. bottom)	-0.37	0.92	1.00	1.55	0.84	0.52	0.78	0.56	0.83	0.79	0.77	0.97	0.41	0.49	0.78
Pretest TOSCRF score, top third (vs. middle)	-1.03	0.59	0.60	0.08	0.57	1.00	-0.14	0.42	1.00	-0.17	0.44	1.00	-0.32	0.33	0.69
Pretest GRADE score, above national norm	-0.79	0.86	0.99	0.15	0.94	1.00	-0.36	0.93	1.00	-1.25	1.33	0.99	-1.14	0.78	0.36
Pretest GRADE score, above sample median	-1.03	0.87	0.93	-0.24	0.89	1.00	-0.66	0.85	1.00	-1.11	1.33	0.99	-1.46	0.76	0.16
Pretest GRADE score, top third (vs. bottom)	-0.44	0.74	1.00	-0.28	0.63	1.00	0.48	0.84	1.00	-0.87	0.82	0.96	-0.89	0.61	0.36
Pretest GRADE score, middle third (vs. bottom)	-0.27	0.68	1.00	-0.13	0.88	1.00	-0.69	0.82	0.99	0.16	1.02	1.00	-1.15	0.62	0.19
Pretest GRADE score, top third (vs. middle)	0.08	0.68	1.00	0.29	0.51	1.00	1.46	0.80	0.53	0.08	0.80	1.00	0.32	0.50	0.88
Classified as ELL	-2.14	1.04	0.35	-2.12	1.22	0.58	1.05	2.14	1.00	-1.52	1.36	0.94	-1.23	0.70	0.22
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	1.34	0.91	0.81	0.54	0.88	1.00	0.60	1.35	1.00	0.23	1.07	1.00	0.64	0.54	0.52
More than 5 Years Teaching Experience	1.16	0.80	0.80	-0.52	0.96	1.00	-1.67	1.36	0.92	-0.84	1.13	1.00	-0.38	0.58	0.86
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-0.20	1.23	1.00	-0.73	0.68	0.96	-2.02	1.15	0.59	-0.24	1.19	1.00	-0.57	0.57	0.63

Table L.6 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-0.32	0.80	1.00	-0.66	0.88	1.00	1.45	0.94	0.76	0.19	0.92	1.00	0.07	0.47	1.00
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	0.85	1.00	0.99	1.17	1.22	0.98	0.80	0.94	0.99	0.25	0.88	1.00	0.76	0.54	0.36
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-1.35	1.20	0.92	-1.18	1.26	0.97	-1.94	1.29	0.71	-0.90	0.93	0.97	-1.42	0.68	0.11
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	1.96	1.81	0.92	3.28	2.12	0.66	1.96	1.94	0.95	1.82	1.61	0.91	1.88	1.15	0.22
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	1.58	0.76	0.37	0.54	0.78	1.00	-1.39	1.14	0.93	-2.01	0.86	0.22	-0.36	0.49	0.82
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-1.05	0.88	0.93	-1.06	0.98	0.96	0.72	1.20	1.00	-0.11	0.95	1.00	-0.52	0.53	0.65
Above Sample Median Classroom Management Scale Score (502.7)	2.15*	0.71	0.04	2.26*	0.69	0.02	1.65	0.87	0.47	0.11	0.95	1.00	1.33*	0.43	0.01

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The *p-values* presented in this table are adjusted for multiple-hypotheses testing. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple hypotheses testing.

TABLE L.7

DIFFERENCES IN EFFECTS ON THE ETS SOCIAL STUDIES FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-3.18	3.43	0.99	3.56	3.15	0.95	4.01	5.13	1.00	-1.84	3.54	1.00	0.66	2.28	0.99
Pretest TOSCRF score, above sample median	-5.53	2.85	0.46	-0.90	2.63	1.00	-0.71	2.49	1.00	0.91	2.01	1.00	-0.83	1.75	0.95
Pretest TOSCRF score, top third (vs. bottom)	-4.08	2.56	0.71	1.26	2.01	1.00	-1.75	2.23	1.00	-0.77	2.04	1.00	-0.81	1.47	0.92
Pretest TOSCRF score, middle third (vs. bottom)	-0.86	2.35	1.00	5.85	2.58	0.25	-0.78	1.75	1.00	-0.15	2.04	1.00	0.35	1.49	0.99
Pretest TOSCRF score, top third (vs. middle)	-2.63	2.93	0.99	-0.35	2.51	1.00	0.62	2.99	1.00	-3.34	1.79	0.50	-0.07	1.64	1.00
Pretest GRADE score, above national norm	1.40	2.47	1.00	-1.76	2.67	1.00	-0.29	2.99	1.00	-0.92	2.74	1.00	-1.53	1.84	0.77
Pretest GRADE score, above sample median	0.82	2.49	1.00	-2.26	2.63	0.99	-0.46	2.93	1.00	-0.60	2.79	1.00	-1.80	1.80	0.66
Pretest GRADE score, top third (vs. bottom)	1.37	1.90	1.00	1.20	2.05	1.00	3.16	2.53	0.91	2.56	2.37	0.96	0.79	1.59	0.94
Pretest GRADE score, middle third (vs. bottom)	-1.64	2.56	1.00	-2.85	2.10	0.85	-1.04	2.19	1.00	-3.75	3.38	0.95	-4.49*	1.69	0.03
Pretest GRADE score, top third (vs. middle)	0.01	2.15	1.00	0.83	1.98	1.00	1.48	2.94	1.00	3.09	2.22	0.84	2.75	1.59	0.23
Classified as ELL	-4.51	3.55	0.88	-1.58	4.47	1.00	-13.26	6.13	0.29	-1.96	3.88	1.00	-3.12	2.46	0.48
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	5.49	2.87	0.48	2.32	1.76	0.89	2.62	3.59	1.00	2.60	2.77	0.99	3.27	1.50	0.09
More than 5 Years Teaching Experience	3.02	3.03	0.98	1.11	2.16	1.00	-1.63	4.32	1.00	0.75	3.62	1.00	1.23	1.60	0.79
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	1.86	2.87	1.00	0.35	2.17	1.00	-2.38	2.71	0.99	2.28	3.33	1.00	0.61	1.51	0.96

Table L.7 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	4.13	2.06	0.41	-1.25	2.79	1.00	1.86	3.17	1.00	-0.59	3.08	1.00	1.05	1.44	0.83
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	5.27	2.81	0.47	-1.28	1.92	1.00	0.63	2.80	1.00	-1.15	3.36	1.00	0.82	1.43	0.89
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-3.24	3.81	0.99	-1.46	3.52	1.00	-5.76	3.46	0.60	-10.21	4.57	0.24	-4.28	2.08	0.12
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	2.01	5.12	1.00	10.33	4.09	0.13	0.24	4.75	1.00	9.14	5.21	0.51	3.86	2.93	0.36
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	4.02	2.81	0.83	3.30	1.50	0.29	-0.04	3.30	1.00	-9.01*	2.96	0.04	0.03	1.65	1.00
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-3.88	2.71	0.81	-5.36	2.05	0.11	-2.88	3.64	1.00	4.52	2.95	0.74	-2.48	1.64	0.31
Above Sample Median Classroom Management Scale Score (502.7)	1.38	2.99	1.00	0.18	2.12	1.00	-0.45	3.85	1.00	4.02	3.32	0.92	1.39	1.58	0.72

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The social studies reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE L.8

DIFFERENCES IN EFFECTS ON THE ETS SCIENCE FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-3.53	3.71	0.98	3.82	2.77	0.85	-7.72	3.24	0.19	-0.12	2.79	1.00	-2.21	2.08	0.62
Pretest TOSCRF score, above sample median	-6.12	2.68	0.25	3.94	2.73	0.82	-2.41	2.71	0.99	0.85	2.70	1.00	-0.64	1.80	0.98
Pretest TOSCRF score, top third (vs. bottom)	-2.29	1.78	0.89	2.85	1.86	0.75	-3.95	1.46	0.09	0.03	2.07	1.00	-0.79	1.27	0.89
Pretest TOSCRF score, middle third (vs. bottom)	-2.50	2.65	0.99	6.39	3.40	0.49	5.53	2.73	0.39	8.88	4.24	0.35	3.20	2.38	0.44
Pretest TOSCRF score, top third (vs. middle)	-0.24	3.00	1.00	-0.58	2.40	1.00	-2.08	1.91	0.96	0.80	2.25	1.00	-0.98	1.56	0.89
Pretest GRADE score, above national norm	-1.67	2.93	1.00	1.56	2.38	1.00	-1.46	3.91	1.00	2.71	3.68	1.00	-0.35	2.44	1.00
Pretest GRADE score, above sample median	-1.60	2.76	1.00	0.77	2.29	1.00	-2.24	3.77	1.00	2.35	3.55	1.00	-1.09	2.32	0.95
Pretest GRADE score, top third (vs. bottom)	-1.26	2.51	1.00	2.24	2.18	0.97	-1.90	2.18	0.99	2.07	2.55	1.00	-0.23	1.99	1.00
Pretest GRADE score, middle third (vs. bottom)	-3.93	2.56	0.74	-0.18	3.42	1.00	-5.73	2.58	0.27	-0.39	2.56	1.00	-4.17	2.16	0.16
Pretest GRADE score, top third (vs. middle)	-0.52	2.55	1.00	0.60	2.46	1.00	2.21	2.04	0.96	-0.97	2.83	1.00	1.86	1.67	0.59
Classified as ELL	-4.72	4.19	0.94	-4.82	3.09	0.70	6.35	4.14	0.72	-9.28	5.21	0.54	-2.87	2.47	0.55
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	1.32	2.67	1.00	0.90	2.71	1.00	-0.26	4.23	1.00	1.29	3.36	1.00	1.30	1.58	0.77
More than 5 Years Teaching Experience	0.83	2.55	1.00	-3.04	3.36	0.99	1.48	4.96	1.00	-3.73	5.02	1.00	-0.36	1.96	1.00
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	2.89	3.36	0.99	1.59	2.42	1.00	0.29	3.24	1.00	0.89	3.99	1.00	1.41	1.59	0.71

Table L.8 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	4.42	2.19	0.40	-1.70	2.52	1.00	3.56	2.67	0.88	1.50	3.39	1.00	1.71	1.55	0.58
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	3.54	3.26	0.95	0.67	3.91	1.00	0.18	2.99	1.00	2.55	4.49	1.00	1.40	1.87	0.79
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-0.98	3.57	1.00	-1.16	3.01	1.00	2.45	3.65	1.00	-4.49	6.64	1.00	-0.40	1.98	0.99
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	5.82	4.20	0.77	7.71	3.28	0.18	1.56	3.65	1.00	1.38	4.52	1.00	3.57	2.33	0.26
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	3.58	2.39	0.78	5.34	2.17	0.16	1.21	4.41	1.00	0.24	2.82	1.00	2.31	1.73	0.44
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-2.22	2.64	0.99	-6.30	2.35	0.10	1.50	2.79	1.00	0.47	3.68	1.00	-1.95	1.54	0.46
Above Sample Median Classroom Management Scale Score (502.7)	3.24	3.41	0.98	5.62	2.24	0.14	1.53	2.79	1.00	0.67	3.46	1.00	3.02	1.62	0.17

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The *p-values* presented in this table are adjusted for multiple-hypotheses testing. The science reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are* adjusted for multiple-hypotheses testing.

TABLE L.9

DIFFERENCES IN EFFECTS ON THE COMPOSITE POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	0.04	0.09	0.70	0.03	0.08	0.65	-0.05	0.08	0.57	0.02	0.07	0.78
Pretest TOSCRF score, above sample median	-0.08	0.06	0.16	-0.06	0.06	0.29	0.00	0.06	1.00	-0.04	0.05	0.45
Pretest TOSCRF score, top third (vs. bottom)	-0.12	0.08	0.14	-0.04	0.07	0.57	0.03	0.08	0.68	-0.04	0.06	0.49
Pretest TOSCRF score, middle third (vs. bottom)	-0.09	0.06	0.12	-0.03	0.05	0.60	-0.08	0.08	0.31	-0.06	0.05	0.22
Pretest TOSCRF score, top third (vs. middle)	0.00	0.05	0.93	0.01	0.04	0.79	0.12*	0.05	0.02	0.03	0.04	0.42
Pretest GRADE score, above national norm	-0.03	0.05	0.59	0.02	0.06	0.78	0.01	0.06	0.88	0.02	0.05	0.71
Pretest GRADE score, above sample median	-0.03	0.05	0.59	0.02	0.06	0.78	0.01	0.06	0.88	0.02	0.05	0.71
Pretest GRADE score, top third (vs. bottom)	0.07	0.08	0.40	0.05	0.09	0.60	0.02	0.07	0.81	0.07	0.07	0.34
Pretest GRADE score, middle third (vs. bottom)	-0.05	0.08	0.58	0.02	0.06	0.68	0.03	0.07	0.70	0.01	0.06	0.87
Pretest GRADE score, top third (vs. middle)	0.12*	0.05	0.01	-0.02	0.04	0.61	0.01	0.06	0.91	0.04	0.04	0.21
Classified as ELL	0.14	0.10	0.14	0.09	0.10	0.33	0.46*	0.13	0.00	0.15	0.08	0.07
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	0.03	0.10	0.74	-0.17	0.10	0.12	0.14	0.08	0.08	-0.02	0.07	0.76
More than 5 Years Teaching Experience	-0.16	0.10	0.14	-0.08	0.12	0.50	0.02	0.11	0.83	-0.06	0.08	0.45
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-0.22	0.11	0.05	-0.01	0.11	0.92	-0.22	0.10	0.04	-0.10	0.09	0.27

Table L.9 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-0.13	0.08	0.11	-0.13	0.10	0.22	-0.13	0.08	0.10	-0.15*	0.06	0.03
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	-0.09	0.11	0.40	-0.12	0.13	0.35	-0.03	0.10	0.74	-0.06	0.08	0.48
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	-0.05	0.12	0.65	-0.08	0.08	0.28	0.11	0.12	0.34	-0.01	0.07	0.88
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	0.01	0.09	0.95	0.15	0.08	0.08	0.19	0.11	0.11	0.06	0.07	0.37
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	0.03	0.07	0.64	-0.13	0.08	0.12	0.11	0.08	0.19	-0.01	0.06	0.87
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-0.10	0.08	0.26	0.01	0.07	0.87	-0.03	0.09	0.77	-0.05	0.06	0.35
Above Sample Median Classroom Management Scale Score (499.9)	-0.07	0.07	0.34	-0.01	0.08	0.91	-0.10	0.11	0.36	-0.06	0.06	0.26

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table *are not* adjusted for multiple-hypotheses testing. The composite is based on the GRADE and the social studies and science reading comprehension tests. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores. The social studies and science reading comprehension assessments were developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that *are not* adjusted for multiple-hypotheses testing.

TABLE L.10

DIFFERENCES IN EFFECTS ON THE GRADE POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	0.37	1.36	0.79	0.69	1.24	0.58	-0.30	1.32	0.82	0.57	1.06	0.59
Pretest TOSCRF score, above sample median	-1.15	0.88	0.20	-0.71	0.87	0.42	-0.40	1.00	0.69	-0.52	0.74	0.48
Pretest TOSCRF score, top third (vs. bottom)	-1.52	1.22	0.22	-0.93	1.17	0.43	-0.68	1.28	0.60	-0.83	0.99	0.40
Pretest TOSCRF score, middle third (vs. bottom)	-1.65	1.11	0.14	-1.50	0.89	0.10	-2.49	1.58	0.12	-1.70	0.84	0.05
Pretest TOSCRF score, top third (vs. middle)	0.08	0.86	0.93	0.55	0.71	0.44	1.70	0.96	0.08	0.64	0.74	0.40
Pretest GRADE score, above national norm	-0.97	0.72	0.18	0.76	0.82	0.36	-1.12	0.84	0.19	0.04	0.68	0.95
Pretest GRADE score, above sample median	-0.97	0.72	0.18	0.76	0.82	0.36	-1.12	0.84	0.19	0.04	0.68	0.95
Pretest GRADE score, top third (vs. bottom)	-0.28	1.01	0.79	0.86	1.09	0.43	-1.22	1.02	0.24	0.31	0.90	0.73
Pretest GRADE score, middle third (vs. bottom)	-1.38	1.10	0.22	0.24	0.91	0.80	-0.60	1.23	0.63	-0.22	0.90	0.81
Pretest GRADE score, top third (vs. middle)	0.82	0.69	0.24	-0.27	0.71	0.70	-0.09	1.06	0.93	0.23	0.59	0.69
Classified as ELL	2.97*	1.07	0.01	0.40	1.30	0.76	6.55*	2.52	0.01	1.79	1.07	0.10
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	1.33	1.21	0.28	-1.39	1.49	0.35	1.99	1.14	0.09	0.40	0.95	0.68
More than 5 Years Teaching Experience	-0.74	1.37	0.59	-1.03	1.48	0.49	-0.85	1.33	0.52	-0.85	1.09	0.43
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-2.26	1.53	0.15	-0.06	1.77	0.97	-2.71	1.57	0.09	-1.03	1.34	0.44

Table L.10 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-0.81	1.07	0.45	-1.43	1.55	0.36	-0.69	1.03	0.50	-1.26	0.94	0.19
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	-1.09	1.62	0.50	-1.86	2.11	0.38	0.06	1.54	0.97	-0.80	1.41	0.58
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	-0.92	1.47	0.54	-1.45	1.17	0.22	0.54	1.82	0.77	-0.63	0.98	0.53
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	0.92	1.14	0.43	2.57*	1.20	0.04	3.51*	1.51	0.03	1.58	1.00	0.12
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	1.77	0.89	0.05	-1.56	1.26	0.22	2.37*	1.05	0.03	0.74	0.85	0.39
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-1.44	1.14	0.21	0.34	0.96	0.72	0.19	1.19	0.87	-0.71	0.78	0.36
Above Sample Median Classroom Management Scale Score (499.9)	-0.51	1.10	0.65	0.75	1.27	0.56	-0.14	1.27	0.91	-0.20	0.72	0.78

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table are *not* adjusted for multiple-hypotheses testing. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE L.11

DIFFERENCES IN EFFECTS ON THE ETS SOCIAL STUDIES POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	1.45	2.68	0.59	-2.23	3.68	0.55	-0.62	2.91	0.83	-0.20	2.49	0.94
Pretest TOSCRF score, above sample median	-0.22	2.76	0.94	-0.47	2.57	0.86	2.57	2.88	0.38	0.01	2.34	1.00
Pretest TOSCRF score, top third (vs. bottom)	0.86	3.27	0.79	2.70	3.06	0.38	7.6*	3.54	0.04	2.69	2.74	0.33
Pretest TOSCRF score, middle third (vs. bottom)	1.24	3.79	0.74	3.95	2.79	0.16	6.08	3.32	0.07	3.48	2.67	0.20
Pretest TOSCRF score, top third (vs. middle)	3.11	3.14	0.33	2.76	2.37	0.25	3.08	2.20	0.17	2.69	2.10	0.21
Pretest GRADE score, above national norm	3.29	3.23	0.31	-0.98	2.92	0.74	4.09	3.19	0.21	1.85	2.57	0.48
Pretest GRADE score, above sample median	3.29	3.23	0.31	-0.98	2.92	0.74	4.09	3.19	0.21	1.85	2.57	0.48
Pretest GRADE score, top third (vs. bottom)	8.64	4.29	0.05	0.40	4.34	0.93	3.82	4.92	0.44	4.46	3.62	0.22
Pretest GRADE score, middle third (vs. bottom)	0.01	4.78	1.00	0.14	3.81	0.97	1.77	4.83	0.72	0.78	3.30	0.81
Pretest GRADE score, top third (vs. middle)	6.67	3.47	0.06	-1.05	2.16	0.63	0.90	3.08	0.77	1.74	2.38	0.47
Classified as ELL	4.07	5.20	0.44	1.45	5.36	0.79	12.87*	5.62	0.03	3.59	4.73	0.45
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	1.40	4.79	0.77	-4.79	3.46	0.17	1.82	4.40	0.68	-0.91	3.14	0.77
More than 5 Years Teaching Experience	-6.62	4.18	0.12	-1.88	4.66	0.69	-5.10	4.08	0.22	-4.07	3.32	0.23
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-4.19	5.06	0.41	0.94	3.45	0.79	-2.64	4.70	0.58	-0.68	3.25	0.83

Table L.11 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-5.08	4.33	0.25	-4.44	3.46	0.20	0.28	4.54	0.95	-5.65	2.56	0.03
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	-6.43	4.70	0.18	-12.09*	4.13	0.01	-10.28*	4.52	0.03	-8.34*	3.27	0.01
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	-3.07	5.26	0.56	-0.24	3.19	0.94	2.32	4.26	0.59	-0.91	3.09	0.77
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	1.40	4.03	0.73	1.76	3.24	0.59	0.68	4.56	0.88	1.86	2.56	0.47
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	-1.23	4.03	0.76	-5.60	3.49	0.12	4.24	4.49	0.35	-0.46	3.00	0.88
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-0.31	4.30	0.94	0.07	3.42	0.98	-2.00	4.40	0.65	-1.89	3.04	0.54
Above Sample Median Classroom Management Scale Score (499.9)	1.24	2.98	0.68	-1.62	3.57	0.65	-1.67	4.55	0.72	0.38	2.80	0.89

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table are *not* adjusted for multiple-hypotheses testing. The social studies reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE L.12

DIFFERENCES IN EFFECTS ON THE ETS SCIENCE POST-TEST BETWEEN SUBGROUPS, SECOND COHORT OF FIFTH GRADERS

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement												
Pretest TOSCRF score, above national norm	-4.15	4.23	0.33	0.69	3.54	0.85	-1.92	3.41	0.58	-1.72	3.06	0.58
Pretest TOSCRF score, above sample median	-4.06	2.93	0.17	-3.76	4.06	0.36	1.51	3.61	0.68	-1.93	3.03	0.53
Pretest TOSCRF score, top third (vs. bottom)	-8.56*	3.65	0.02	-3.64	3.96	0.36	0.51	4.43	0.91	-4.01	3.46	0.25
Pretest TOSCRF score, middle third (vs. bottom)	-5.02	3.20	0.12	1.04	3.21	0.75	-1.59	4.34	0.72	-1.46	2.97	0.63
Pretest TOSCRF score, top third (vs. middle)	-3.48	2.00	0.09	-3.94	2.21	0.08	2.05	2.61	0.43	-1.80	1.87	0.34
Pretest GRADE score, above national norm	0.57	3.16	0.86	0.56	3.55	0.88	0.21	3.98	0.96	1.76	3.06	0.57
Pretest GRADE score, above sample median	0.57	3.16	0.86	0.56	3.55	0.88	0.21	3.98	0.96	1.76	3.06	0.57
Pretest GRADE score, top third (vs. bottom)	2.79	4.18	0.51	2.64	4.28	0.54	0.68	5.42	0.90	3.55	3.93	0.37
Pretest GRADE score, middle third (vs. bottom)	-2.82	4.62	0.54	-0.04	4.05	0.99	-0.28	5.23	0.96	-0.62	3.94	0.88
Pretest GRADE score, top third (vs. middle)	5.67*	2.22	0.01	0.25	2.03	0.90	1.09	2.77	0.70	2.76	1.82	0.14
Classified as ELL	0.79	3.59	0.83	4.42	4.31	0.31	4.73	3.75	0.21	4.01	3.22	0.22
Subgroups Defined by Teacher Characteristics												
Above Sample Median Teaching Experience (11 Years)	-0.49	4.10	0.91	-5.91	5.33	0.27	5.09	4.89	0.30	-1.46	3.71	0.70
More than 5 Years Teaching Experience	-3.21	5.95	0.59	1.13	7.19	0.88	7.51	8.96	0.41	2.03	5.82	0.73
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-5.58	5.18	0.29	0.17	5.69	0.98	-10.93	6.67	0.11	-2.77	5.11	0.59

Table L.12 (*continued*)

	Project CRISS			ReadAbout			Read for Real			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-3.76	4.39	0.40	-4.71	5.04	0.35	-15.09*	4.89	0.00	-7.05	4.02	0.09
Subgroups Defined by School Characteristics												
In Schools with Professional Culture Scale Score Above Sample Median (5.68)	3.29	4.35	0.45	3.49	5.10	0.50	1.17	5.26	0.83	3.46	3.66	0.35
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (69 percent)	2.42	4.23	0.57	-4.22	4.35	0.34	4.33	4.04	0.29	0.69	3.13	0.83
In Schools with Proportion of Students Classified as ELLs Above Sample Median (15.5 percent)	-2.17	4.16	0.61	7.81	5.47	0.16	4.25	4.99	0.40	1.76	3.84	0.65
Subgroups Defined by Teacher Practices												
Above Sample Median Traditional Interaction Scale Score (499.7)	1.13	3.92	0.77	-1.51	4.88	0.76	2.84	4.60	0.54	0.84	3.65	0.82
Above Sample Median Reading Strategy Guidance Scale Score (500.4)	-4.22	3.61	0.25	-0.85	4.24	0.84	-3.86	4.34	0.38	-3.51	3.20	0.28
Above Sample Median Classroom Management Scale Score (499.9)	-7.59*	2.91	0.01	-0.70	3.91	0.86	-3.22	4.21	0.45	-4.12	2.87	0.16

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 2 students who are (and are not) classified as ELL is statistically significant. In other words, in the second year, do Cohort 2 ELL students experience larger impacts of the interventions than Cohort 2 students not classified as ELL? The *p-values* presented in this table are *not* adjusted for multiple-hypotheses testing. The composite is based on the GRADE and the social studies and science reading comprehension tests. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores. The science reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE L.13

DIFFERENCES IN EFFECTS ON THE COMPOSITE FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-0.15*	0.06	0.02	0.07	0.06	0.24	-0.13	0.11	0.23	-0.08	0.06	0.21	-0.06	0.04	0.14
Pretest TOSCRF score, above sample median	-0.16*	0.07	0.03	0.05	0.06	0.37	-0.02	0.05	0.68	0.01	0.05	0.88	-0.01	0.03	0.67
Pretest TOSCRF score, top third (vs. bottom)	-0.13*	0.05	0.01	0.07*	0.04	0.04	-0.07	0.05	0.15	-0.04	0.04	0.36	-0.03	0.03	0.22
Pretest TOSCRF score, middle third (vs. bottom)	-0.04	0.06	0.52	0.17*	0.05	0.00	0.07	0.04	0.11	0.10	0.06	0.09	0.05	0.03	0.12
Pretest TOSCRF score, top third (vs. middle)	-0.06	0.05	0.24	0.00	0.04	0.98	-0.04	0.06	0.45	-0.04	0.03	0.31	-0.03	0.03	0.32
Pretest GRADE score, above national norm	-0.05	0.06	0.36	-0.02	0.06	0.70	-0.06	0.05	0.27	-0.04	0.08	0.58	-0.08	0.05	0.07
Pretest GRADE score, above sample median	-0.05	0.06	0.36	-0.02	0.06	0.70	-0.06	0.05	0.28	-0.04	0.08	0.58	-0.09	0.05	0.07
Pretest GRADE score, top third (vs. bottom)	-0.02	0.04	0.66	0.00	0.04	0.96	0.02	0.05	0.69	-0.01	0.05	0.84	-0.03	0.04	0.39
Pretest GRADE score, middle third (vs. bottom)	-0.06	0.05	0.22	-0.03	0.05	0.51	-0.10	0.04	0.02	-0.03	0.06	0.66	-0.13*	0.04	0.00
Pretest GRADE score, top third (vs. middle)	0.01	0.04	0.80	0.01	0.03	0.72	0.09	0.06	0.10	0.01	0.05	0.83	0.05	0.03	0.06
Classified as ELL	-0.15	0.08	0.07	-0.15	0.08	0.06	0.01	0.10	0.90	-0.15	0.09	0.11	-0.11*	0.04	0.01
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	0.10	0.06	0.14	0.05	0.05	0.35	0.05	0.09	0.61	0.04	0.06	0.58	0.06	0.03	0.10
More than 5 Years Teaching Experience	0.07	0.04	0.10	-0.05	0.06	0.45	-0.06	0.11	0.56	-0.04	0.09	0.70	-0.01	0.04	0.81
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	0.03	0.09	0.76	-0.01	0.04	0.88	-0.12	0.08	0.15	0.00	0.07	0.96	-0.01	0.04	0.82

Table L.13 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	0.05	0.05	0.33	-0.06	0.06	0.29	0.12	0.06	0.07	0.01	0.06	0.86	0.02	0.03	0.46
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	0.09	0.08	0.26	0.03	0.08	0.70	0.01	0.07	0.84	0.02	0.08	0.78	0.04	0.04	0.31
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-0.08	0.10	0.43	-0.06	0.09	0.49	-0.10	0.07	0.19	-0.15	0.11	0.19	-0.08	0.05	0.08
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	0.15	0.13	0.24	0.27	0.13	0.05	0.11	0.11	0.32	0.16	0.12	0.17	0.13	0.08	0.09
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	0.10	0.05	0.08	0.10*	0.04	0.02	-0.04	0.08	0.58	-0.14*	0.06	0.02	0.01	0.03	0.80
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-0.09	0.06	0.15	-0.15*	0.06	0.02	0.03	0.08	0.73	0.02	0.07	0.73	-0.07	0.04	0.08
Above Sample Median Classroom Management Scale Score (502.7)	0.12	0.06	0.06	0.14*	0.05	0.01	0.09	0.07	0.24	0.05	0.07	0.49	0.09*	0.03	0.01

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The *p-values* presented in this table are *not* adjusted for multiple-hypotheses testing. The composite is based on the GRADE and the social studies and science reading comprehension tests. Each test score is converted into a z-score by subtracting the mean and dividing by the standard deviation of the variable for students in the sample. The composite is the simple average of the three z-scores. The social studies and science reading comprehension assessments were developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE L.14

DIFFERENCES IN EFFECTS ON THE GRADE FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-2.66*	0.72	0.00	0.14	0.86	0.87	-1.56	0.92	0.09	-1.45	0.85	0.09	-1.35*	0.50	0.01
Pretest TOSCRF score, above sample median	-2.10	1.22	0.09	0.64	0.93	0.49	0.08	0.80	0.92	-0.33	0.83	0.69	-0.46	0.59	0.43
Pretest TOSCRF score, top third (vs. bottom)	-2.12*	0.60	0.00	0.91	0.45	0.05	-0.33	0.54	0.55	-0.66	0.53	0.22	-0.64	0.39	0.10
Pretest TOSCRF score, middle third (vs. bottom)	-0.37	0.92	0.69	1.55	0.84	0.07	0.78	0.56	0.17	0.79	0.77	0.30	0.41	0.49	0.41
Pretest TOSCRF score, top third (vs. middle)	-1.03	0.59	0.09	0.08	0.57	0.89	-0.14	0.42	0.74	-0.17	0.44	0.70	-0.32	0.33	0.34
Pretest GRADE score, above national norm	-0.79	0.86	0.36	0.15	0.94	0.88	-0.36	0.93	0.70	-1.25	1.33	0.35	-1.14	0.78	0.15
Pretest GRADE score, above sample median	-1.03	0.87	0.24	-0.24	0.89	0.78	-0.66	0.85	0.44	-1.11	1.33	0.41	-1.46	0.76	0.06
Pretest GRADE score, top third (vs. bottom)	-0.44	0.74	0.56	-0.28	0.63	0.66	0.48	0.84	0.57	-0.87	0.82	0.29	-0.89	0.61	0.15
Pretest GRADE score, middle third (vs. bottom)	-0.27	0.68	0.69	-0.13	0.88	0.88	-0.69	0.82	0.41	0.16	1.02	0.88	-1.15	0.62	0.07
Pretest GRADE score, top third (vs. middle)	0.08	0.68	0.91	0.29	0.51	0.57	1.46	0.80	0.07	0.08	0.80	0.92	0.32	0.50	0.52
Classified as ELL	-2.14	1.04	0.04	-2.12	1.22	0.09	1.05	2.14	0.63	-1.52	1.36	0.27	-1.23	0.70	0.09
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	1.34	0.91	0.15	0.54	0.88	0.54	0.60	1.35	0.66	0.23	1.07	0.83	0.64	0.54	0.24
More than 5 Years Teaching Experience	1.16	0.80	0.15	-0.52	0.96	0.59	-1.67	1.36	0.22	-0.84	1.13	0.46	-0.38	0.58	0.52
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	-0.20	1.23	0.87	-0.73	0.68	0.29	-2.02	1.15	0.09	-0.24	1.19	0.84	-0.57	0.57	0.32

Table L.14 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	-0.32	0.80	0.69	-0.66	0.88	0.45	1.45	0.94	0.13	0.19	0.92	0.84	0.07	0.47	0.88
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	0.85	1.00	0.40	1.17	1.22	0.34	0.80*	0.94	0.40	0.25	0.88	0.78	0.76	0.54	0.16
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-1.35	1.20	0.27	-1.18	1.26	0.35	-1.94	1.29	0.14	-0.90	0.93	0.34	-1.42*	0.68	0.04
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	1.96	1.81	0.29	3.28	2.12	0.13	1.96	1.94	0.32	1.82	1.61	0.26	1.88	1.15	0.11
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	1.58*	0.76	0.04	0.54	0.78	0.49	-1.39	1.14	0.23	-2.01*	0.86	0.02	-0.36	0.49	0.46
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-1.05	0.88	0.24	-1.06	0.98	0.28	0.72	1.20	0.55	-0.11	0.95	0.91	-0.52	0.53	0.33
Above Sample Median Classroom Management Scale Score (502.7)	2.15*	0.71	0.00	2.26*	0.69	0.00	1.65	0.87	0.06	0.11	0.95	0.91	1.33*	0.43	0.00

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The *p-values* presented in this table are *not* adjusted for multiple-hypotheses testing. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE L.15

DIFFERENCES IN EFFECTS ON THE ETS SOCIAL STUDIES FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-3.18	3.43	0.36	3.56	3.15	0.26	4.01	5.13	0.44	-1.84	3.54	0.61	0.66	2.28	0.77
Pretest TOSCRF score, above sample median	-5.53	2.85	0.06	-0.90	2.63	0.73	-0.71	2.49	0.78	0.91	2.01	0.65	-0.83	1.75	0.64
Pretest TOSCRF score, top third (vs. bottom)	-4.08	2.56	0.12	1.26	2.01	0.53	-1.75	2.23	0.43	-0.77	2.04	0.71	-0.81	1.47	0.58
Pretest TOSCRF score, middle third (vs. bottom)	-0.86	2.35	0.72	5.85*	2.58	0.03	-0.78	1.75	0.66	-0.15	2.04	0.94	0.35	1.49	0.82
Pretest TOSCRF score, top third (vs. middle)	-2.63	2.93	0.37	-0.35	2.51	0.89	0.62	2.99	0.84	-3.34	1.79	0.07	-0.07	1.64	0.96
Pretest GRADE score, above national norm	1.40	2.47	0.57	-1.76	2.67	0.51	-0.29	2.99	0.92	-0.92	2.74	0.74	-1.53	1.84	0.41
Pretest GRADE score, above sample median	0.82	2.49	0.74	-2.26	2.63	0.39	-0.46	2.93	0.87	-0.60	2.79	0.83	-1.80	1.80	0.32
Pretest GRADE score, top third (vs. bottom)	1.37	1.90	0.47	1.20	2.05	0.56	3.16	2.53	0.22	2.56	2.37	0.28	0.79	1.59	0.62
Pretest GRADE score, middle third (vs. bottom)	-1.64	2.56	0.52	-2.85	2.10	0.18	-1.04	2.19	0.64	-3.75	3.38	0.27	-4.49*	1.69	0.01
Pretest GRADE score, top third (vs. middle)	0.01	2.15	1.00	0.83	1.98	0.68	1.48	2.94	0.62	3.09	2.22	0.17	2.75	1.59	0.09
Classified as ELL	-4.51	3.55	0.21	-1.58	4.47	0.72	-13.26*	6.13	0.03	-1.96	3.88	0.62	-3.12	2.46	0.21
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	5.49	2.87	0.06	2.32	1.76	0.19	2.62	3.59	0.47	2.60	2.77	0.35	3.27*	1.50	0.03
More than 5 Years Teaching Experience	3.02	3.03	0.32	1.11	2.16	0.61	-1.63	4.32	0.71	0.75	3.62	0.84	1.23	1.60	0.44
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	1.86	2.87	0.52	0.35	2.17	0.87	-2.38	2.71	0.38	2.28	3.33	0.50	0.61	1.51	0.69

Table L.15 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	4.13	2.06	0.05	-1.25	2.79	0.66	1.86	3.17	0.56	-0.59	3.08	0.85	1.05	1.44	0.47
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	5.27	2.81	0.06	-1.28	1.92	0.51	0.63	2.80	0.82	-1.15	3.36	0.73	0.82	1.43	0.57
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-3.24	3.81	0.40	-1.46	3.52	0.68	-5.76	3.46	0.10	-10.21*	4.57	0.03	-4.28	2.08	0.05
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	2.01	5.12	0.70	10.33*	4.09	0.02	0.24	4.75	0.96	9.14	5.21	0.09	3.86	2.93	0.20
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	4.02	2.81	0.16	3.30*	1.50	0.03	-0.04	3.30	0.99	-9.01*	2.96	0.00	0.03	1.65	0.98
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-3.88	2.71	0.16	-5.36*	2.05	0.01	-2.88	3.64	0.43	4.52	2.95	0.13	-2.48	1.64	0.13
Above Sample Median Classroom Management Scale Score (502.7)	1.38	2.99	0.65	0.18	2.12	0.93	-0.45	3.85	0.91	4.02	3.32	0.23	1.39	1.58	0.38

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The *p-values* presented in this table are *not* adjusted for multiple-hypotheses testing. The social studies reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.

TABLE L.16

DIFFERENCES IN EFFECTS ON THE ETS SCIENCE FOLLOW-UP TEST BETWEEN SUBGROUPS, FIRST COHORT

	Project CRISS			ReadAbout			Read for Real			Reading for Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Subgroups Defined by Student Characteristics and Prior Achievement															
Pretest TOSCRF score, above national norm	-3.53	3.71	0.34	3.82	2.77	0.17	-7.72*	3.24	0.02	-0.12	2.79	0.97	-2.21	2.08	0.29
Pretest TOSCRF score, above sample median	-6.12*	2.68	0.03	3.94	2.73	0.15	-2.41	2.71	0.38	0.85	2.70	0.75	-0.64	1.80	0.72
Pretest TOSCRF score, top third (vs. bottom)	-2.29	1.78	0.20	2.85	1.86	0.13	-3.95*	1.46	0.01	0.03	2.07	0.99	-0.79	1.27	0.54
Pretest TOSCRF score, middle third (vs. bottom)	-2.50	2.65	0.35	6.39	3.40	0.06	5.53	2.73	0.05	8.88*	4.24	0.04	3.20	2.38	0.18
Pretest TOSCRF score, top third (vs. middle)	-0.24	3.00	0.94	-0.58	2.40	0.81	-2.08	1.91	0.28	0.80	2.25	0.72	-0.98	1.56	0.53
Pretest GRADE score, above national norm	-1.67	2.93	0.57	1.56	2.38	0.51	-1.46	3.91	0.71	2.71	3.68	0.46	-0.35	2.44	0.89
Pretest GRADE score, above sample median	-1.60	2.76	0.57	0.77	2.29	0.74	-2.24	3.77	0.55	2.35	3.55	0.51	-1.09	2.32	0.64
Pretest GRADE score, top third (vs. bottom)	-1.26	2.51	0.62	2.24	2.18	0.31	-1.90	2.18	0.39	2.07	2.55	0.42	-0.23	1.99	0.91
Pretest GRADE score, middle third (vs. bottom)	-3.93	2.56	0.13	-0.18	3.42	0.96	-5.73*	2.58	0.03	-0.39	2.56	0.88	-4.17	2.16	0.06
Pretest GRADE score, top third (vs. middle)	-0.52	2.55	0.84	0.60	2.46	0.81	2.21	2.04	0.28	-0.97	2.83	0.73	1.86	1.67	0.27
Classified as ELL	-4.72	4.19	0.26	-4.82	3.09	0.12	6.35	4.14	0.13	-9.28	5.21	0.08	-2.87	2.47	0.25
Subgroups Defined by Teacher Characteristics															
Above Sample Median Teaching Experience (10 Years)	1.32	2.67	0.62	0.90	2.71	0.74	-0.26	4.23	0.95	1.29	3.36	0.70	1.30	1.58	0.41
More than 5 Years Teaching Experience	0.83	2.55	0.75	-3.04	3.36	0.37	1.48	4.96	0.77	-3.73	5.02	0.46	-0.36	1.96	0.86
Above Sample Median Teacher Reading Instruction Professional Development (12.5 hours)	2.89	3.36	0.39	1.59	2.42	0.51	0.29	3.24	0.93	0.89	3.99	0.82	1.41	1.59	0.38

Table L.16 (*continued*)

	Project CRISS			ReadAbout			Read For Real			Reading For Knowledge			Combined Treatment Group		
	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*	Estimate	Standard Error	*p-value*
Above Sample Median of Teacher Efficacy Scale Score (4.16)	4.42	2.19	0.05	-1.7	2.52	0.50	3.56	2.67	0.19	1.50	3.39	0.66	1.71	1.55	0.27
Subgroups Defined by School Characteristics															
In Schools with Professional Culture Scale Score Above Sample Median (5.67)	3.54	3.26	0.28	0.67	3.91	0.86	0.18	2.99	0.95	2.55	4.49	0.57	1.40	1.87	0.46
In Schools with Proportion of Students Eligible for Free or Reduced-Price Lunch Above Sample Median (68 percent)	-0.98	3.57	0.78	-1.16	3.01	0.70	2.45	3.65	0.51	-4.49	6.64	0.50	-0.40	1.98	0.84
In Schools with Proportion of Students Classified as ELLs Above Sample Median (12 percent)	5.82	4.20	0.17	7.71*	3.28	0.02	1.56	3.65	0.67	1.38	4.52	0.76	3.57	2.33	0.13
Subgroups Defined by Teacher Practices															
Above Sample Median Traditional Interaction Scale Score (499.5)	3.58	2.39	0.14	5.34*	2.17	0.02	1.21	4.41	0.79	0.24	2.82	0.93	2.31	1.73	0.19
Above Sample Median Reading Strategy Guidance Scale Score (500.0)	-2.22	2.64	0.40	-6.03*	2.35	0.01	1.50	2.79	0.59	0.47	3.68	0.90	-1.95	1.54	0.21
Above Sample Median Classroom Management Scale Score (502.7)	3.24	3.41	0.35	5.62*	2.24	0.01	1.53	2.79	0.59	0.67	3.46	0.85	3.02	1.62	0.07

SOURCE: Reading comprehension tests administered by study team.

NOTE: The estimates presented in this table reflect whether there is a differential impact in the second year for the subgroup listed. For example, for ELL status, the estimates in this row allow one to determine whether the difference in impacts of the interventions between Cohort 1 students who are (and are not) classified as ELL is statistically significant in the second year of the study, when they were in sixth grade. In other words, in the second year, do Cohort 1 ELL students experience larger sustained impacts of the interventions than Cohort 1 students not classified as ELL? The *p-values* presented in this table are *not* adjusted for multiple-hypotheses testing. The science reading comprehension assessment was developed by ETS. Variables in the regression model include pretest GRADE and TOSCRF scores, student ethnicity and race, student ELL status, school location, teacher race, and district indicators.

ELL = English language learner; ETS = Educational Testing Service; GRADE = Group Reading Assessment and Diagnostic Evaluation; TOSCRF = Test of Silent Contextual Reading Fluency.

*Significantly different from zero at the .05 level. This measure of statistical significance is based on *p-values* that are *not* adjusted for multiple-hypotheses testing.