NCEE 2010–4035

U.S. DEPARTMENT OF EDUCATION

Impact of the Thinking Reader® Software Program on Grade 6 Reading Vocabulary, Comprehension, Strategies, and Motivation

Final Report

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

REL NORTHEAST & ISLANDS
Regional Educational Laboratory

Impact of the *Thinking Reader*® Software Program on Grade 6 Reading Vocabulary, Comprehension, Strategies, and Motivation

April 2011

Authors

Kathryn Drummond, American Institutes for Research

Marjorie Chinen, American Institutes for Research

Teresa Garcia Duncan, American Institutes for Research

H. Ray Miller, American Institutes for Research

Lindsay Fryer, American Institutes for Research

Courtney Zmach, American Institutes for Research

Katherine Culp, Education Development Center, Inc.

NCEE 2010-4035

U.S. Department of Education

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
Rebecca A. Maynard
Commissioner

April 2011

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED-01-CO-0060/0025 with Education Development Center, Inc.

While permission to reprint this publication is not necessary, the citation should read: Drummond, K., Chinen, M., Duncan, T.G., Miller, H.R., Fryer, L., Zmach, C., & Culp, K. (2011). *Impact of the Thinking Reader® software program on grade 6 reading vocabulary, comprehension, strategies, and motivation* (NCEE 2010-4035). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

To order copies of this report,

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

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

Alternate Formats Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, call the Alternate Format Center at 202-205-8113.

Disclosure of Potential Conflict of Interest[1]

Data collection, data analysis, and report-writing for this evaluation were conducted by the American Institutes for Research® (AIR®) and its subcontractor Sun Associates. Neither the organizations nor their key staff have financial interests that could be affected by findings from the evaluation. Also involved in the study was The Center for Applied Special Technology (CAST), the developer of *Thinking Reader*®, and Tom Snyder Productions®, the software distributor, who provided teacher training. These two organizations do have financial interest that could be affected by the evaluation but were not involved in the research activities.

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

Acknowledgments

This study represents a collaborative effort among the authors and staff from the participating schools, the Center for Applied Special Technology (CAST), Tom Snyder Productions, Sun Associates, and our colleagues at AIR and Educational Development Center (EDC). The authors thank Hans Bos and Anja Kurki for their constructive comments and review of the report. We also acknowledge the helpful input of Peggy Clements, Craig Hoyle, and Jessica Heppen during the course of the study, as well as Nick Sorenson and Suzanne Stachel, who assisted with data analysis. The expert advice of Mike Garet at various stages of data analysis is much appreciated. The authors acknowledge contributions of Melissa Collier and Judith Littman who served in key roles with data collection and data cleaning activities. Andy Peterman and Holly Baker, along with the entire AIR production department, were instrumental in editing and preparing the document. The authors also thank the many additional individuals who conducted data collection.

The Regional Educational Laboratory Northeast and Islands' (REL-NEI) technical working group provided insight and advice on the evaluation design. We thank Larry Aber, Larry Hedges, Steve Klein, Don Leu, Mike Nettles, Aline Sayer, and Barbara Schneider for their guidance. Support for the study in its early stages from REL NEI's Governing Board helped make the study possible.

We particularly acknowledge and thank the hard-working students, teachers, principals, and district and state leaders who made this study possible.

Contents

Tables

Figures

Boxes

Exhibits

Executive Summary

Improving adolescent literacy is a critical step toward improving adolescent academic achievement (Kamil , Borman, Dole, Kral, Salinger, & Torgesen, 2008). “Adolescent literacy” commonly refers to the skills that students in Grades 4–12 need in order to successfully learn by reading, as opposed to learning how to read, which is emphasized in earlier grades (Kamil, 2003; Kamil et al., 2008; National Governors Association, 2005). Recent policy reports emphasize the need to build students’ reading vocabulary and comprehension skills to meet the increased literacy demands that begin in Grade 4 (Carnegie Council on Advancing Adolescent Literacy, 2010; Meltzer, Smith, & Clark, 2001). Experts who drafted the Common Core State Standards for English Language Arts have emphasized that students must show a steadily increasing ability to discern more from text to become successful readers (National Governors Association & Council of Chief State School Officers, 2010). The current study evaluates an intervention (*Thinking Reader*®) designed to improve middle school students’ reading vocabulary and comprehension (Tom Snyder Productions, 2006a). It responds to an interest expressed by stakeholders to the Regional Educational Laboratory Northeast and Islands in improving literacy outcomes for students beyond elementary school.

Thinking Reader is a software program for students in Grades 5–8 that incorporates elements commonly identified in policy reports as being key components of effective adolescent literacy instruction. These reports prioritize elements such as instruction in comprehension strategies (Biancarosa & Snow, 2004; Carnegie Council on Advancing Adolescent Literacy, 2010; Meltzer et al., 2001); attention to motivation and self-directed learning (Biancarosa & Snow, 2004; Meltzer et al., 2001; National Council of Teachers of English, 2006); ongoing formative assessment (Biancarosa & Snow, 2004; Carnegie Council on Advancing Adolescent Literacy, 2010; National Council of Teachers of English, 2006); and inclusion of technology as an instructional tool (Biancarosa & Snow, 2004; Kamil, 2003; National Council of Teachers of English, 2006).

When using *Thinking Reader*, students read novels on computers and respond to prompts. The software aims to teach students to use comprehension strategies through a reciprocal teaching approach in which the strategies are taught while the teacher and students explore the meaning of text (Palincsar & Brown, 1984). Students’ progress is assessed regularly. *Thinking Reader* also aims to motivate students to read and to make self-directed use of strategies. The software has a limited but positive evidence base, including statistically significant findings of one quasi-experimental study (Dalton, Pisha, Eagleton, Coyne, & Deysher, 2002) and empirical evidence supporting instruction on comprehension strategy use (for example, RAND Reading Study Group, 2002), particularly on strategy instruction through reciprocal teaching approaches (for example, Rosenshine & Meister, 1994). The current study is the first rigorous, randomized controlled trial on the program.

Research Questions and Measures

This evaluation of the impact of *Thinking Reader* use by Grade 6 students focused on two confirmatory research questions about the effect of the program on two measures of students’ reading achievement:

1. What is the effect of *Thinking Reader* on students' reading vocabulary?

2. What is the effect of *Thinking Reader* on students' reading comprehension?

A statistically significant impact on either outcome measure would signal the program's success.

The study also examined whether *Thinking Reader* has an effect on two ancillary, but important, measures of students' approaches to reading:

1. What is the effect of *Thinking Reader* on students' use of reading comprehension strategies?

2. What is the effect of *Thinking Reader* on students' motivation to read?

The answers to these questions provide information that may be useful to educators who see these factors as being important precursors or supplements to improved achievement. But without a direct, measurable effect on reading achievement itself (vocabulary, comprehension), such effects would be insufficient to judge the program's effectiveness.

The vocabulary and reading comprehension subtests of the Gates-MacGinitie Reading Tests (GMRT) (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 1999) served as the achievement measures for the primary research questions. Two self-report student surveys—the Metacognitive Awareness of Reading Strategies Inventory (MARSI; Mokhtari & Reichard, 2002) and the Motivation for Reading Questionnaire (MRQ; Wigfield & Guthrie, 1997)—served as the measures for the ancillary research questions. Each was collected as both a pretest fall baseline measure before the start of the intervention and as a posttest spring outcome measure.

This study also addressed four exploratory research questions. These questions investigate whether the impact of the *Thinking Reader* intervention on students' reading achievement varied across subgroups of students formed on the basis of baseline reading vocabulary, baseline reading comprehension, and baseline motivation to read measures:

1. Does the effect of *Thinking Reader* on students' reading vocabulary vary according to their baseline reading vocabulary scores?

2. Does the effect of *Thinking Reader* on students' reading comprehension vary according to their baseline reading comprehension scores?

3. Does the effect of *Thinking Reader* on students' reading vocabulary vary according to their baseline reading motivation scores?

4. Does the effect of *Thinking Reader* on students' reading comprehension vary according to their baseline reading motivation scores?

The outcomes of interest for these exploratory research questions are the vocabulary and comprehension subtests of the GMRT that served as the achievement measures for the primary research questions.

These exploratory research questions were selected because the *Thinking Reader* program is intended to provide differentiated support to students with different skill levels. Specific features of the program are intended to help teachers monitor progress and facilitate teachers' individualization of instruction based on students' demonstrated performance. Exploratory Research Questions 1 and 2 are important empirical questions because the literature contains very limited evidence to indicate whether an intervention like *Thinking Reader* might be more or less beneficial for students with lower or higher baseline achievement scores. Exploratory Research Questions 3 and 4 were selected because, although an interactive software program could be motivating for adolescent readers, rigorous evidence does not exist to indicate whether an intervention like *Thinking Reader* might be more or less beneficial for students with lower or higher self-reported baseline motivation levels.

To support the exploratory analyses, we partitioned students' baseline achievement scores into subgroups. The sample was divided into three groups representing the lowest, middle, and highest achieving tertiles. Tertiles[2] were used instead of the continuous score distribution in order to allow the analysis to demonstrate a possible non-linear effect of the treatment across the baseline achievement score distribution. The intervention could be ineffective for the majority of students, which could drive the results demonstrated for the sample as a whole, but results from the tertiles may be able to detect a non-zero effect on a particular, smaller subgroup of students.

The study collected student, teacher, and school data as covariates for the analyses and collected data from classroom observations and electronic report data from *Thinking Reader* to help understand program implementation. Student data included age, gender, and ethnicity, as well as English language learner and special education status. Teacher data included information about years of teaching experience, educational attainment, and certifications or endorsements held. School data included type of school (elementary or middle school), state, enrollment size, and the poverty level and ethnicity of students. For a subset of classrooms, structured observations in the winter and spring in one intervention and one control classroom at each school provided descriptive information about instruction. To examine fidelity of implementation, electronic report data were collected from the *Thinking Reader* program at the end of the year. These data provided information about students' exposure to the software (such as number of books started and completed, total number of minutes using the software, and number of weeks spent on each book), as well as students' program levels.

Description of *Thinking Reader*

Students using *Thinking Reader* read novels on computers and respond to prompts that support a range of strategies for understanding text. Users can choose from nine novels with a range of difficulty appropriate for middle school readers.[3] Development of the program by the Center for

[2] Tertiles were used rather than quartiles or quintiles to avoid having cells with very small sample sizes. The original power analysis was not calculated to accommodate dividing the sample into subgroups.

[3] Software licenses for *Thinking Reader* are available for individual computers, for bundles of 10 licenses, and as an unlimited site license. Discounts are given on orders for multiple *Thinking Reader* novels and for multiple schools within a district. As of spring 2010, prices per novel ranged from $250 to $2,200, depending on the type of license and number of different novels ordered. Participating schools received unlimited site licenses, at no cost, for three *Thinking Reader* titles during the study year and an additional fourth title for the subsequent year.

Applied Special Technology (CAST) began in 2002, and marketing of the program by Tom Snyder Productions/Scholastic began in 2004.[4]

The program was designed to embody reciprocal teaching (Brown & Palincsar, 1985; Palincsar, 1982; Palincsar & Brown, 1984), part of a pedagogical approach for explicitly teaching students cognitive strategies to help them understand text. Reciprocal teaching guides students in applying four concrete strategies for comprehending text—summarizing, questioning, clarifying, and predicting. In *summarizing*, students explain important information from a piece of text. *Questioning* involves querying a concept in the text to identify key information. *Clarifying* teaches students to question concepts or words that are unclear. *Predicting* involves using what is known from the story to hypothesize what will happen next.

In addition to these four strategies, *Thinking Reader* also uses three others—visualizing, feeling, and reflecting. *Visualizing* instructs readers to create mental images of what they are reading. *Feeling* aims to connect students emotionally with what they are reading. *Reflecting* asks students to monitor their progress as readers. Although the last three strategies are not formally part of reciprocal teaching, the *Thinking Reader* program uses the reciprocal teaching approach to teach all seven strategies.

In reciprocal teaching, teachers model specific comprehension strategies and support students' efforts to recall and employ those strategies in their reading. Teachers and students engage in an instructional dialogue about text in which teachers initially lead—demonstrating, modeling, and providing feedback about strategy use. As students become more competent using the strategies, teachers decrease their modeling and students increasingly apply the strategies on their own. In *Thinking Reader*, animated coaches and peers on the computer act as instructors, modeling the seven comprehension strategies and prompting students to use them.

Thinking Reader allows for individualization of instruction. Teachers can customize the amount of support to each student by choosing among five levels. The leveling system varies the representation of the strategy task, students' response options, and the availability of the animated coaches. Students can progress from responding to highly structured strategy prompts (Level 1) to independently selecting their own strategies (Level 5).

Teachers are encouraged to use *Thinking Reader* in a three-phase instructional routine. In the first phase, teachers introduce the program offline through activities, such as modeling a strategy or asking for a summary from the previous session. In the second phase, students read the *Thinking Reader* novel on the computer while the teacher observes, reviews the computer work, and conducts conferences with students about the quality of responses. The third phase involves additional offline teacher–student interaction, discussing the book or completing an activity to illustrate understanding.

[4] According to Tom Snyder Productions, prior to the start of the study, *Thinking Reader* was being used at approximately 760 schools in 46 states by 67,000 students. These figures are estimates based on the number of software licenses sold.

Recruitment, Statistical Power, and Study Conditions

High-need schools were targeted for the study because of the link between high economic need and low reading achievement (National Center for Education Statistics, 2009a). Recruitment was limited to schools with more than 33% of students eligible for free or reduced-price lunch (high-poverty schools). A total of 92 teachers participated from 32 elementary and middle schools in 16 districts in Connecticut, Massachusetts, and Rhode Island. With a sample size of 2,407 students and assumptions about the correlation between pretest and posttest scores, the study has the statistical power to detect an effect size of 0.19–0.24 standard deviations.

In each school, Grade 6 reading/English language arts teachers were randomly assigned to intervention and control groups after students had been scheduled into classes using the typical school procedures. The random assignment produced two groups that did not have statistically significant differences in teacher characteristics—including education, certification, and experience—or in student achievement measures of reading vocabulary and comprehension and self-report measures of comprehension strategy use and motivation for reading.

Intervention teachers received three *Thinking Reader* digital novels to read with their students during the 2008–09 academic year. Teachers were asked to participate in professional development (two, 6-hour group sessions and three individual coaching sessions totaling 7.5–8.5 hours) to learn the software and how to use the three-phase instructional routine. They were asked to incorporate *Thinking Reader* into their regular English language arts or reading instruction, which was to last 110–165 minutes a week during the time a novel was being covered. Each novel was to take 4–6 weeks to complete, with two additional class sessions for a culminating activity. Trainers suggested that one novel be covered in the fall, one in the winter, and one in the spring—with potential weeks off, as needed or desired, between novels.

Control teachers used each school's regular curriculum (business as usual). Students in these classrooms engaged in their usual reading/English language arts curriculum and instructional program (e.g., reading short stories, newspaper and magazine articles, and novels).

Analysis and Results

The final analysis included 90 teachers and a minimum of 2,140 students (89% of the overall baseline sample, 90% of the intervention group, and 88% of the control group). A three-level model of students nested within teachers nested within schools was used to estimate impact. To improve the precision of impact estimates, covariates at Level 1 included students' pretest scores, English language learner status, and special education status; at Level 2 included teacher education and years of teaching experience; and at Level 3 included school poverty and school size. The analysis also explored whether the intervention effect varied for each outcome or was homogeneous across schools. Sensitivity analyses were conducted to test the robustness of the results under different scenarios.

The impact results for the primary research questions indicate that *Thinking Reader* was no more effective than business as usual in improving students' reading vocabulary (effect size of –0.04) or reading comprehension (effect size of 0.03). Results for the ancillary research questions indicate that *Thinking Reader* was also no more effective than business as usual in improving students' use of reading comprehension strategies (effect size of 0.03) or their motivation to read (effect size of –0.03). None of these results are statistically significant. Sensitivity analyses found no changes in the direction or magnitude of the intervention effects.

The data from the classroom observations—conducted twice for each classroom in a subset of classrooms—showed statistically significant differences between the intervention and control conditions on 47 of 57 measured classroom variables, indicating a contrast in the nature of instruction between the intervention and control groups during observations.

Data from the program's electronic records showed that students used the *Thinking Reader* program for considerably fewer minutes per week on average—60 minutes (Book 1), 56 minutes (Book 2), and 42 minutes (Book 3)—than the recommended 110–165 minutes. The average number of weeks per book—8.3 (Book 1), 7.1 (Book 2), and 1.7 (Book 3)—also differed from the recommendation (4–6 weeks). Book completion fell off from the first to the third book, from 74% for Book 1, to 53% for Book 2, to 9% for Book 3. Thus the software was not always used as prescribed by the developers and modeled in the professional development for intervention teachers.

With regard to the exploratory research questions detailed in Chapter 5, we investigated whether the impact of *Thinking Reader* on student outcomes was different for students with different baseline achievement scores or baseline motivation to read, and whether or not impacts for each tertiles were statistically different from zero. The multilevel results revealed that the *Thinking Reader* program had no statistically significant effects on any of the subgroups formed from baseline scores. In other words, these results confirmed the impact findings. Exploratory findings suggest that there is no strong evidence supporting the hypotheses that the *Thinking Reader* program might have differential impact effects on students from different achievement and motivation to read subgroups.

Eleven out of 12 interactions tested across the four exploratory research questions were not statistically significant. For the reading comprehension outcome, we found one statistically significant interaction (5.77, p =.03). This interaction resulted from the fact that in the lowest tertile, intervention students performed 2.15 points higher than control students while in the middle tertile, control students outperformed intervention students by 3.61 points.

However, this statistically significant interaction is difficult to interpret because this finding was not replicated on any other contrast or outcome. Due to the large number of post hoc analyses involved in the exploratory analyses, this statistically significant interaction may be due to chance. Furthermore, when we tested whether the impact effect of the intervention was different from zero in each tertile, we found that the *Thinking Reader* program had no statistically significant effects for any of the baseline achievement or motivation to read tertiles.

Conclusions

This study was the first randomized controlled trial of *Thinking Reader*. The study maintained the integrity of the randomization throughout. The lack of statistically significant positive effects contrasts with the findings of a quasi-experimental study by Dalton et al. (2002). The intent-to-treat analytical approach, which analyzes participants on the basis of how they are randomly assigned, yielded unbiased estimates of program effectiveness as implemented. Program impact should be interpreted as the effect of being assigned to the intervention condition, not necessarily of actually receiving the intervention (Shadish, Cook, & Campbell, 2002). Implementation data show that actual take up of the intervention in the way the developer intended was low. The fact that dosage and usage varied from the developers' recommendations reflects the choices that practitioners make and the challenges they face when implementing an instructional program. The study is limited in that it was not designed to collect more in-depth information on implementation.

The results reported here apply to the implementation of the *Thinking Reader* software after modest professional development, used as a partial substitute for the regular Grade 6 curriculum in a whole-group setting—just one of the ways in which the program can be implemented. Use of a volunteer sample limits the findings to the schools, teachers, and students that participated in the study in Connecticut, Massachusetts, and Rhode Island.

Chapter 1.
Introduction and Study Overview

Improving adolescent literacy is a critical step in improving adolescent academic achievement (Kamil et al., 2008). “Adolescent literacy” commonly refers to the skills that students in Grades 4–12 need in order to successfully learn by reading, as opposed to learning how to read, which is emphasized in earlier grades (Kamil, 2003; Kamil et al., 2008; National Governors Association, 2005). Recent policy reports emphasize the need to build students’ reading vocabulary and comprehension skills to meet the increased literacy demands that begin in Grade 4 (for example, Carnegie Council on Advancing Adolescent Literacy, 2010; Meltzer, Smith, & Clark, 2001). Experts drafting the Common Core Standards for English Language Arts emphasize that students must show a steadily increasing ability to discern more from text to become successful readers (National Governors Association & Council of Chief State School Officers, 2010).

On the 2009 National Assessment of Educational Progress (NAEP), 68% of Grade 8 students read below the proficient level—84% among low-income students (National Center for Education Statistics, 2009a). The three states in the Northeast Region that participated in the current study (Connecticut, Massachusetts, and Rhode Island) reflect this national problem (National Center for Education Statistics, 2009a; Table 1.1).

Table 1.1 Percentage of Students Testing Below the Proficient Level on the 2009 National Assessment of Educational Progress Grade 8 Reading Assessment and Percentage Eligible for Free or Reduced-Price Lunch in Connecticut, Massachusetts, Rhode Island, and Nationally

State	Below proficient on NAEP—Overall[a]	Eligible for free or reduced-price lunch
Connecticut	58	82
Massachusetts	57	81
Rhode Island	72	86
National	68	84

[a]The number of students assessed (rounded to the nearest hundred by the source) was 2,900 in Connecticut, 4,000 in Massachusetts, 2,800 in Rhode Island, and 169,100 nationally. NAEP uses weighted sampling to generalize findings to the larger population.
Source: National Center for Education Statistics (2009a); National Center for Education Statistics (2009b).

Difficulties with comprehension and self-monitoring of understanding are common among middle and high school readers (RAND Reading Study Group, 2002). Because of these comprehension problems, many students struggle to learn from advanced texts and may disengage from reading (National Governors Association, 2005; RAND Reading Study Group, 2002). Certain teaching and learning approaches have been shown empirically to improve middle and high school students’ reading achievement, including direct teaching of vocabulary and teaching a combination of comprehension strategies (Kamil, 2003). The current evaluation examines the effectiveness of *Thinking Reader*®, which uses a software program to directly teach vocabulary and evidence-based reading comprehension strategies, supported by specific teacher practices (Tom Snyder Productions 2006a, b).

The evaluation was conducted by Regional Educational Laboratory Northeast and Islands (REL-NEI) partners: the American Institutes for Research® (AIR®), the Center for Applied Special

Technology (CAST), Tom Snyder Productions®, and Sun Associates. AIR led the evaluation—recruiting schools, testing students, and analyzing data. Staff from CAST, a nonprofit research and development organization and developer of *Thinking Reader*, oversaw implementation and conducted teacher training along with staff from Tom Snyder Productions, the software distributor. To minimize a potential source of bias, the developer and distributor were not involved in random assignment, data collection, or data analysis; CAST staff served as instructional coaches, and Tom Snyder Productions provided the software and technical support. Sun Associates, an education evaluation organization, conducted classroom observations.

Description of *Thinking Reader*

Key features. *Thinking Reader* is a software program for improving the reading vocabulary and comprehension of students in Grades 5–8. CAST began developing the program in 2002, and Tom Snyder Productions/Scholastic began marketing it in 2004. With *Thinking Reader*, students read novels on computers and respond to prompts that support seven comprehension strategies (explained in detail below). *Thinking Reader* offers a choice of nine novels with a range of difficulty appropriate for middle school readers.[5]

The program embodies an approach to reading instruction known as reciprocal teaching (Palincsar, 1982; Palincsar & Brown, 1984; Brown & Palincsar, 1985), which requires teachers to model comprehension strategies and support students' efforts to recall and employ those strategies in their reading. In *Thinking Reader*, animated coaches and peers play the instructional role, modeling the seven comprehension strategies and prompting students to use them at appropriate points in the text.

Teachers are expected to play an active role—monitoring and individualizing instruction, assessing students' progress, and customizing the support provided to each student. The program has five levels of support, which vary the representation of the strategy prompt, students' response options, and the availability of the animated coaches. Students can progress from responding to highly structured strategy prompts (Level 1) to independently selecting strategies (Level 5). According to the distributor, when the program is used at lower levels, supports focus mostly on helping students understand text using direct hints on to how to answer prompts (e.g., sentence starters). Higher levels focus more generally on how to use and apply comprehension strategies and include more questions requiring open-ended responses (Tom Snyder Productions, 2006). Representative screenshots are presented in Appendix A (see Exhibits A.1–A.7). According to the distributor, teachers are expected to adjust the level of support to individualize instruction for each student based on their progress and development. As students gain more skills and confidence as readers, they should become more independent and use the program supports less (Tom Snyder Productions, 2006).

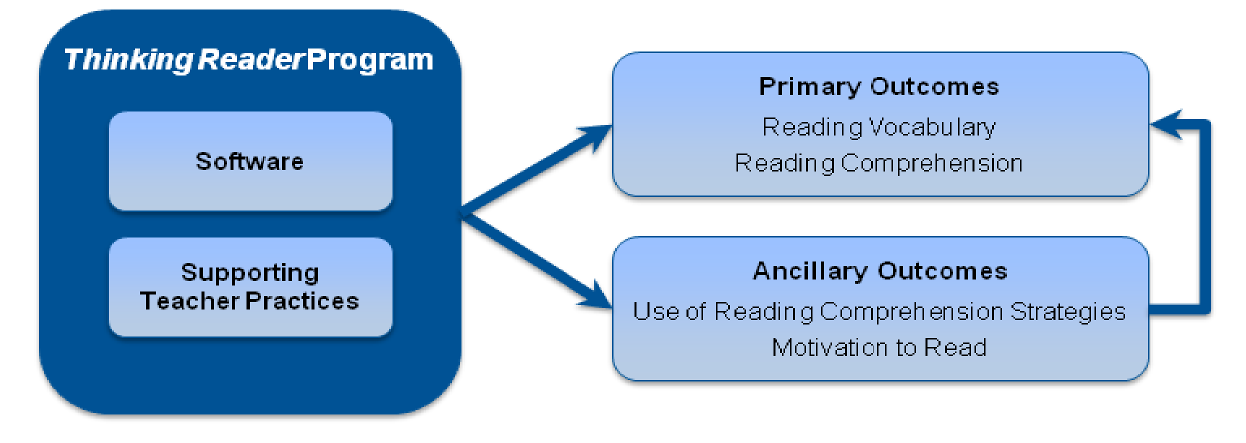
Teachers are encouraged to use *Thinking Reader* in a three-phase instructional routine based on the developer's guidelines and commonly recommended reading pedagogy of before, during, and

[5] Participating schools received unlimited site licenses for four *Thinking Reader* titles at no cost. For reference, the following price information, as of spring 2010, was provided by the distributor: Prices per novel range from $250 to $2,200, depending on the type of license and number of different novels ordered. Licenses for *Thinking Reader* are available for individual computers, for bundles of 10 licenses, and as an unlimited site license. Discounts are given on orders for multiple *Thinking Reader* novels and for multiple schools within a district.

after reading activities (Duke & Pearson, 2002). The routine calls for the teacher to introduce the program offline through activities, such as modeling a strategy or asking for a summary from the previous session. In the second phase, students read the *Thinking Reader* novel on the computer while the teacher observes them, reviews their work, and meets with students about the quality of their responses. The third phase involves additional offline teacher–student interaction, discussing the book or completing an activity to illustrate understanding.

Figure 1.1 is a logic model that depicts how the *Thinking Reader* program of instruction—a combination of software use and supporting teacher practices—is intended to improve student outcomes. Supporting teacher practices include student–teacher conferencing, comprehension strategy instruction, and implementing the three-phase instructional routine. The *Thinking Reader* program of instruction is intended to improve two related measures of students' reading achievement: students' reading vocabulary and comprehension skills. In addition, the program is intended to directly increase students' knowledge and use of seven reading comprehension strategies (described in detail below) and increase students' motivation to read. The two ancillary outcomes, use of reading comprehension strategies and motivation to read, may also be important precursors or supplements to improved reading outcomes in vocabulary and comprehension.

Figure 1.1 Logic Model for *Thinking Reader* Software Program



Source: Study team compilation.

Additional components. Engaged readers have been described as readers who are motivated to read and are socially interactive in their approach to reading (Guthrie & Wigfield, 2000). The influential policy report *Reading Next* identified motivation and self-directed learning as being important to adolescent reading achievement (Biancarosa & Snow, 2004). *Thinking Reader* aims to address engagement, motivation, and self-directed learning in at least four ways. First, scaffolding supports in the program alter the level of challenge for students, providing students with choices and control over access to supports and options for their response. Second, self-assessment allows students to reflect on their progress and set goals. Third, the program sets up students to interact with one another around a text. *Thinking Reader* is intended to be integrated with classroom discussion and peer interaction. Last, the novels in *Thinking Reader* are selected to be age appropriate to middle school students and relevant to their real-life experiences.

Thinking Reader has features that can be tailored to match a set of principles called universal design for learning. Although this set of principles has not been rigorously tested, it is based on the premise that teachers should individualize instruction and use a flexible learning environment to meet students' needs (Rose & Meyer, 2002). In addition to the levels of support, *Thinking Reader* incorporates ways to adjust the text size, text contrast, and screen color; a text-to-speech tool so text can be read aloud; a highlighter function to track text as it is read aloud; the option of a human recorded narration or a synthetic voice whose speed and pitch can be adjusted; and an option for recording rather than typing answers.

Literature Review

Evidence base for* Thinking Reader*. The research relevant to this evaluation includes one rigorous study on the *Thinking Reader* program (Dalton, Pisha, Eagleton, Coyne, & Deysher, 2002) and studies that examined the effectiveness of the approaches to reading instruction used by *Thinking Reader* (e.g., Deno, 2003; Gambrell & Jawitz, 1993; Palincsar & Brown, 1984; Rosenshine, Meister, & Chapman, 1996; What Works Clearinghouse, 2010). A review of this body of knowledge showed enough positive (if not always highly rigorous) evidence to warrant an evaluation of the *Thinking Reader* software with a randomized controlled trial design.

The direct evidence base for *Thinking Reader* is limited. The current study is the first randomized field trial to test its effectiveness. A previous quasi-experimental study evaluated a pilot version of the software among a sample of 102 struggling middle school readers (average age, 12.5 years). After controlling for pretest scores on reading vocabulary and comprehension, the study found that students in the *Thinking Reader* intervention demonstrated significantly greater gains in comprehension than the comparison group on the Gates-MacGinitie Reading Tests (GMRT) (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 1999; Dalton et al., 2002). The effect size was 0.29, approximately half a grade level of reading achievement gain.

That evaluation differed from the current work in several ways. First, a program of traditional (i.e., noncomputer-based) reciprocal teaching methods was delivered to the control group.[6] Second, because the software was a pilot program, it had fewer features—less scaffolding and fewer assessments. Finally, the analysis focused only on struggling readers who had performed at the 25th percentile or lower on the pretest. In that study, 80% of the students were in special education and 79% of the teachers taught the intervention in a special education class of 12–14 students (B. Dalton, personal communication, December 7, 2009)—a smaller setting than 83% of the classes in the current study.

Evidence base for instructional features used by* Thinking Reader*. Evaluating *Thinking Reader* in an experimental study is warranted based on the statistically significant findings of one quasi-experimental study of the program and its use of strategies and instructional approaches with empirical evidence. In addition, the high priority given to improving adolescent literacy by stakeholders in the northeast states when the REL-NEI study was in its start-up phase and the call for more research on educational technology in general (Moran, Ferdig, Pearson, Wardrop, & Blomeyer, 2008) also warranted conducting a randomized controlled trial on *Thinking Reader*.

[6] This was a deliberate decision to make the intervention and control conditions as comparable as possible so that the main difference between the groups was whether *Thinking Reader* was used.

Comprehension strategies. *Thinking Reader* aims to teach comprehension strategies through reciprocal teaching (developed by Palincsar 1982; also Brown & Palincsar, 1985; Palincsar & Brown, 1984), which involves guided practice in applying four strategies to reading comprehension—summarizing, questioning, clarifying, and predicting. In an instructional dialogue about text between a teacher and students, the teacher initially takes the lead—demonstrating, modeling, and providing feedback on strategy use. As students gain competence in using the strategies, the teacher does less modeling and students apply the strategies on their own (Au & Raphael, 1998; Duke & Pearson, 2002; Pearson & Gallagher, 1983). With scaffolded learning, the teacher adjusts instruction to meet the needs of the student based on the demands of the task, gradually releasing responsibility to the student (Vygotsky, 1978; Wood, Bruner, & Ross, 1976). As McKenna (1998) suggests, this type of guidance and scaffolding customized to the student can be offered by talking book computer software.

Rigorous studies indicate that reciprocal teaching strategies can show positive results for reading (Brown & Palincsar, 1989; Palincsar & Brown, 1984, 1989; Rosenshine & Meister, 1994; What Works Clearinghouse, 2010). In two early experiments using groups of students randomly assigned to a reciprocal teaching condition or a control condition, Palincsar and Brown (1984) reported that students receiving reciprocal teaching strategy instruction showed statistically significant positive effects on comprehension scores on daily assessments. Research syntheses have shown mixed results for the strategies. In a meta-analysis of experimental or quasi-experimental studies examining the effect of reciprocal teaching, Rosenshine and Meister (1994) found an average effect size of 0.20 (range: –0.55 to 0.77) using standardized tests of reading comprehension as outcome measures. The What Works Clearinghouse (2010) reviewed six studies on reciprocal teaching and found mixed results for comprehension outcomes. Across the studies, participants included 316 students from Grades 4 to 12 who experienced reciprocal teaching interventions at their schools for periods of time ranging from 6 to 30 days of intervention. The group of studies included both teacher- and researcher-implemented sessions. The What Works Clearinghouse rated reciprocal teaching with improvement indices in comprehension from –23 to 42 percentile points, with an average of 6 percentile points.

Thinking Reader expands on the reciprocal teaching approach by providing guided practice in seven reading comprehension strategies (Table 1.2). The first four are the strategies commonly used in reciprocal teaching. In summarizing, students explain information gleaned from a piece of text. Questioning involves identifying key information about a concept in the text by asking questions. Clarifying involves asking about concepts or words that are unclear. Predicting entails using what is known from the story to hypothesize what will happen next. The three additional strategies are visualizing, feeling, and reflecting. Visualizing involves instructing students to create mental images of what they are reading. The feeling strategy aims to connect students emotionally with what they are reading. The reflecting strategy involves students in monitoring or reflecting on their progress as readers, helping them to become more strategic and efficient in their approach to reading.

The reciprocal teaching literature generally looks at the four strategies applied in combination; however, two of the strategies—summarizing and questioning—have been studied individually. The National Reading Panel (National Institute of Child Health and Human Development, 2000) conducted a comprehensive review of experimental and quasi-experimental reading-related studies. Eighteen studies examined summarizing, 72% of them for students in Grades 5 and 6. Of the 18 studies, 11 reported positive effects on the quality of written summaries, with improved recall and question answering. Only two studies looked at effects on standardized tests. Overall, the comprehensive review concluded that the summarization strategy led to improvement in students' memory for text and their ability to identify main ideas. The questioning strategy had the strongest body of evidence, with 27 studies, in the National Reading Panel report. Reporting on 26 of these studies (conducted in Grade 3 through college) in a meta-analysis of quasi-experimental and randomized controlled trial studies, Rosenshine , Meister, and Chapman (1996) found that 17 studies used questioning as the single intervention.[7] Of those, 7 used a standardized outcome measure (mean effect size of 0.25, with a range of –0.25 to 0.70); 11 used a research-specific outcome measure (mean effect size of 0.79, with a range of 0.00 to 1.37). Thus, summarizing and questioning have more evidence of transfer to researcher-generated tests (near transfer) than to standardized tests (far transfer). Clarifying and predicting have not been studied rigorously in isolation.

No rigorous studies have examined the feeling or reflecting strategies. However, Rosenblatt (1994) theorized, for example, that if students think about how they are feeling about a story, its characters, and the plot, they will be more engaged. The National Reading Panel (National Institute of Child Health and Human Development, 2000) reported on seven experimental studies (for Grades 2–8) that examined visualizing and concluded that it can help students improve their memory for the text. Joffe, Cain, and Marić (2007) theorized that making a mental image of what is read can help with memory because information is being encoded into two separate mental systems: one verbal and one non-verbal. Studies have shown that, after being trained to create mental pictures while reading, students in intervention groups outperform those in control groups on story-related questions (Gambrell & Jawitz, 1993; Oakhill & Patel, 1991; Pressley, 1976).

Table 1.2 Key Reading Comprehension Strategies Used in *Thinking Reader*

Strategy	Definition	Purpose	General example	*Thinking Reader* example
Summarizing	Students retell the main points of what they have read.	Helps students remember and understand what they are reading.	After students read a book, the teacher asks them to write a short summary.	The software prompts students to select the best summary.
Questioning	Students pose a question about something that is important to learn and remember from what they have read.	Requires students to identify key information.	While reading a book as a class, students are asked to write questions about important parts of the chapter and then to answer the questions.	The software prompts students to select the most important question from a list and then to answer it.
Clarifying	Students ask about something they do not understand in what they have read.	Helps students resolve confusion.	While reading a book as a class, students ask questions about something that is confusing.	The software asks students to identify something that is confusing in the text.

[7] Nine studies used questioning in the context of an intervention that used two or four of the reciprocal teaching strategies.

Strategy	Definition	Purpose	General example	*Thinking Reader* example
Predicting	Students use what they have read to predict what will happen next in the novel.	Helps students analyze text content.	After reading a few chapters, students are asked to predict what will happen next in the novel.	The software asks students to make a prediction about what is going to happen next in the text.
Visualizing	Students use the information they have read to create a picture in their minds.	Gives students a way to remember what they read.	After reading a chapter, students are asked to create a mental picture of a key occurrence in the chapter.	The software asks students to look at highlighted words in the text and visualize what is happening. Then, the software prompts students to write or record what they see in their minds.
Feeling	Students make explicit their feelings about the story, a character, or the plot.	Helps students connect with the text.	The teacher asks students to describe how they feel about a particular event in the story.	The software prompts students to describe their feelings or how they would feel if they were one of the characters.
Reflecting	Students track their progress as a reader, monitoring their comprehension.	Helps students decide which reading strategies work for them and why.	Teachers ask students to write self-evaluations of their own progress as readers in journals.	The software prompts students to describe and evaluate their progress.

Note: The first four strategies are part of the reciprocal instruction model.
Source: Duke and Pearson (2002); Harris and Hodges (1995); Tom Snyder Productions (2006b).

Vocabulary instruction. In addition to comprehension strategy instruction, *Thinking Reader* aims to build students' vocabulary by exposing them to new words in the context of authentic literature (Tom Snyder Productions, 2006b). The program provides the opportunity for explicit instruction when teachers introduce and discuss vocabulary pertinent to the story. Target words are highlighted on the screen. Students can click on these words, which are hyperlinked to a glossary. Definitions are also available in Spanish.

Research suggests that word knowledge is correlated with measures of reading comprehension (Pearson, Hiebert, & Kamil, 2007); however, establishing the causal link between increasing vocabulary and increasing reading ability has been difficult (Stanovich, 2000). Nevertheless, students with good comprehension skills know many more words than struggling readers and are better able to use contextual cues and word structure knowledge to determine word meaning (Nagy & Scott, 2004). In their review, Blachowicz, Fisher, Ogle, and Watts-Taffe (2006) suggest that exposure to new and sophisticated words, including ones presented in a read-aloud environment, help to build a "word consciousness" for students. A meta-analysis of vocabulary interventions (conducted at Grade 2 through college) by Stahl and Fairbanks (1986) suggests that vocabulary interventions can have a statistically significant positive effect on reading comprehension outcomes (mean effect size of 0.97 for 41 analyses of comprehension measures that contain words targeted in the intervention; mean effect size of 0.30 for 15 analyses of standardized tests not designed to contain targeted words). These interventions also had a statistically significant positive effect on measures of vocabulary knowledge (effect size of 0.20 for 17 analyses using global vocabulary measures). The most effective vocabulary interventions have been "mixed-methods" that provided both definitions and contextual information for words (Blachowicz , Fisher, Ogle, & Watts-Taffe, 2006). The *Thinking Reader* program uses this approach.

In addition to support for direct vocabulary instruction in general, the National Reading Panel (National Institute of Child Health and Human Development, 2000) concluded that a small literature base supports the use of computers for vocabulary instruction. For example, Reinking and Rickman (1990) tested whether providing Grade 6 students with definitions on a computer screen had an impact on their vocabulary and comprehension outcomes. Students randomly assigned to interventions in which they read passages on a computer and received optional or mandatory computer vocabulary assistance outperformed comparison groups who read excerpts on paper with assistance from either a dictionary or glossary. Statistically significant differences were found between the two groups for experimenter-devised vocabulary and comprehension tests. Vocabulary assistance from a computer may be particularly helpful as students read longer and more difficult material (Greenlee-Moore & Smith, 1996).

Assessment and feedback. *Thinking Reader* is designed to support teachers' assessment of students and students' self-evaluation. Computerized comprehension quizzes provide teachers and students with information to monitor students' understanding. Responses to strategy prompts are collected in an electronic worklog for review during student–teacher conferences. Teachers can also provide students with electronic feedback and requests for revisions to their strategy prompts. Responding to certain strategy prompts gives students opportunities to reflect on their progress as readers.

The policy report *Reading Next* suggests that teachers should informally assess their students regularly to understand how adolescent readers are progressing (Biancarosa & Snow, 2004). Ongoing classroom-based assessments are intended to give teachers immediate feedback about student performance and to help them adjust the instruction that students are receiving, as needed (Torgesen & Miller, 2009; Winograd, Flores-Dueñes, & Arrington, 2003). Frequent monitoring of students' progress is intended to prompt teachers to intervene as soon as a learning difficulty is encountered and to help students learn to self-assess and set appropriate goals (Cioffi & Carney, 1997; Lipson & Wixson, 2003). Deno's (2003) review of both experimental and correlational studies supports teachers' use of progress monitoring—associating it with high levels of student achievement, improved teacher decision making, and increases in students' awareness of their own performance (see also Fuchs & Fuchs, 1999 for an earlier review of monitoring student progress).

Learner control. As Proctor, Dalton, and Grisham (2007) point out, a key issue in designing digital texts with embedded learning supports is learner control and whether computer support should be "pushed" to the student, which help to ensure that features are experienced. The alternative is to have students be more self-directed, allowing support features to be "pulled" by the student when he or she judges the support to be needed. This design question centers on whether students in a more learner-controlled environment will appropriately access supports. In their review, Shin, Schallert, and Savenye (1994) describe several studies that show it is only students with higher skill and inquiry levels that do better and work efficiently using programs with more learner control. Indeed, some researchers claim that the students who would benefit the most from computer support are least likely to access supports appropriately (e.g., Anderson-Inman, Horney, Chen, & Lewin, 1994). Struggling readers may find a high number of features and options in a program distracting, doing things like accessing definitions for words they already know (Boone & Higgins, 1993; Dalton & Strangmann, 2006). In one study, high- and low-skilled fifth- and sixth-graders read passages in both printed and computer presentations. Students who were not the highest achieving benefitted more from computer-controlled environments, perhaps because their metacognitive skills were not as developed (Reinking, 1988). One option is to give students control of their environment, thus reinforcing self-direction, while having the computer advise them on their choices. In a review of studies, Shin et al. (1994) found that students in learner-controlled environments with computer guidance receive higher scores than student in control conditions. *Thinking Reader* uses a combined "push–pull" approach. Some features—such as text-to-speech, glossary hyperlinks, and hints by animated coaches—are accessed only when students choose to use them. Other features, such as explanation of and prompts to use comprehension strategies and periodic comprehension checks, are built so that students automatically experience them.

Research Questions

This evaluation of the impact of *Thinking Reader* use by Grade 6 students focused on two confirmatory research questions about the effect of the program on two measures of students' reading achievement:

1. What is the effect of *Thinking Reader* on students' reading vocabulary?

2. What is the effect of *Thinking Reader* on students' reading comprehension?

A statistically significant impact on either outcome measure would be a signal of the program's success.

The study also examined whether *Thinking Reader* has an effect on two ancillary, but important, measures of students' approaches to reading:

1. What is the effect of *Thinking Reader* on students' use of reading comprehension strategies?

2. What is the effect of *Thinking Reader* on students' motivation to read?

The answers to these questions provide information that may be useful to educators who consider these factors to be important precursors or supplements to improved achievement. But without a direct measurable effect on reading achievement (vocabulary, comprehension), such effects would be insufficient to judge the program's effectiveness. The impact results are presented in Chapter 4.

We also specified a set of four exploratory research questions that examined whether the impact of the *Thinking Reader* intervention on students' reading achievement varied across different subgroups of students. The subgroups were formed on the basis of baseline achievement and motivation to read measures:

1. Does the effect of *Thinking Reader* on students' reading vocabulary vary according to their baseline reading vocabulary scores?

2. Does the effect of *Thinking Reader* on students' reading comprehension vary according to their baseline reading comprehension scores?

3. Does the effect of *Thinking Reader* on students' reading vocabulary vary according to their baseline reading motivation scores?

4. Does the effect of *Thinking Reader* on students' reading comprehension vary according to their baseline reading motivation scores?

The outcomes of interest for these exploratory research questions are the vocabulary and comprehension subtests of the GMRT that served as the achievement measures for the primary research questions. The exploratory results are presented in Chapter 5.

Study Design

The study, a multisite cluster randomized controlled trial, randomly assigned Grade 6 teachers to intervention or control groups within schools. The study involved reading/English language arts teachers in 32 elementary and middle schools in 16 districts in Connecticut, Massachusetts, and Rhode Island. The study sample consisted of 92 teachers and 2,407 students. Teachers and their intact classes were randomly assigned to one of the two groups after students had been scheduled into classes using the typical school procedures. Teachers assigned to the control group were to use the regular school curriculum (business as usual). A well-conducted randomized controlled trial design that compares an intervention group with a business-as-usual control group generates statistically unbiased estimates of the effects of an intervention on the outcomes of interest (Shadish, Cook, & Campbell, 2002).

The primary outcomes of interest—reading vocabulary and comprehension—were measured using two subtests of the standardized GMRTs (MacGinitie et al., 1999). Results for two additional, ancillary outcomes—students' use of comprehension strategies and students' motivation to read—were measured using two existing self-report instruments: the Metacognitive Awareness of Reading Strategies Inventory (MARSI) (Mokhtari & Reichard, 2002) and the Motivation for Reading Questionnaire (MRQ) (Wigfield & Guthrie, 1997). Data from all measures given at the beginning and end of the year were used to estimate the impact of the intervention. Sensitivity analyses were conducted, and usage data from the software were analyzed to examine implementation fidelity. An existing instrument was modified for conducting structured classroom observations for a subset of classrooms to check fidelity of implementation and to explore possible differentiation between intervention and control classrooms in classroom instruction.

Intervention group. Schools received the three *Thinking Reader* digital novels that they selected during the 2008–09 academic year. Intervention teachers were asked to participate in professional development (two, 6-hour group sessions and three individual coaching sessions totaling 7.5–8.5 hours) to learn the *Thinking Reader* software. Intervention teachers were trained to use the three-phase instructional routine. They were asked to incorporate *Thinking Reader* into their regular English language arts or reading instruction for 110–165 minutes a week while a novel was being covered. Teachers were told that each novel should take 4–6 weeks to complete, with two additional class sessions for the culminating end-of-novel activity. Trainers also suggested that one novel be covered in the fall, one in the winter, and one in the spring—with weeks off, as needed or desired, between novels.

Control group. Classrooms in the control group used their schools' regular curriculum (business as usual). Students in these classrooms engaged in the regular activities of their usual English language arts literacy curriculum and instructional program (e.g., reading short stories, newspaper and magazine articles, and novels).

Organization of This Report

Chapter 2 describes the study methodology. Chapter 3 provides an overview of implementation of the *Thinking Reader* intervention, including data from computer worklogs and classroom observations. Chapter 4 describes the results of the impact analyses. Chapter 5 summarizes the study's findings. Appendices provide detailed information about the *Thinking Reader* program, the technical approach to the study, and the analytical decisions made.

Chapter 2.
Study Methodology

This chapter describes the approach to securing and randomizing the sample, collecting data, and the administering measures during the study. It also presents characteristics of the study sample and details of the statistical analyses.

Recruitment

The target population for this study was Grade 6 teachers and students in high-need schools in Connecticut, Massachusetts, and Rhode Island that had two or more reading/English language arts teachers in Grade 6 during the 2008–09 school year. Grade 6 was selected because the transition from elementary to middle school is the time when reading for content-area comprehension becomes increasingly important for academic success. Eligible schools had to have the technology to implement the intervention in a whole-group setting, with a computer for each student.

High-need schools were targeted because of the link between high economic need and low reading achievement. For example, students eligible for free or reduced-price lunch have scored an average of 27–33 scale score points (7–13%) below their counterparts on each of the last six National Assessment of Educational Progress tests (National Center for Education Statistics, 2009a). Additionally, each of the participating states has a statistically significant negative correlation between school-level free or reduced-price lunch rates and reading achievement, indicating that a higher rate of poverty is associated with lower achievement: –0.87 in Connecticut (*p* value of .00),[8] –0.84 in Massachusetts (*p* value of .00),[9] and –0.89 in Rhode Island (*p* value of .00).[10]

High-need was defined as having more than 33% of students eligible for free or reduced-price lunch (based on the 2006–07 school data available at the time of recruitment in spring 2008). The proportion of schools that enrolled Grade 6 students and met this criterion varied slightly: 41% in Connecticut, 44% in Massachusetts, and 46% in Rhode Island. The cut point of greater than 33% was used to ensure a sufficiently large sampling frame of schools that could be considered economically disadvantaged.

Recruitment activities built on the Education Development Center's (EDC) longstanding relationships with education leaders in technology and reading at state and local education agencies. The American Institutes for Research (AIR) worked with EDC's state liaisons in

[8] Correlation between the 2006–07 school-level percentage of students eligible for free or reduced-priced lunch and the percentage of students meeting the state goal on the Grade 6 reading section of the Connecticut Mastery Test.

[9] Correlation between the 2006–07 school-level percentage of students eligible for free or reduced-priced lunch and the percentage of students scoring at or above proficient on the English language arts section of the Massachusetts Comprehensive Assessment System.

[10] Correlation between the October 2007 school-level percentage of students eligible for free or reduced-priced lunch and the percentage of students scoring at or above proficient on the reading section of the 2006–07 New England Common Assessment Program for all schools enrolling Grade 6 students in October 2007.

Connecticut, Massachusetts, and Rhode Island to identify key contacts and set up meetings with school, district, or state leaders to elicit interest in the study.

Recruitment began in spring 2008. School data for 2006–07 from state departments of education were used to identify schools that met the free or reduced-price lunch criterion. Because the study required at least two Grade 6 reading/English language arts teachers in each building, schools with fewer than 48 students enrolled in Grade 6 were excluded. The tally of eligible schools was 145 in Connecticut, 180 in Massachusetts, and 3 in Rhode Island.[11] Recruitment efforts relied on making eligible schools aware of the study and the intervention. If they were interested, further communication was pursued. Information packets were mailed to principals, reading/English language arts teachers, and literacy coaches at the schools. For those schools expressing interest, e-mail exchanges and phones calls were used to provide more information. Not every school communicated back to the study team during this phase. Schools that communicated they were not interested most often cited that they were too busy with other curricular initiatives. For those schools that were still eligible and interested, meetings were arranged to provide more information about the study, answer questions, and determine the feasibility of doing the study at the school.

All eligible schools received equal priority for inclusion in the study. All eligible schools were accepted that were interested in participating, would have at least two reading/English language arts teachers during the 2008–09 school year, and had the required technological infrastructure.[12] Of the 328 schools initially contacted, 32 schools participated. The number of schools that participated compared to the number of schools that were contacted is low (approximately 10%); another Regional Educational Laboratory randomized controlled trial also experienced low recruitment rates (approximately 1%) (Wijekumar, Hitchcock, Turner, Lei, & Peck, 2009). In the current study, 92 of the 98 Grade 6 reading/English language arts teachers at participating schools agreed to take part in the *Thinking Reader* study (Table 2.1).

Table 2.1 Numbers of Teachers in Study, by State

State	Teachers
Connecticut	45
Massachusetts	42
Rhode Island	5
Total	92

Source: Study administrative records.

[11] The decision to recruit in Rhode Island came close to the end of the 2007–08 school year. Although 29 Rhode Island schools were identified as potentially eligible, recruitment materials were not mailed to all schools. Instead, individual schools were targeted based on identification by the EDC state liaison. In the end, fewer than four Rhode Island schools were identified, responded, met with the study team to confirm eligibility and agreed to participate.

[12] *Thinking Reader* uses server-based software. Tom Snyder Productions recommends the following technical specifications for robust performance: Server: Must run Tom Snyder Server software on Windows 2000/2003, Mac OS X 10.3/10.4, or Netware 6/6.5; Processor: 1 Ghz or better; RAM: 1GB or more; Disk space: 200 MB; Workstations: Windows XP or Mac OS 10.3.4 or later; Processor: 800 Mhz or better; RAM 128 Mb or more; Disk space: 250 Mb per title; Network: 100 Mbit wired network or 802.11 a/g/n wireless network.

Incentives for Participation

Districts and schools agreeing to participate in the study received *Thinking Reader* materials free of charge, a key incentive for participating. Tom Snyder Productions donated the *Thinking Reader* software kits, each of which included an installation disk, three–five hard copies of the novels, a reading strategies wall chart, reading strategies bookmarks for students, a teacher's guide, and a novel-specific discussion guide. Tom Snyder Productions also provided free lifetime technical assistance by phone and e-mail.

In fall 2008, participating schools received three *Thinking Reader* software kits, supporting equipment (headphones for each student and microphones if not built into the computers), and hard copies of the novels. Because personnel at some participating schools expressed the desire that all their students read the same novels, hard copies of the novels were provided to the schools in the same number as there were students in the control group. That way, control group students at all schools had the opportunity to read the novels being read via computer by the treatment group.

In fall 2009, Tom Snyder Productions gave each school another *Thinking Reader* novel (in addition to the three provided in fall 2008). Additional headphones and microphones were given to the schools, in the same number as there were students in the previous year's control group, in order to ensure that enough materials were available to enable all sixth graders to use *Thinking Reader.*

In addition, the 92 intervention teachers were offered free professional development in using *Thinking Reader.* Intervention teachers were encouraged to attend two, day-long workshops and participate in three follow-up coaching visits during 2008–09. Well after the study had ended, control group teachers were invited to attend a day-long workshop in October 2009, after which they could start using the software if they so desired. Additional information about professional development is described in Chapter 3.

Random Assignment

AIR determined eligibility and completed recruitment of all schools prior to conducting random assignment. Random assignment was then conducted by AIR on a rolling basis, starting in late August 2008 and ending in late September 2008, when the school information sheets with the current list of teacher names were returned. To assign teachers to intervention and control conditions, a random number generator was used to create a separate allocation sequence for each school. Random assignment should evenly allocate intervention and control teachers to any preexisting imbalance that might exist (e.g., in classroom size, student achievement level, and demographic characteristics). In each school, teachers and their existing classrooms were randomly assigned to intervention or control conditions; no attempt was made to randomize students to teachers. In schools with an odd number of teachers, classrooms had a slightly higher probability of being assigned to the intervention group.[13] Specifically, in schools with three or

[13] In the 23 schools with an even number of teachers, the teachers had a 50/50 probability of being assigned to intervention or control conditions. In 9 schools with an odd number of teachers, the teachers had a probability between 0.60 and 0.66 (depending whether the school had 5 or 3 teachers for randomization) of being assigned to the intervention group. To examine whether these different randomization procedures affected the impact estimates,

five teachers, we randomly picked two or three teachers, respectively, to go to the intervention group. If the availability of computer labs was limited, then only one teacher was assigned to use the intervention. This occurred in fewer than four instances. Comparisons between students in the treatment and control groups are shown later in this chapter. A total of 92 teachers were eligible and randomized.

Institutional Review and Informed Consent

The study team used various methods to ensure that students and their parents were notified about the study and had a chance to opt out. Two institutional review boards (IRBs) approved the study procedures. Considering a variety of factors, including the potential risk to students, both IRBs agreed that the research team could use an opt-out consent process rather than having to obtain "active" parental consent to include students. Schools were asked to distribute information about the study to parents, who could notify the team if they did not want their children to be in the study. The IRB decision was shared with all participating districts. Most districts allowed the study team to use this type of consent. No students were excluded from the study as a result of this notification process. In any district that required active parental consent, the team used alternative consent forms. Fewer than four districts had this requirement. In those districts, eight students were excluded for lack of parental consent.

To inform students directly that taking the tests and completing the surveys were voluntary, the study team distributed an assent form and read it aloud to all students immediately before baseline and follow-up data collections. Students were told that they were not required to answer questions (complete measures) if they did not want to. Six students refused to complete one or more of the measures but did not refuse to complete testing altogether. At pretest, some students refused eight measures in total. At posttest, some students refused seven measures in total. In addition, the team used a written assent form to inform participating teachers that completing the questionnaire was voluntary. No teachers withheld assent. The student and teacher assent statements are shown in Appendix B1.

Data Collection

This study required data on students, teachers, schools, classroom instruction, and the fidelity of the intervention's implementation. As the evaluator, AIR conducted all data collection; to avoid introducing potential bias, neither the developer nor the distributor were involved. In this section, we describe how and when these data were collected for the study. We also present information on the consent process that we followed before data collection and on response rates to the various data collections.

Data on students. Data on students were gathered from three sources: Study teachers provided student demographic and enrollment information on classroom rosters; students provided

we conducted two sets of sensitivity analyses. In the first analysis, the benchmark impact models were run with only the 23 schools that had an even number of classrooms (dropping 9 schools with an odd number of teachers). In the second analysis, we ran the same impact analyses with all schools but added a dummy covariate at the school-level (or Level 3) that was equal to 1 if the schools had an odd number of teachers and 0 if the school had an even number of teachers. None of the impact estimates presented in the report changed in either analysis. Appendix D1 details these sensitivity analyses and Tables D1.1-D1.6 presents the multilevel model results of these sensitivity analyses.

background characteristics and completed the outcome measures; and states or schools provided student background characteristics in data files from state achievement tests.

Class rosters from teachers. Class rosters were collected from all study teachers to gather information on student demographics for determining eligibility for the tests to be administered (described below) and information on student mobility. Class rosters were collected from study teachers at three times: at the start of the school year, before the baseline data collection; at midyear; and near the end of the school year, before follow-up data collection. On the rosters, teachers indicated which students were enrolled in the class and whether each student was considered an English language learner (ELL) or had an individualized education program (IEP). This information was obtained for 98% of students.[14] Students who would normally be excluded from the regular state achievement tests (because of very low English language levels or an indication on their special education IEPs) were excluded from the achievement tests. Other ELL students and students with IEPs who required accommodations on state reading achievement tests were also excluded from testing unless the accommodations could be fulfilled by field staff or staff at the school.[15] All remaining enrolled students were considered eligible for testing.

Information from the class rosters was also used to document student movement across conditions (from intervention to control classrooms and vice-versa) and attrition from the study (when students left study classrooms, schools, or districts and could not be part of the follow-up data collection).

Measures from students. The study team visited each school twice during the study year to collect outcome measures to address the primary and ancillary research questions. The baseline data collection (pretesting) took place from September 8, 2008, to October 10, 2008 after random assignment.[16] For baseline testing, the data collectors were unaware of the condition to which any teacher had been assigned. The follow-up data collection took place from May 4, 2009, to June 5, 2009. At follow-up, data collectors were aware of each teacher's condition. To economize, the same data collectors who collected student data also collected data (worklogs) from the *Thinking Reader* program itself. Therefore, the data collectors needed to know the condition of teachers in order to know which teachers would have the *Thinking Reader* computer records.

The four paper-and-pencil instruments described below were administered each time in English. Data collection time per session was 120 minutes. The measures are described briefly in this section; information about psychometric properties is presented in Appendix B2. Before testing, students were asked to fill out background information on their Gates-MacGinitie Reading Tests (GMRT) answer sheet. Students filled in their gender and birthdate at both pretest and posttest; at posttest, they filled in race/ethnicity. This information helped to better describe

[14] For the other 2% of students, we were able to gather information from state data files (see below).

[15] At pretest, 66 students (44 intervention, 22 control) received accommodations; at posttest, 44 students (33 intervention, 11 control) received accommodations.

[16] Collecting pretest data after random assignment can bias the posttest impact estimates; this is a design imperfection. Schochet (2008) explicitly addressed the late pretest problem in randomized control trials of education interventions. He found that for randomized control trials of interventions such as this one, estimates obtained when the late pretests were included are normally preferred to estimates that excluded them or instead included baseline test score data from other sources. The author argued that these results hold when growth in test score impacts do not grow very quickly early in the school year. This partially depends on when and how quickly the intervention is implemented. In our case, the intervention was not implemented until well after the pretest was given.

the sample and verify that random assignment had created equivalent intervention and control groups. Response rates for the background information and for each of the four measures are reported in Table 2.2.

Reading achievement. Two standardized, multiple-choice subtests from the GMRT (MacGinitie et al., 1999), were administered at both pre- and posttest. All GMRT data were analyzed using extended scale scores (transformed raw test scores that put scores on an equal-unit scale so that the data could be used in statistical analyses to measure growth over time). Extended scale scores represent a single, continuous scale that can be used to track growth and identify the location of a specific score relative to a range of achievement. The GMRT is a vertically equated test, suggesting that changes in a student's score over time indicate improvement in the student's competency level.

- *Reading vocabulary.* The GMRT vocabulary subtest, which measures word knowledge, was used to assess vocabulary. Forty-five target words are presented in a brief context that suggests the part of speech. Students are asked to select the word or phrase that most closely approximates the meaning of the test word. Students had the standard 20 minutes to complete this subtest.
- *Reading comprehension.* The GMRT comprehension subtest, a 48-item subtest that measures a student's ability to read and understand different types of prose and to understand the meaning of words in context, was used to measure reading comprehension. Students are asked a variety of questions. Some of them tap into literal understanding of the text, while others require students to make inferences or draw conclusions. Students had the standard 35 minutes to complete this subtest.
- *Reading strategies.* The Metacognitive Awareness of Reading Strategies Inventory (MARSI) (Mokhtari & Reichard, 2002) was used to measure the strategies that students use to understand what they read. The self-report survey consists of items that focus on strategies for global analysis of text, strategies for problem-solving when text is too difficult, and the use of outside reference materials or other support strategies (see Box B2.1 in Appendix B2). Students rate items on a scale of 1–5, where 1 means "I never or almost never do this" and 5 means "I always or almost always do this." A score is derived by calculating the mean of the 30 items. According to Mokhtari and Reichard (2002), strategy use may be considered high if the mean is 3.5 or higher; medium if scores are 2.5–3.4; and low if scores are 2.4 or lower. The survey is not a timed measure and takes approximately 15 minutes to complete. The survey was read aloud during administration to help struggling readers complete the instrument.
- *Reading motivation.* The Motivation for Reading Questionnaire (MRQ) (Guthrie & Wigfield, 2000; Wigfield & Guthrie, 1997) was used to measure students' reading motivation. Students rate each of the 52 questionnaire items on a scale of 1–4, where 1 means "Very different from me" and 4 means "A lot like me." An overall mean is calculated from all items. This survey is not a timed measure and typically takes 20 minutes to complete. The questionnaire was read aloud during administration. Box B2.2 in Appendix B2 displays the items that make up the questionnaire.

Data files from state achievement tests. Data files from state achievement tests[17] were used to help fill in missing student demographic data and to verify the ELL and IEP status listed by teachers on class rosters.[18] State personnel in Connecticut and Massachusetts provided data for tests conducted in spring 2008, prior to the study, which contained information on IEP and ELL status; the data from Massachusetts also contained information about gender and ethnicity. School personnel in Rhode Island sent us state test scores directly; however, no additional student data were provided. Details on the missing data filled in using state files are presented in Table 2.2.

Table 2.2 Response Rates on Student Data

	Total (*n* = 2,407)		Intervention (*n* = 1,286)		Control (*n* = 1,121)	
Data	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**
Pretest measures						
Reading vocabulary	2,388	99.2	1,276	99.2	1,112	99.2
Reading comprehension	2,388	99.2	1,280	99.5	1,108	98.8
Reading strategies: MARSI	2,388	99.2	1,278	99.4	1,110	99.0
Reading motivation: MRQ	2,387	99.2	1,275	99.1	1,112	99.2
Posttest measures						
Reading vocabulary	2,156	89.6	1,160	90.2	996	88.9
Reading comprehension	2,149	89.3	1,155	89.8	994	88.7
Reading strategies: MARSI	2,217	92.1	1,188	92.4	1,029	91.8
Reading motivation: MRQ	2,225	92.4	1,192	92.7	1,033	92.2
Demographic variables[a]						
Gender	2,407[b]	100.0	1,286	100.0	1,121	100.0
Age	2,392	100.0	1,279	100.0	1,113	100.0
Race/ethnicity	2,269[c]	94.3	1,216	94.6	1,053	93.9
Individualized education program status	2,407[d]	100.0	1,286	100.0	1,121	100.0
English language learner status	2,407[d]	100.0	1,286	100.0	1,121	100.0

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
[a] Collected by the study team from student pre- and posttest measures and classroom rosters.
[b] Massachusetts state data used to fill in three missing cases.
[c] Massachusetts state data used to fill in 141 missing cases.
[d] Massachusetts and Connecticut state data used to fill in 51 missing cases.
Source: Study administrative records.

Data on teachers. Limited information was collected on teacher backgrounds for assessing the baseline equivalence of intervention and control groups and strengthening the analytic models. A brief teacher questionnaire asked about years of teaching experience, educational attainment, and certifications/endorsements held (see box B2.3 in Appendix B2). The questionnaire took approximately 10 minutes to complete, and all teachers returned it.

Data on schools. To strengthen the analytic models, data on school characteristics were collected from state education agency websites. Data included type of school (elementary or middle), state, enrollment size, and the poverty level and ethnicity of students. The most current

[17] State reading test score data were collected for use in the imputation models, but the data had too many missing cases to be helpful.
[18] ELL and IEP status had to be verified because some data provided by teachers was incomplete or contradictory.

data available at the time of data analysis were from the 2007–08 school year and were collected for 100% of the study schools.

Data on classroom instruction. To document any differences in instruction in intervention and control classrooms and to examine the fidelity of *Thinking Reader* use, structured observations were conducted twice in a sample of study classrooms. An observer visited the classroom of one intervention and one control teacher at each school and one section/class for each teacher, for a total of 64 teachers and 128 observations. If a school had more than two teachers (one intervention, one control), one from each group was picked randomly. If a selected teacher taught multiple sections, one classroom was selected randomly. Sixteen of the 32 schools (50%) had more than two teachers, and 30 of the 92 teachers (32.6%) taught multiple classes.

Observers contacted teachers directly to schedule observation visits between January and June 2009. Observers asked to schedule visits when intervention classrooms would be doing "typical" *Thinking Reader* instruction and when control classrooms would be doing "typical" reading instruction. Because the observers in intervention classrooms saw *Thinking Reader* being used, observers were aware of the condition to which a classroom had been assigned. During the second round of visits, four classrooms had substitute teachers. These data were not included in the final analysis because instruction by a substitute teacher might not represent typical instruction. With 124 valid observations, the response rate was 97%.

Fifty-two (81%) of the first-round observations took place in February and March 2009. Fifty-two (81%) of the second-round observations took place in May and June 2009. An average of 70 days passed between observations (range: 26–107 days). Table 2.3 shows the timing of the observations.

Table 2.3 Classroom Observations by Month, 2009

	First observation		Second observation	
Month	**Number**	**Percent**	**Number**	**Percent**
January–March	54	84	4	6
April–June	10	16	60	94
Total	64	100	64	100

Source: Study administrative records.

The Center for the Improvement of Early Reading Achievement (CIERA) Classroom Observation Scheme (Taylor, 2004; Taylor & Pearson, 2000), a reliable, low-inference tool, was used for the study because it could be easily adapted to document reading instruction. The instrument provides a structured protocol for trained classroom observers to record the presence or absence of different aspects of classroom instruction. The protocol is designed for use in Grades K–6 (Taylor & Pearson, 2000) and was used previously in studies with a low-income sample of students, who, as in the current study, were heterogeneous in linguistic and racial/ethnic background (Bitter, O'Day, Gubbins, & Socias, 2009). Taylor, Pearson, Peterson, and Rodriguez (2004) used the measure in a study of 13 schools that had student populations in which 70–95% of students were eligible for free or reduced-price lunch (mean of 81%), 0–78% were ELLs (mean of 20%), and 48–92% were racial/ethnic minority students (mean of 71%).

The measure addresses seven dimensions of instruction: who in the classroom is providing instruction and working with students, observed instructional groupings, major academic area covered, materials used by the teacher and students, specific literacy activity, interaction style used by the teacher, and expected mode of student response.

Observers take notes on the classroom environment for 4 minutes and then complete the coding scheme in 7-minute segments—observing and taking notes on classroom instruction for 5 minutes, followed by completing on-task counts and ratings for the seven dimensions for 2 minutes—before resuming observation. If students are engaged in different activities or pacing themselves at different rates through an activity, observers are instructed to walk around the class to ensure that they capture what individual students are doing. The protocol calls for observers to record all codes that are relevant for each segment. Thus, each segment's codes represent all activities that are happening across all students even if not all students are engaging in all activities.

Although the CIERA Classroom Observation Scheme fits the needs of the study, modifications were necessary to reflect key aspects of the *Thinking Reader* software. Primary adaptations included adding the "computer" as a prime medium of instruction, the "*Thinking Reader* digital text" as instructional material, and the "*Thinking Reader* worklog" as an activity to represent worklog tasks—such as students reading their worklogs, students reading teacher comments made to the worklogs, and teachers providing feedback to students about their worklog responses. Codes that were less relevant to Grade 6 readers were omitted (such as letter identification, phonics, and phonemic awareness). In addition, examples from the original CIERA coding scheme that were not age-appropriate were edited to reflect what observers were likely to see in Grade 6 classrooms. With the goal of maintaining consistency with the original coding scheme, the study team consulted with one of the authors of the CIERA coding scheme, who provided feedback on the proposed modifications. The final set of codes and definitions—which are presented in Table B3.1 in Appendix B3 and Box B4.1 in Appendix B4—provides an example of narrative notes and related codes for an observation segment.

Training and inter-rater reliability. Three study team members, along with another expert experienced with training observers on the CIERA coding scheme, delivered an initial 1.5-day training to the five-member classroom observation team. The goal was to introduce the observer trainees to each dimension of the CIERA coding scheme so that they would become reliable users. On the first day, the trainer explained each of the seven coding scheme dimensions, and trainees coded practice video clips of authentic reading/English language arts classrooms. Trainees compared their codes with the expert trainer's codes for accuracy, discussed the process and results, and asked clarifying questions. Trainees received a DVD with additional practice clips to complete at home.

On the second day, trainees corrected their homework and practiced using the CIERA coding scheme for *Thinking Reader*. Because videotapes of students using *Thinking Reader* were not available, trainees received 13 screenshots of *Thinking Reader* that highlighted the components of the program.

The computer program has specific options, so trainees were guided explicitly on how to code all the program components. For example, blue-highlighted text on the computer screen indicated that the computer was reading aloud to the student so the observer would code that the student was expected to be listening to the text. If students use one of the program supports to help them answer a comprehension prompt, an animated "peer coach" appears on the screen to indicate that students are receiving modeling, coaching, or scaffolding. After learning the components, the trainees practiced using the CIERA protocol in a live *Thinking Reader* classroom in a middle school that was not part of the study. Trainees (along with trainers) walked around the *Thinking Reader* class, coding what individual students were doing. Trainees then met with trainers to discuss their codes, ask questions, and check codes for accuracy. Additional time was spent reviewing the program components and their codes. Observers received a demonstration version of *Thinking Reader* to further familiarize themselves with the program and accompanying protocol codes.

To increase reliability, classroom observers independently coded a videotaped lesson at the conclusion of training. These codes were reviewed and discussed by phone. Before any field visits, observers coded additional video clips of classroom instruction that did not involve computers or *Thinking Reader* instruction. The video clips had been coded in advance by the study team and the veteran CIERA trainer, who reached consensus on coding decisions so that the videos could be considered "expert-coded." The classroom observers rated clips until every observer attained 80% coding accuracy for each dimension, which was the benchmark rate used in previous work with the CIERA coding scheme (Taylor et al., 2003, 2004). Four of the five observers required three rounds of coding to attain this level of accuracy. Final percentage agreement between expert and observer codes ranged from 80.6% to 100% across the dimensions; pairwise Cohen's kappa values ranged from 0.61 to 1.00. One coder entered a fourth round of practice, which resulted in 100% agreement for all dimensions. Booster training was provided throughout the observation period to help observers differentiate similar codes and clarify codes that proved difficult to apply even though the codebook provided definitions, examples, and information about code nuances.

To ensure continuing reliability, an expert observer read all observation notes and examined the associated codes to assess the consistency of code use. The expert noted some instances of coding that were not consistent with codebook instructions. In these cases, the expert recorded new codes based on the observers' narrative notes. Inter-rater reliability was calculated between the observers' codes and the expert's codes for a random sample of 15% of observations for each observation round (n = 20; kappa = 0.60). Because of the variability between the observers and the expert, a second expert reviewed the codes of the first expert, and inter-rater reliability was high (kappa = 0.98). Because of inconsistencies in how observers used some codes and a high level of agreement between experts, codes from experts were used when there were disagreements between the expert and observers. The same procedures were followed in previous studies using the CIERA coding scheme (Taylor et al., 2003, 2004).

Data on implementation fidelity. Three kinds of data were collected to gauge implementation fidelity. First, data from the *Thinking Reader* electronic reports were downloaded from the computers of intervention teachers. The data showed the degree to which students were exposed to the software (e.g., number of books started and completed and number of days and weeks spent on each book). Second, although the software itself does not collect detailed information about students' use of specific program features, some features could be seen during observations. These features, including use of text-to-speech and worklogs, were built into the coding scheme. Although limited, observation data gives at least some indication of the prevalence of the use of features. Third, the narrative observation notes from *Thinking Reader* classrooms were analyzed to gauge whether observed classrooms followed the three-phase instructional routine of *Thinking Reader* recommended in the first workshop and described in the *Thinking Reader* teacher guide (prereading discussion before students login, student use of the *Thinking Reader* software, and postreading activities after students exit *Thinking Reader*).

Study Sample

The baseline sample consisted of 49 intervention teachers and 43 control teachers (N = 92). In October 2008, a small number of teachers (less than four) experienced circumstances, such as school transfers or layoffs due to budget cuts. The schools distributed these teachers' students among the other teachers. The study continued to track these students, and the impact analyses followed an intent-to-treat approach by maintaining the students' original group assignments and including them in the impact analyses. The robustness of the impact results was tested by running the impact models on a reduced sample that excluded the students who were in these teachers' classrooms (see sensitivity analyses section, later in this chapter).

Following randomization (in September 2008), a small number of intervention teachers (fewer than four) refused to implement the intervention but agreed to let the study collect data for the intent-to-treat analyses. Data from all teachers present at the end of the study were used in the impact and sensitivity analyses. With this sample size and the assumptions made about the correlation between pretest and posttest scores, this study has the statistical power to detect a minimum effect of 0.19–0.24 standard deviations (see Appendix B5).

The student sample is defined as follows:

- Students listed on the classroom rosters at the time of the pretest (September–October 2008).
- Students eligible to be tested (based on roster information; see above).
- Students identified as special needs before the pretest and for whom testing accommodations could be provided without extraordinary effort (e.g., allowing the student to write in the test booklet or offering extended time).
- Students who joined a classroom after pretesting (n = 169) were excluded from the study because of possible bias in how the students might have been assigned to classrooms.

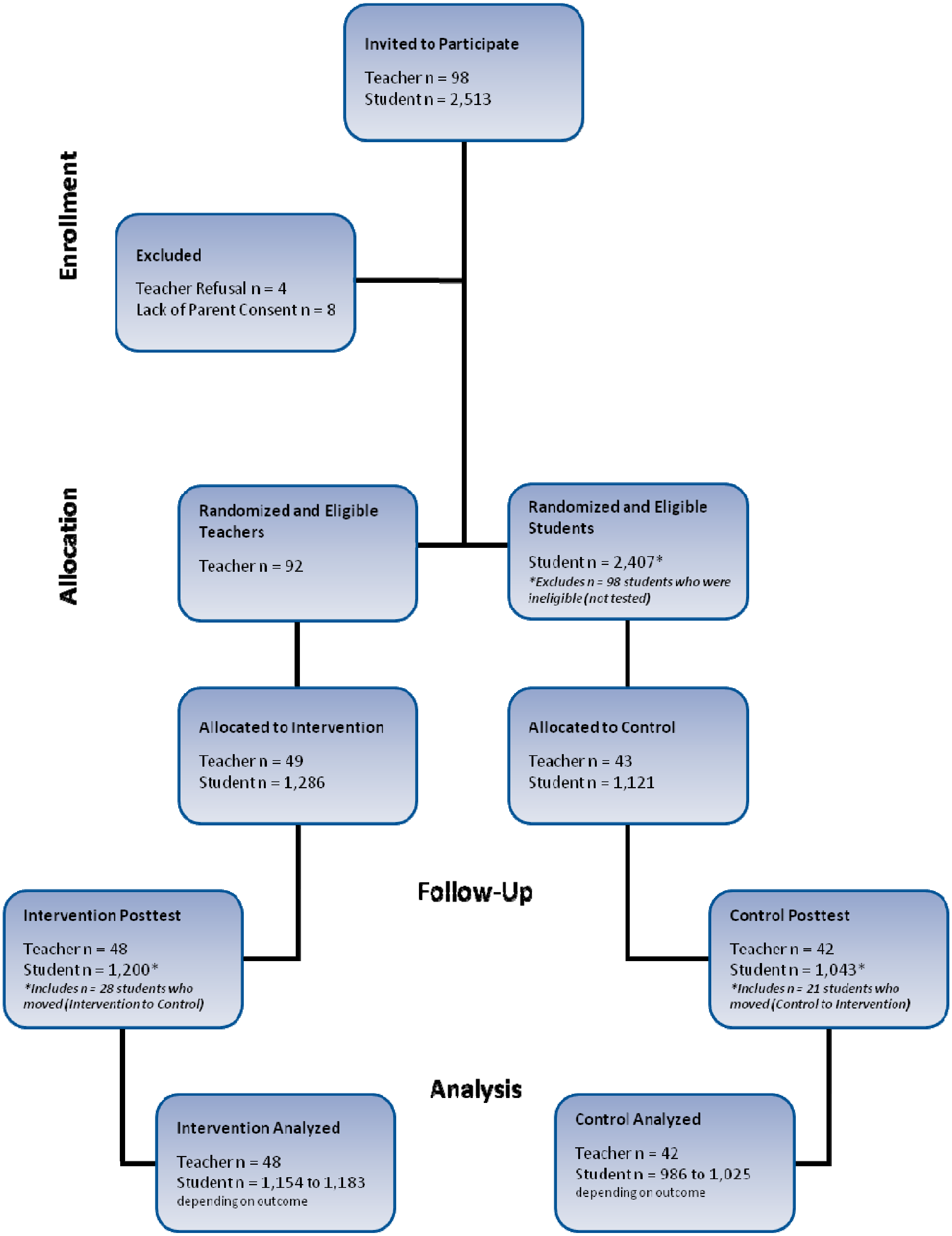
The study sample is summarized in Table 2.4 which shows the numbers of students, teachers, and classrooms in the sample and their distribution across study groups, and in Figure 2.1, which shows the study sample from recruitment through analysis.

Table 2.4 Students, Teachers, and Classes, by Study Condition

		Intervention		Control	
Study sample	**Total**	**Number**	**Percent**	**Number**	**Percent**
Teachers	92	49	53.4	43	46.7
Classes	129	67	51.9	62	48.1
Students					
Consented to participate	2,505	1,343	53.6	1,162	46.4
Eligible for testing[a]	2,407	1,286	53.4	1,121	46.6
Attrition, pretest to posttest[b]	164	86	6.7	78	6.9
Postattrition sample	2,243	1,200	53.5	1,043	46.5
Student movement[c]					
Changed teachers, not condition	20	13	1.0	7	0.6
Changed schools, not condition	5	≤3	≤0.2	≤3	≤0.3
Moved, control to intervention[d]	21	0	0.0	21	1.9
Moved, intervention to control[e]	28	28	2.2	0	0.0

[a] Students who normally would be excluded from state tests (e.g., because of their English language or special education status) were excluded from testing. Students with disabilities who required accommodations on their state tests were deemed ineligible for study testing, unless the accommodations could be fulfilled by field staff or staff at the school.
[b] Percentages were calculated within each condition by dividing the number of students who left the sample by the number of eligible students.
[c] Percentages were calculated within each condition by dividing the number of students within each movement category by the number of eligible students.
[d] Students who moved from control to intervention class between pre- and posttesting.
[e] Students who moved from intervention to control class between pre- and posttesting.
Source: Study administrative records.

Figure 2.1 Flowchart for the Study Sample, From Enrollment to Analysis



Source: Study administrative records.

Student characteristics. At baseline, no statistically significant differences were found between intervention and control groups for the primary outcomes of reading vocabulary and reading comprehension (Table 2.5). The pretest GMRT scores in the 500–502 range correspond to normal curve equivalent scores of 42–43 and grade equivalents of 5.2–5.3, which indicate that the study sample had lower than average achievement compared with the national norm (for fall for Grade 6).

Table 2.5 Reading Achievement Pretest Scores, by Study Condition

Primary outcomes: Gates-MacGinitie Reading Tests extended scale scores	**Intervention (*n* = 1,286)**			**Control (*n* = 1,121)**			***t*-statistic (standard error)[a]**	***p* value**
	Number	**Mean**	**Standard deviation**	**Number**	**Mean**	**Standard deviation**		
Reading vocabulary	1,276	502.14	32.63	1,112	502.19	35.00	0.03 (1.38)	.97
Reading comprehension	1,280	500.19	32.20	1,108	502.06	32.63	1.41 (1.33)	.16

Note: Calculations do not account for the clustering of students by teacher or teacher by school.
[a]Numbers in parentheses are standard errors of the differences between the two means for each *t*-statistic. We used Levene's test (1960) to assess the equality of variance assumption across the two groups.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

At baseline, no statistically significant differences were found between the intervention and control groups for the ancillary outcomes of student use of reading comprehension strategies or motivation for reading (Table 2.6). Reported strategy use by the sample may be considered medium, according to ranges provided by MARSI developers (Mokhtari & Reichard, 2002). Although developers of the MRQ do not provide similar guidelines for judging the motivation score, the students' mean scores are in the mid-range of the 1 (lower motivation) to 4 (higher motivation) point scale.

Table 2.6 Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Pretest Scores, by Study Condition

	Intervention (*n* = 1,286)			**Control (*n* = 1,121)**				
Ancillary outcomes	**Number**	**Mean**	**Standard deviation**	**Number**	**Mean**	**Standard deviation**	***t*-statistic (standard error)[a]**	***p* value**
Reading strategies: MARSI	1,278	3.18	0.67	1,110	3.15	0.71	–1.03 (.03)	.30
Reading motivation: MRQ	1,275	2.83	0.47	1,112	2.85	0.48	0.73 (.02)	.47

Note: Calculations do not account for the clustering of students by teacher or teacher by school. MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
[a]Numbers in parentheses are standard errors of the differences between the two means for each *t*-statistic. We used Levene's test (1960) to assess the equality of variance assumption across the two groups.
Source: MARSI and MRQ surveys administered by study team.

Intervention and control groups differed significantly in the percentages of students identified as ELLs or who had IEPs (Table 2.7). These differences were addressed in the impact analysis by including the two covariates in the analytical model. (In the analyses reported in Chapter 4, the treatment coefficient was always interpreted after adjusting for the effects of these covariates.)

Table 2.7 Student Characteristics, by Study Condition

	Intervention (n = 1,286)		Control (n = 1,121)			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	**χ^2 (degrees of freedom)**	***p* value**
Female	649	50.5	562	50.1	χ^2 (1) = 0.03	.87
Ethnicity[a]					χ^2 (4) = 5.40	.25
African American	150	12.3	140	13.3		
Asian	74	6.1	87	8.3		
Hispanic	346	28. 5	279	26.5		
White	455	37.4	393	37.3		
Other ethnicity	191	15.7	154	14.6		
Individualized education program	164	12.8	84	7.5	χ^2 (1) = 17.93	.00
English language learner	89	6.9	148	13.2	χ^2 (1) = 26.62	.00

Note: Calculations do not account for the clustering of students by teacher or teacher by school.

[a] This variable has 138 missing cases (intervention = 70; control = 68).

Source: Student self-report section on Gates-MacGinitie Reading Tests administered by study team; student rosters completed by study teachers.

Teacher characteristics. In the sample of 90 teachers, 100% had at least a bachelor’s degree. They averaged 13.4 years of teaching experience (range: 0–45) and 7.7 years of experience teaching Grade 6 (range: 0–34). Of the 90 teachers, 86 were fully certified, and 4 had probationary, provisional, or temporary certifications. No statistically significant differences in teacher characteristics were found between the intervention and control groups. Table 2.8 presents the two teacher characteristics that were used in analyses—educational attainment level and years of teaching experience. Table B2.1 in Appendix B2 presents the frequencies of additional questions and responses included in the teacher questionnaire (the disaggregated categories of the teachers’ academic degrees, types of certification, and areas of certification), by condition.

Table 2.8 Teacher Characteristics, by Study Condition

	Intervention (n = 48)		Control (n = 42)			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	**Test statistic[a]**	***p* value**
Highest degree held						
Master’s degree or higher[b]	37	77.1	28	66. 7	χ^2 (1) = 1.21	.27
Bachelor’s degree	11	22.9	14	33. 3		
Years teaching (average)						
Average, total years teaching	48	13.48 (9.07)	42	13.31 (9.32)	t = –0.09 (1.94)	.93
Average, total years teaching grade 6	48	8.25 (6.73)	42	7.11 (6.40)	t = –0.82 (1.39)	.41

Note: Frequencies of additional questions and responses from the teacher background questionnaire are presented in Table B2.1 in Appendix B2. Calculations do not account for the clustering of teacher by school.

[a] Numbers in parentheses are standard deviations (for group means), degrees of freedom (for chi-squared), or standard errors of the differences between the two means (for *t*-statistics). We used Levene’s test (1960) to assess the equality of variance assumption across the two groups.

[b] Includes teachers with master’s degrees, first professional degrees, or education specialist/professional diplomas.

Source: Teacher background questionnaire administered by study team.

School characteristics. The 19 elementary and 13 middle schools in the study had an average enrollment of 553 students, with a range of 212–1,162 students (Table 2.9). Race/ethnicity of the student population averaged 32.51% Hispanic, 22.98% Black, and 37.26% white, with lower numbers of Asian (5.85%) and American Indian (0.29%) students. Student poverty (defined by the percentage of students who were eligible for free or reduced-price lunch) in schools ranged from 29% to more than 95% and averaged 70.9%**.**

When schools were recruited during spring 2008, school poverty was based on information from the 2006–07 school year, and all 32 schools met the criterion of having more than 33% of their students eligible for free or reduced-price lunch. School characteristics reported in Table 2.9 were obtained in fall 2009, from the most current data available from state departments of education (2007–08 school year). The more recent snapshot of these schools' populations is presented here to describe the schools that participated in the study. The data from 2007–08 were used in the analytic models.

Table 2.9 Demographics of Participating Schools as of the 2007–08 School Year (Percent, Except Where Noted)

Characteristic	Mean	Standard deviation	Minimum	Maximum
School type[a]				
Middle school	41			
Elementary school	59			
Enrollment size (number)	552.75	222.23	212.00	1162.00
Student characteristics				
Eligible for free or reduced-price lunch	70.92	23.93	29.00	>95.00[b]
Special education[c]	13.69	4.89	2.90	27.00
English language learners[d]	11.78	13.53	0.00	39.40
American Indian/Alaskan Native	0.29	0.46	0.00	2.45
Asian/Pacific Islander	5.85	11.85	0.00	53.05
Hispanic	32.51	23.02	2.13	83.11
Black, non-Hispanic	22.98	21.30	1.71	84.97
White, non-Hispanic	37.26	33.45	1.65	92.55
N	32			

Note: Ethnic composition was derived from total school enrollment and reported number of students within each ethnicity. Percentages reported from each school did not always sum to 100% because of missing data.

[a] Middle school is defined as having Grades 5/6–8. Elementary school is defined as all other grade configurations (K–6; 1–8).

[b] When percentage eligible for free or reduced-price lunch was greater than 95%, the data source for Connecticut schools reported >95 instead of the actual percentage. In these instances, 95% was used in the calculations.

[c] Defined in Massachusetts as students with an individualized education program and in Connecticut and Rhode Island as students receiving special education services.

[d] English language learners are referred to in Massachusetts as limited English-proficient students and in Connecticut and Rhode Island as students receiving English as a second language/bilingual services.

Source: Connecticut Department of Education (n.d. b); Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b).

The percentages of students who scored below the proficient level on the state achievement test in participating schools and statewide were used to provide context on student achievement at the participating schools (Table 2.10). Data for 2007–08 show that the percentage of students who scored below proficient was higher on average in the study schools than the statewide average.

Table 2.10 Average Percentage of Students Scoring Below Proficient Level on 2007–08 State Grade 6 Reading/English Language Arts Assessments

		Study schools		State overall	
State	**Assessment**	**Number of students tested**	**Percent scoring below proficient**	**Number of students tested**	**Percent scoring below proficient**
Connecticut	Connecticut Mastery Test: Reading	1,262	45	42,131	22
Massachusetts	Massachusetts Comprehensive Assessment System: English language arts	1,551	46	71,575	33
Rhode Island	New England Common Assessment Program: Reading	294	59	11,129	38

Note: Average percentages for study schools within each state were calculated by dividing the number of students in all of the study schools that scored below proficient by the number of students in all of the study schools who took the assessment. State averages were provided directly by the source.
Source: Connecticut Department of Education (n.d. a); Massachusetts Department of Education (n.d. a); Rhode Island Department of Education (n.d. a).

Missing Data

The amount of missing data was examined to ascertain whether it was reducing the overall statistical power to detect intervention effects and to determine whether the amount differed between treatment and control groups. Any nonrandom differences in the amount of missing data could introduce bias in the findings.

The amount of missing data was low: 0.47–1.16% of students had missing data at pretest and 7.31–11.33% at posttest, depending on the measure. No statistically significant differences in the amount of missing data were found between intervention and control conditions. Thus, evidence does not suggest that missing data were related to the *Thinking Reader* intervention or would lead to biased impact results. Further information about missing data is presented in Appendix C1. Tables C1.1 and C1.2 show the percentage of missing data for each primary and ancillary outcome, by condition.

Estimating Impacts

The basic analytic strategy for assessing the impacts of *Thinking Reader* was to compare achievement outcomes for students whose teachers were randomly assigned to the intervention condition with those assigned to a business-as-usual control condition. The average outcome in the control group is an estimate of the achievement level that would have been observed in the intervention group if it had not used *Thinking Reader*. Thus, the difference in outcomes provides an unbiased estimate of the impact of the intervention.

Listwise deletion was used for the impact analyses presented in Chapter 4. Students were removed from the analysis of the intervention effect on each outcome if their pre- or posttest scores for that outcome were missing. The percentage of students with missing data across pre- and posttests did not exceed 11.4%, and no statistically significant differences were found between intervention and control groups; no other covariates in the impact models had missing data. Simulations reported by Puma, Olsen, Bell, & Price (2009) indicated that, with missing data at these levels and with similar assumptions, dropping cases with missing data yields unbiased impact estimates that are comparable to findings from other approaches for handling missing data. More information about the missing data can be found in Appendix C1. Tables C2.1–C2.3 in Appendix C2 present the baseline equivalence of the analytic sample for the primary achievement outcomes and the ancillary outcomes and present student characteristics for the analytic sample, by condition.

Because of the research design and hierarchical data structure (students nested within teachers and teachers nested within schools), a multilevel model was used to estimate the impact of *Thinking Reader*. The multilevel model accounts for the nested structure of the data and yields appropriate standard errors for the effects of interest (Raudenbush & Bryk, 2002).[19] Effect sizes were computed to show the magnitude of the effect of *Thinking Reader*, expressed in standard deviation units. The impact model is presented in detail in Appendix C3.

Adjusting for Multiple Comparisons

This study addresses four research questions that involve the two primary outcomes of reading vocabulary and reading comprehension. To account for the two comparisons and to avoid spurious positive findings, a multiple comparison adjustment (a Bonferroni correction) was used that divided the critical *p* value (alpha) in half. This means that an impact result on the measures of vocabulary and comprehension must have an observed *p* value lower than .025, instead of the more standard .05, to be considered statistically significant.

Sensitivity Analyses

The robustness of the results was tested through analyses that examined how sensitive the impact estimates were to the assumptions made about listwise deletion (by reanalyzing the data using two multiple imputation models), specification of the treatment coefficient as a random effect (by reanalyzing the data using a fixed-effect treatment coefficient), specification of a three-level model (by reanalyzing the data using a two-level model), the number of classes per teacher (multiple classrooms or one classroom), and retention in the sample of students whose original teachers left their schools at the beginning of the school year (see Chapter 4).

Estimating Impacts for Exploratory Research Questions

To answer the four exploratory research questions, we used a similar analytic strategy to the one we used to respond to the main research questions. Three-level multilevel models using listwise deletion samples were created to address these questions.

[19] The HLM 6.08 program was used to analyze the multilevel models (Raudenbush, Bryk, & Congdon, 2008).

Chapter 3.
Implementation of the *Thinking Reader* Intervention

This chapter describes the professional development and training provided and dosage/fidelity data on how *Thinking Reader* was implemented. It looks at whether all students in the intervention classes read all three digital novels and whether teachers used the recommended instructional routine. The chapter also considers the intensity of the intervention delivery, examining whether the books were read for at least the recommended minimum number of minutes per week and whether the books were completed within the expected number of weeks. Finally, it presents findings from classroom observations and compares literacy instruction in intervention and control classrooms.

Professional Development

Intervention teachers received two, 6-hour group professional development sessions in the fall; three individual coaching sessions in the fall, winter, and spring (totaling 7.5–8.5 hours); and other opportunities for communication with coaches throughout the year. The training and coaching provided were typical of that provided to teachers in Dalton and colleagues' (2002) quasi-experimental study. Systematic information about the amount of training that typical *Thinking Reader* customers implement is not available.[20]

Intervention teachers attended group trainings at one of four locations selected to be geographically convenient to clusters of study schools. Training was conducted by four trainers from the Center for Applied Special Technology (CAST) and a trainer from Tom Snyder Productions. Support staff (e.g., special education teachers, reading specialists/consultants, and paraprofessionals) who worked regularly with the intervention teachers were invited to attend the training sessions so that they could understand the principles of the intervention.

At the first round of introductory workshops in September and early October 2008, the training outlined the role of reading strategies in improving comprehension and the value of universal design features to support diverse learners, featured interactive demonstrations and guided practice in using *Thinking Reader*, and shared strategies for launching and managing *Thinking Reader*.[21] Of the 49 intervention teachers, most attended the first group training or received individual make-up training; fewer than 4 teachers elected not to participate.

A second 6-hour follow-up training was conducted in November, some 6–8 weeks after the first workshop, and focused on further integrating reading strategies into instruction. Teachers shared implementation successes and challenges and had the opportunity to ask technical questions. They analyzed program assessment results and practiced holding conferences with students to discuss and scaffold progress. Teachers also received information about how to organize a culminating activity with students after they would finish reading a book. Trainers shared ideas

[20] According to Tom Snyder Productions (personal communication, July 24, 2007), although customers are encouraged to purchase professional development, schools vary in the amount of training they conduct. Because a definition of "typical" training was not available, the study implemented the model used in past research on the program (Dalton et al., 2002), which is the amount of training recommended by the developers of the program.

[21] Training materials are available online at http://www.literacyintervention.org/participantspage_training.asp.

for activities, such as painting a class mural about the novel, adapting the novel as a play, and creating a newspaper about the novel. Most teachers attended or received makeup training; fewer than four teachers elected not to participate.

During the first two coaching sessions (held December 2008–February 2009 and February–March 2009), CAST coaches conducted individual sessions followed by a debriefing in the teacher's classroom or the school's computer lab. Each component lasted 2–3 hours. This session included time for the coach to watch the teacher during *Thinking Reader* instruction. Teachers could also observe the coach teaching a particular instructional process in conjunction with the program. During debriefing, the coach and the teacher reflected on the instruction, engaged in problem-solving, and set goals for the next phase of implementation. Most teachers engaged in the individual meetings—with fewer than four teachers who did not attend. Each coaching visit also included a 2-hour after-school group meeting of teachers from one or more schools in the same area to enable teachers to share successes, strategies, and teaching ideas. Attendance was 81% at the first visit and 77% at the second. The third set of coaching sessions took place in May 2009 and consisted of 1.5-hour conference calls between several teachers and two coaches. On the calls, each teacher highlighted a student who had shown reading improvement; teachers also reflected on their overall learning in strategy instruction. Attendance for the telephone sessions was 63%.

Several other methods were used to facilitate information exchange between teachers and coaches. Coaches responded to teachers' questions by e-mail, and teachers also received a biweekly e-mail check-in asking them to share their progress on implementation and to report any technical or instructional challenges that coaches could help troubleshoot. An average of 61% of teachers responded to the 17 probes during the year. Of those who responded, an average of 94% for each probe reported that the software was working technically "well" or "very well."

In addition, an electronic mailing list for trainers and coaches was set up to answer common questions for the whole group and to share instructional ideas or concerns. CAST staff sent 22 mailing list posts, and teachers posted 35 responses.

Implementation of *Thinking Reader*

Program use. According to CAST, optimal implementation of *Thinking Reader* involves 110–165 minutes of program use per week. In training, intervention teachers were told to adhere to these guidelines during the time a novel was being covered. This gave teachers flexibility to schedule *Thinking Reader* use around their regular curriculum and computer availability. Trainers estimated that each of the novels would take 4–6 weeks to complete. Trainers also recommended that teachers use the program two or three times per week, with two additional class sessions for end-of-novel activities. If these recommendations were followed and three novels were covered, students' exposure to *Thinking Reader* would range between 1,320 and 2,970 total minutes of use across 24–54 days.

Trainers suggested initiating the first novel in fall, the second novel in the winter, and the third novel in the spring—with weeks off, as needed or desired, between novels. Teachers were advised to work around their class schedules, school vacation, testing schedules, and computer availability to complete the three novels. Trainers acknowledged that part of implementing the program includes planning how *Thinking Reader* would be integrated with other instruction.

While a *Thinking Reader* novel was being covered, teachers might plan days “off” the computer to cover curricular components unrelated to the novel (e.g., spelling, grammar, or word study) or they could choose to incorporate the novel with activities (e.g., reviewing the plot or studying vocabulary). When a *Thinking Reader* novel was not being covered, teachers could continue to cover reading material that would regularly be part of the curriculum. Coaches collected biweekly probes and visited classrooms, making suggestions or inquiries about program use. They did not try to enforce suggested time guidelines in any way, in line with the developer’s philosophy that teachers need to adopt innovation without coercion (P. Coyne and B. Dalton, personal communication, April 22, 2010).

Information from the *Thinking Reader* electronic reports was used to establish the number of novels initiated by each teacher; the number of novels completed by students of each teacher; the average number of minutes, days, and weeks that students of each teacher spent on each novel; and the level of support (ranging from 1 to 5) that teachers set for each student across each book. The program usage data presented below are based on the intervention sample (*n* = 1,286) although 158 (12%) of these students did not have usable *Thinking Reader* electronic report data.

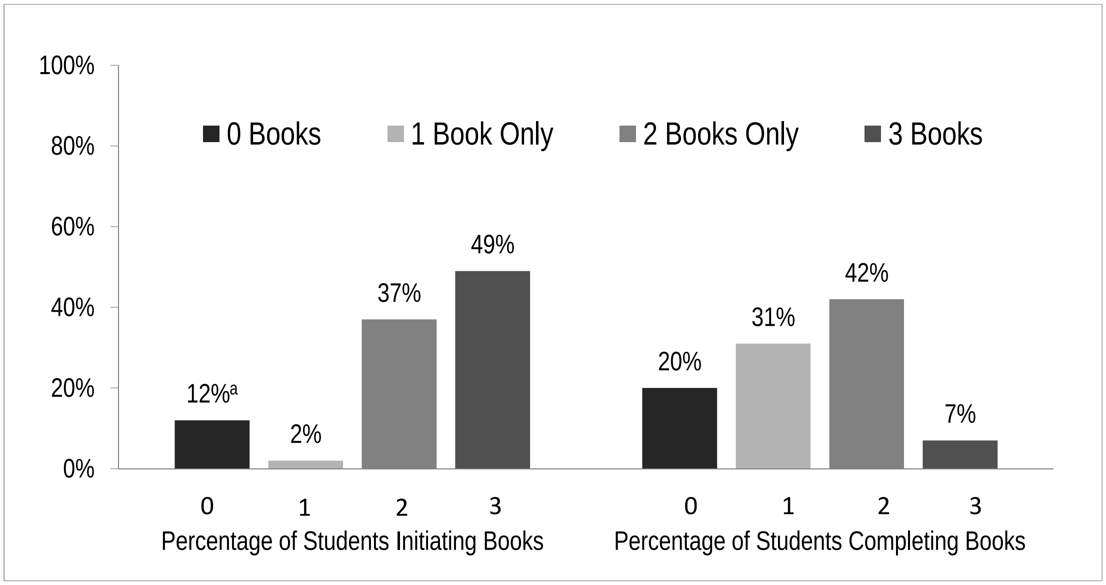
Most teachers initiated a first book and a second book (fewer than four teachers did not). Twenty-seven teachers initiated a third book. Overall, 22 teachers started the program later than planned, so intervention use was not spread across the year as intended. For the first book, 25 teachers began implementation in October, as originally planned; 14 in November; and 8 in December or January.

Figure 3.1 shows the percentage of students initiating and completing books and the number of books (regardless of sequence). Students were not required to complete one book before starting another.

Over the school year, students in the intervention group averaged 1,013.7 minutes (about 17 hours) using *Thinking Reader* across 25.5 days and 16.9 weeks. This was an average of 57.7 minutes across 1.4 days per week. The total number of minutes students spent using the software program was lower than expected (69.1% of the sample was below the recommended range). Additionally, although students were accessing the program within the expected range of total days of use, usage was on the lower end of the expected range (37.2% of the sample was below the recommended range). For the books with higher completion rates (Books 1 and 2), students exceeded the expected 4- to 6-week completion time on average, resulting in less concentrated use of the software. When situated in the context of total days in a school year (approximately 180), students used *Thinking Reader* for an average of 14.2% of school days. Table 3.1 presents information about software use in minutes, days, and weeks by book, as well as average minutes per week and average days per week.

Program usage varied by book. Students spent similar amounts of time in *Thinking Reader* for their first and second books, but spent approximately half as much time on their third book. More than half (50.9%) of students did not start the third book, and completion rates dropped markedly from Book 1 (73.7%) to Book 3 (8.9%). Novel completion details, by book, are shown in Table 3.2.

Figure 3.1 *Thinking Reader* Use by Students in the Intervention Group, by Number of Books



Note: Percentages are based on the treatment sample (*n* = 1,286).
[a]Twelve percent of students (158 students) did not have usable student progress reports and/or data from class worklogs. These 158 students are also included in the 20% of students who did not complete any book.
Source: Student progress reports from *Thinking Reader* software collected by study team.

Twenty-one control group students (1.9% of the control group and 0.9% of the total sample) used the software at some point during the study year. These were students who had crossed over from control to treatment classrooms; all maintained their control group standing in the impact analyses, for intent-to-treat purposes.[22]

Because a number of students did not use *Thinking Reader*, we next present software usage information for only those 1,128 intervention students who actually used the program. This group had higher usage than the intent-to-treat sample as a whole. Over the school year, students in the intervention group who used the software averaged 1,195.8 total minutes (about 20 hours) using *Thinking Reader* across 29 days and 19.2 weeks. This was an average of 68.4 minutes across 1.6 days per week (see Table 3.3).

[22] For Book 1, 17 control group students used the software and 13 completed the book. For Book 2, 21 control group students used the software and 11 completed the book. For Book 3, 11 control group students used the software and none completed or almost completed the book. The control group students who used the program spent an average of 437.44 minutes across 11.76 days and 8.68 weeks on Book 1, 423.48 minutes across 10.70 days and 7.35 weeks on Book 2, and 189.20 minutes across 6.00 days and 2.35 weeks on Book 3.

Table 3.1 *Thinking Reader* Use, by Book (for Entire Intervention Group)

	First book (*n* = 1,286)			Second book (*n* = 1,286)			Third book (*n* = 1,286)			Total (*n* = 1,286)			
Average use	**Mean**	**Standard deviation**	**Range**	**Mean**	**Standard deviation**	**Range**	**Mean**	**Standard deviation**	**Range**	**Mean**	**Standard deviation**	**Range**	**Expected use**
Minutes[a]	469.8	337.6	0–1,840	424.9	364.2	0–1,795	119.0	234.3	0–1,775	1,013.7	635.1	0–3,639	1,320–2,970
Days[b]	11.8	6.3	0–30	10.8	7.1	0–39	2.8	4.5	0–27	25.5	13.0	0–63	24–54
Weeks[c]	8.3	5.2	0–25	7.0	4.3	0–17	1.7	2.8	0–13	16.9	8.4	0–50	12–18
Minutes/week[d]	59.6	58.5	0–1,202	56.0	53.2	0–683	42.2	81.3	0–969	57.7	40.9	0–366	110–165
Days/week[e]	1.4	0.8	0–5	1.4	0.9	0–5	1.0	1.3	0–6	1.4	0.8	0–5	2–3

Note: Calculations do not account for the clustering of students by teacher or teacher by school.

[a] Because logging out of the software was not automatic and failure to log out could result in a large number of total minutes used, outliers above the 95th percentile for the sample were excluded from estimates. CAST verified that these outliers were likely errant data.

[b] Average day's calculations involved counting the number of days a student logged into *Thinking Reader* software for each book.

[c] Average week's calculations involved subtracting the start date from the end date for each student and dividing by 7. The span of weeks could have included time for vacation, testing, or some other extended break from program use.

[d] Average minutes per week's calculations involved dividing students' total minutes by total weeks.

[e] Average days per week's calculations involved dividing students' total days by total weeks.

Source: Student progress reports and data from class worklogs from *Thinking Reader* software collected by study team.

Table 3.2 *Thinking Reader* Completion, by Book (for Entire Intervention Group)

	First book		Second book		Third book	
Novel completion	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**
Completed the novel	948	73.7	680	52.9	115	8.9
Completed three-fourths of the novel	125	9.7	129	10.0	68	5.3
Completed less than three-fourths of the novel	48	3.7	295	22.9	449	34.9
Did not initiate the novel	165	12.8	182	14.2	654	50.9
Total	1,286	100.0	1,286	100.0	1,286	100.0

Source: Student progress reports from *Thinking Reader* software collected by study team.

Table 3.3 *Thinking Reader* Use, by Book (for Intervention Students Who Used the Software)

	First book (*n* = 1,121)			Second book (*n* = 1,104)			Third book (*n* = 632)			Total (*n* = 1,128)			
Average use	Mean	Standard deviation	Range	Mean	Standard deviation	Range	Mean	Standard deviation	Range	Mean	Standard deviation	Range	Expected use
Minutes[a]	595.2	264.1	35–1,840	549.7	321.0	14–1,795	279.5	290.3	1–1,775	1,195.8	507.7	14–3,639	1,320–2,970
Days[b]	13.4	4.8	1–30	12.7	5.9	1–39	6.0	5.0	1–27	29.0	9.6	1–63	24–54
Weeks[c]	9.5	4.4	1–25	8.2	3.5	1–17	3.5	3.2	1–13	19.2	6.1	1–50	12–18
Minutes/week[d]	75.9	56.0	6–1,202	72.4	49.8	9–683	102.4	99.4	5–969	68.4	36.5	1–463	110–165
Days/week[e]	1.6	0.7	0–4	1.7	0.7	0–5	2.0	1.2	0–6	1.6	0.6	0–5	2–3

Note: Calculations do not account for the clustering of students by teacher or teacher by school. A total of 158 students were without usable student progress reports and/or data from class worklogs.

[a] Average minute's calculations account for the students who started a book. Because logging out of the software was not automatic and failure to log out could result in a large number of total minutes used, outliers above the 95% percentile for the sample were excluded from estimates. CAST verified that these outliers were likely errant data.

[b] Average day's calculations involved counting the number of days a student logged into *Thinking Reader* software for each book.

[c] Average week's calculations involved subtracting the start date from the end date for each student and dividing by 7. The span of weeks could have included time for vacation, testing, or some other extended break from program use.

[d] Average minutes per week's calculations involved dividing students' total minutes by total weeks.

[e] Average days per week's calculations involved dividing students' total days by total weeks.

Source: Student progress reports and data from class worklogs from *Thinking Reader* software collected by study team.

Calibration of level of support to student achievement. A key feature of the *Thinking Reader* program is that teachers can adjust the level of support provided to individual students so that each can progress to higher levels and become less reliant on program supports (Tom Snyder Productions, 2006). Trainers suggested that all students begin at Level 1, which offers more support, when initiating the first novel to give them a chance to become familiar with *Thinking Reader*. Additionally, the software program defaults to Level 1 at the beginning of each book. Teachers could give students access to change their own level. Table 3.4 presents the initial and final levels of support provided to students for the first strategy prompt and the last prompt they finished within each book.

Table 3.4 *Thinking Reader* Use, by Level of Support and Book Order (for Intervention Students Who Used the Software)

Level	First book: Initial level		Second book: Initial level		Third book: Initial level	
	Number	**Percent**	**Number**	**Percent**	**Number**	**Percent**
1	1,081	95.6	968	87.8	486	79.5
2	36	3.2	79	7.2	69	11.3
3	<15	<1.3	45	4.1	22	3.6
4	≤3	≤0.3	6	0.5	21	3.4
5	—	—	5	0.5	13	2.1
Total	1,131	100.0	1,103	100.0	611	100.0
Level	**First book: End level**		**Second book: End level**		**Third book: End level**	
	Number	**Percent**	**Number**	**Percent**	**Number**	**Percent**
1	852	75.3	610	55.3	392	64.2
2	191	16.9	268	24.3	111	18.2
3	76	6.7	188	17.0	54	8.8
4	6	0.5	25	2.3	39	6.4
5	6	0.5	12	1.1	15	2.5
Total	1,131	100.0	1,103	100.0	611	100.0

Note: Calculations are based on 1,131 students because 155 students from the intervention group did not have data from class worklogs. Percent is the valid percent not accounting for the students who did not start the book. Calculations do not account for the clustering of students by teacher or teacher by school.
Source: Class worklogs from *Thinking Reader* software collected by study team.

For the most part, teachers followed the trainers' recommendation to start all students at Level 1; more than 95% of students used the highest level of support when beginning their first book. Most students also began Books 2 (88%) and 3 (80%) at Level 1. Additionally, although most students (75%) finished their first book at Level 1, fewer students remained at Level 1 at the end of their second (55%) and third (64%) books. Table 3.5 presents a summary of change from initial to final levels of support for each book. Overall, most students did not change levels within each book (79% in the first book, 63% in the second book, and 84% in the third book).

Table 3.5 *Thinking Reader* Use, by Change in Level and Book Order (for Intervention Students Who Used the Software)

	First book		Second book		Third book	
Change in level	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**
Decreased 1–2 levels	≤3	≤0.3	5	0.5	≤3	≤0.5
No level change	888	78.5	695	63.0	513	84.0
Increased 1 level	177	15.6	261	23.7	44	7.2
Increased 2 levels	53	4.7	125	11.3	34	5.6
Increased 3–4 levels	<15	<1.3	17	1.5	<20	<3.3
Total	1,131	100.0	1,103	100.0	611	100.0

Note: Calculations are based on 1,131 students because 155 students from the intervention group did not have data from class worklogs. Percent is the valid percent not accounting for the students who did not start the book. Calculations do not account for the clustering of students by teacher or teacher by school.
Source: Class worklogs from *Thinking Reader* software collected by study team.

Although only 21% of students increased levels in Book 1, 37% of students who initiated a second book increased at least one level. Few students changed levels in Book 3, perhaps because many students did not work long on or finish the third book.

Of those students who progressed to a higher level in any book, most increased only one or two levels; only a small fraction of students increased more than two levels in each book (0.9% in Book 1, 1.5% in Book 2, and 3.1% in Book 3). Changing students' level of support was concentrated among 69% of teachers for Books 1 and 2 and 33% for Book 3. In other words, more than 1 out of 4 teachers made no level adjustments for students during Books 1 and 2 and more than twice as many teachers (67%) made no level adjustments during Book 3.

A next question is whether these adjustments were calibrated to students' reading performance and development. Although trainers did not explicitly direct teachers on how they should determine level adjustments, trainers suggested that teachers meet with each student regularly over the course of the year to discuss reading growth. Additionally, electronic reports for each student included responses to mandatory prompts and quizzes as progress was made through a novel. Each book contained 4–6 quizzes, consisting of 10-multiple choice items: 5 fact recall questions, 2 vocabulary questions, and 3 inference questions (Tom Snyder Productions, 2006). The difficulty of each quiz was standardized across levels (i.e., not tailored to levels of support).

To examine how students' level of support might be related to their progress, we examined the relationship between students' final level of support and average quiz scores for each book. Findings are presented in Table 3.6. These correlational analyses demonstrate a positive and statistically significant relationship between students' level of support and their average quiz scores. With a few exceptions, students who finished each book at higher levels generally had higher average quiz scores. Caution is warranted in interpreting data for students who ended at Levels 4 and 5 given the small number of students at each of these levels. One possible explanation for the positive associations between students' final level of support and their average quiz scores for each book is that teachers could have been calibrating students' levels of support according to their performance.

Table 3.6 *Thinking Reader* Use, by Level of Support for Each Book and Average Quiz (for Intervention Students Who Used the Software)

	Quiz score								
	First book			Second book			Third book		
Final level	Number	Mean	Standard deviation	Number	Mean	Standard deviation	Number	Mean	Standard deviation
1	834	65.84	0.16	579	64.11	0.17	312	65.39	0.20
2	189	75.75	0.13	261	70.96	0.14	90	68.26	0.19
3	76	72.16	0.14	176	75.40	0.14	46	69.22	0.19
4	6	74.00	0.16	25	80.44	0.08	34	76.97	0.14
5	6	72.33	0.10	11	71.73	0.14	14	76.93	0.13
Total	1,111	68.04	0.16	1,052	68.17	0.16	496	67.38	0.19
Pearson's correlation[a]	0.19	p = .00		0.28	p = .00		0.17	p = .00	

Note: A total of 158 students were without usable student progress reports and/or data from class worklogs. Calculations do not account for the clustering of students by teacher or teacher by school.

[a] Pearson's correlation is the correlation between average quiz score and final level for each book. Spearman's correlation was also calculated: Book 1, 0.25 (*p* value of .00); Book 2, 0.30 (*p* value of .00); and Book 3, 0.17 (*p* value of .00).

Source: Student progress reports and data from class worklogs from *Thinking Reader* software collected by study team.

Summary. Although students in the treatment group accessed the *Thinking Reader* software program within the expected range of total days of use, their overall time spent using the program, as measured by total minutes, was lower than the time suggested by trainers. Additionally, these total minutes spent using the program were diffused across a larger number of weeks for Books 1 and 2 than expected, potentially resulting in less concentrated usage than advised by program trainers. Book completion rates were also lower than expected and dropped substantially from Book 1 to Book 3, with less than 1 in 10 students completing Book 3. The study is limited in that actual take up of the intervention in the way the developer intended was low.

With regard to level of support, teachers followed instructions to start students at Level 1 but did not advance the majority of students to higher levels (less support) that require more general use of the focal reading comprehension strategies in the *Thinking Reader* program. Also, at least 1 out of 4 teachers did not make any adjustments to students' levels of support for each book. When teachers made adjustments, they appeared to be associated to some extent with students' average quiz scores. The data on program use are limited in that they do not capture the extent to which students made use of the software support features and do not capture the quality and depth of teacher implementation. The fact that the study was not designed to collect more in-depth information on implementation is a limitation.

Classroom Observations

Adherence to three-phase instructional routine in intervention classrooms. The observers' notes were examined to assess whether teachers applied the three-component instructional routine during the two observations. One expert coded each of the 61 *Thinking Reader* observations to tabulate how many components of the instructional routine were conducted (Table 3.7). A second expert reviewed the codes, agreeing with 99% of them (the pairwise Cohen's kappa value was .98).

During 12 observations (20%), intervention teachers used all three components of the instructional routine. During 17 observations (28%), teachers completed two components of the instructional routine. During 12 of the 17, teachers conducted a prereading discussion before students logged in to *Thinking Reader* but did not engage in postreading activities. During 4 of these 17, teachers had students use *Thinking Reader* without a prereading discussion, but ended the class with a postreading activity. During 31 observations (51%), teachers used just one component, directing students to use the computer during the *Thinking Reader* portion of class without engaging in pre- or postreading discussions.

Table 3.7 Number of Components of the Three-Phase Instructional Routine Used Across Observation Visits

Observation visit	None	One	Two	Three
First	0	14	10	8
Second	1	17	7	4
Total	1	31	17	12

Source: Classroom observations conducted by the study team.

Consistency in use of the three-phase routine across the first and the second observations was also examined, to see whether teachers as a group decreased or increased their use of the routine over the school year. Twenty-eight teachers who were observed during both visits were included in this analysis (56 observations). Overall, use of the routine decreased over the year. Sixteen teachers used the same number of routine components during the two visits. Of these, fewer than four teachers used the three-phase instructional routine during both observed classes, fewer than four teachers used two phases during each visit, and 10 teachers used one phase. The remaining teachers changed the number of routine components they used from one observation to the other (fewer than four increased and nine decreased).

Whether teachers used the three-phase instructional routine more or less often during the school year compared to the class periods observed is unknown. Additionally, teachers observed using incomplete instructional routines might have covered the remaining components of the instructional routine during unobserved class periods. This practice would be consistent with the advice of trainers, who suggested that teachers with class time constraints or scheduling conflicts conduct components of the instructional routine over more than 1 day (B. Dalton, personal communication, April 12, 2010). For example, a teacher could conduct prereading activities and *Thinking Reader* instruction on 1 day, and postreading debriefing the next day.

Nature of instruction in intervention and control classrooms. The analyses combined segment data across the first and second observations, for a total of 413 segments in the intervention condition and 433 segments in the control condition. This section outlines the data for each dimension of the Center for the Improvement of Early Reading Achievement (CIERA) Classroom Observation Scheme. Percentages were calculated by dividing the number of occurrences by the number of segments. Because more than one category (variable) could be observed in a segment, percentages do not total to 100. Chi-squared statistics were used to determine whether the distribution of categorical variables for each dimension differed statistically ($\alpha < .05$) between intervention and control groups.

Dimension 1: Provider of instruction. Statistically significant differences were found between intervention and control groups in the number of segments during which the main instructor was the classroom teacher, the computer, a specialist, or an aide. During two segments, no one provided instruction; no statistically significant difference was found between the two groups for this variable (Table 3.8).

Table 3.8 Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Dimension 1: Provider of Instruction, by Study Condition

	Intervention		Control			
Provider of instruction	**Number**	**Percent**	**Number**	**Percent**	$\chi^2(df = 1)$	***p* value**
Classroom teacher	335	81.1	430	99.3	80.81	.00
Computer	337	81.6	3	0.7	575.61	.00
Specialist	79	19.1	13	3.0	56.72	.00
Aide	23	5.6	9	2.1	7.08	.01
No one	0	0.0	2	0.5	1.91	.17
Total 5-minute segments	413		433			

Note: Percentages do not sum to 100 because more than one type of instructor could be observed in each 5-minute segment. Analysis includes segments from both classroom observations for each teacher.
Source: Classroom observations conducted by study team.

The intervention group (81.1% of segments) was less likely than the control group (99.3% of segments) to have the teacher as the main instructor and more likely (81.6%) than the control group (0.7%) to have the computer as the main source of instruction. The greater number of segments in which the computer was observed to be the main provider of instruction was likely due to the fact that *Thinking Reader* is a software intervention and observations of intervention classrooms were scheduled for times when classes would be using *Thinking Reader*.

Compared with teachers and the computer, specialists and aides were observed providing instruction in fewer segments. However, they were more likely to be doing so in intervention classrooms (19.1% and 5.6% of segments) than in control group classrooms (3.0% and 2.1%).

Dimension 2: Instructional groupings. Statistically significant differences between the intervention and control groups were found for the number of segments during which students were working in large groups, small groups, individually, or individually with the teacher (Table 3.4). Control group students were more frequently observed to be working in large groups than were intervention group students (75.3% for control versus 23.2% for intervention), as well as in small groups (30.0% versus 5.3%). Intervention group students more frequently worked individually (83.8%) than control group students (35.8%). Similarly, intervention group students more frequently worked individually with the teacher (13.6%) than control group students (5.3%). No statistically significant difference was found between groups working in pairs (3.9% intervention versus 6.0% control).

Intervention teachers did not engage in whole class/large group prereading discussion or postreading activities in 51% of the observations, which could account for the higher frequency of segments in which students worked individually (83.8% intervention versus 35.8% in the control) or individually with the teacher (13.6% intervention versus 5.3% control).

Table 3.9 Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Dimension 2: Instructional Groupings, by Study Condition

	Intervention		Control			
Instructional grouping	Number	Percent	Number	Percent	$\chi^2(df = 1)$	*p* value
Whole class/large group	96	23.2	326	75.3	229.02	.00
Small group	22	5.3	130	30.0	87.47	.00
Pairs	16	3.9	26	6.0	2.03	.15
Individual	346	83. 8	155	35.8	201.50	.00
Individual with teacher	56	13. 6	23	5.3	16.98	.00
Total 5-minute segments	413		433			

Note: Percentages do not sum to 100 because more than one instructional grouping could be observed in a 5-minute segment. Analysis includes segments from both classroom observations for each teacher.
Source: Classroom observations conducted by study team.

Dimension 3: Academic content focus. Statistically significant differences between groups were found in content focus for reading, other language-related content,[23] and non-literacy-related academic content (Table 3.10). The intervention classes covered reading in more segments (95.6%) than control classes (87.8%). The control condition classes were more likely than the intervention conditions classes to cover other language-related content (4.6% versus 1.5%) and non-literacy-related content (4.9% versus 1.2%). In only one instance did an entirely nonacademic segment cross both groups, so no statistically significant differences existed between groups for that variable. No statistically significant difference was found between groups for content focused on composition/writing (3.9% intervention versus 6.2% control).

The *Thinking Reader* intervention focuses academic content on reading, and intervention students used *Thinking Reader* during 83.1% of classroom observation segments (see results for Dimension 4). Control classrooms did not have a single common curricular component, such as *Thinking Reader*, that focused instruction so specifically. Although reading was a large academic content focus in control condition classrooms, the focus of instruction was more likely to fall across additional areas of literacy, such as writing or other language-related content.

Table 3.10 Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Dimension 3: Academic Content Focus, by Study Condition

	Intervention		Control			
Academic content focus	Number	Percent	Number	Percent	$\chi^2(df = 1)$	*p* value
Reading	395	95.6	380	87.8	17.08	.00
Composition/writing process	16	3.9	27	6.2	2.44	.12
Other language	6	1. 5	20	4.6	7.11	.01
Other academic (not literacy)	5	1.2	21	4.9	9.40	.00
Not academic	0	0.0	1	0.2	0.95	.33
Total 5-minute segments	413		433			

Note: Percentages do not sum to 100 because more than one academic focus could be observed in a 5-minute segment. Analysis includes segments from both classroom observations for each teacher.
Source: Classroom observations conducted by study team.

[23] Examples are grammar, mechanics, oral expression, spelling, and handwriting.

Dimension 4: Instructional materials. Statistically significant differences were observed between intervention and control condition groups in all but three categories: use of hard-copy versions of novels in the *Thinking Reader* program, computer use other than *Thinking Reader*, and the not-applicable category (Table 3.11).

The only instructional material that was observed in more segments for the intervention group than for the control group was digital narrative text in the context of the *Thinking Reader* program (83.1% versus 0.0%). Ten other instructional material categories were observed in more segments for the control group than for the intervention group. For three categories, no statistically significant differences were found between the two groups: hard-copy versions of novels in the *Thinking Reader* program (15.3% intervention versus 17.3% control), computer use other than *Thinking Reader* (7.3% versus 4.9%), and the not-applicable category (0.5% versus 1.9%). Thus, none of the control condition classrooms were observed using *Thinking Reader* in any segment, and compared with the intervention classrooms, the control classrooms used a wider variety of instructional materials across segments—perhaps because they did not have a common tool that restricted or focused the range of materials.

Table 3.11 Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Dimension 4: Instructional Materials, by Study Condition

	Intervention		Control			
Instructional materials	**Number**	**Percent**	**Number**	**Percent**	**χ^2(*df* = 1)**	***p* value**
Narrative text						
Thinking Reader novel, digital text	343	83.1	0	0.0	604.83	.00
Thinking Reader novel, hard-copy	63	15.3	75	17.3	0.66	.42
Textbook	4	1.0	42	9.7	31.34	.00
Other	15	3.6	171	39.5	158.48	.00
Informational text						
Textbook	2	0.5	25	5.8	19.14	.00
Other	1	0.2	31	7.2	27.79	.00
Other materials						
Computer, not *Thinking Reader*	30	7.3	21	4.9	2.17	.14
Video/television/audio	0	0.0	7	1.6	6.73	.01
Overhead projector	13	3.2	55	12.7	26.11	.00
Student writing	24	5.8	65	15.0	19.01	.00
Worksheet	98	23.7	257	59.4	110.15	.00
Board/chart/posters	28	6. 8	115	26. 6	58.88	.00
Other	31	7.5	79	18.2	21.55	.00
Not applicable	2	0.5	8	1.9	3.36	.07
Total 5-minute segments	413		433			

Note: Percentages do not sum to 100 because more than one item of instruction material could be observed in a 5-minute segment. Analysis includes segments from both classroom observations for each teacher.
Source: Classroom observations conducted by study team.

Dimension 5: Literacy activities. Among the 15 literacy activities, all comparisons were statistically significant, except spelling activities (Table 3.12).

Compared with the control group, six literacy activities were observed in more segments for the intervention group: listening to connected text[24] (80.9% intervention versus 25.9% control), vocabulary (41.4% versus 30.3%), higher level comprehension (88.6% versus 66.7%), metacognition (11.4% versus 5.3%), quiz/assessment (37.1% versus 5.3%), and worklog activities (47.5% versus 0.0%).

Compared with the intervention group, eight literacy activities were observed in more segments for the control group: reading connected text (45.5% control versus 21.6% intervention), lower level comprehension (20.1% versus 10.4%), identification (25.4% versus 8.5%), text elements (15.2% versus 1.5%), language development (13.6% versus 2.7%), writing (13.4% versus 5.3%), word work (1.4% versus 0.0%), and other activities (45.7% versus 26.2%).

These results suggest that observers witnessed the main elements of *Thinking Reader* in use. The frequencies of the "listening to connected text" variable indicate that students were using the text-to-speech feature of the software, which allows students to listen to text being read. The higher incidences of vocabulary, higher level comprehension, metacognition, and assessment codes in the intervention group likely reflect *Thinking Reader*'s built-in glossary, prompts to engage students in comprehension strategies, comprehension quick checks, and chapter quizzes. Worklogs, which are used to help assess student progress, are built into *Thinking Reader* and were not used by any students in the control condition. The range of literacy activities was broader in control classrooms than in intervention classrooms, reflecting the lack of a common curricular component that would focus literacy instruction on certain activities.

Table 3.12 Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Dimension 5: Literacy Activities, by Study Condition

	Intervention		Control			
Literacy activity	**Number**	**Percent**	**Number**	**Percent**	**χ^2(*df* = 1)**	***p* value**
Reading connected text	89	21.6	197	45.5	54.17	.00
Listening to connected text	334	80.9	112	25.9	256.58	.00
Vocabulary	171	41.4	131	30.3	11.45	.00
Comprehension						
Lower level	43	10.4	87	20.1	15.23	.00
Higher level	366	88.6	289	66.7	57.87	.00
Identification	35	8.5	110	25.4	42.66	.00
Metacognition	47	11.4	23	5.3	10.26	.00
Text elements	6	1.5	66	15.2	51.62	.00
Language development	11	2.7	59	13.6	33.47	.00
Writing	22	5.3	58	13.4	16.07	.00
Word work	0	0.0	6	1.4	5.76	.02
Spelling	3	0.7	3	0.7	0.00	.95
Quiz/assessment	153	37.1	23	5.3	129.20	.00
Worklog related	196	47.5	0	0.0	267.46	.00
Other	108	26.2	198	45.7	35.09	.00
Total 5-minute segments	413		433			

Note: Percentages do not sum to 100 because more than one literacy activity could be observed in a 5-minute segment. Analysis includes segments from both classroom observations for each teacher.
Source: Classroom observations conducted by study team.

[24] Connected text refers to reading sentences and passages, as opposed to reading at the word level.

Dimension 6: Teacher's interaction style. Statistically significant differences were found on all seven categories of teacher interaction except discussion (Table 3.13). Intervention teachers were observed in more segments than control teachers in telling or giving information (96.6% intervention versus 87.8% control); modeling, coaching, or scaffolding (63.9% versus 57.0%); reading aloud (80.6% versus 25.9%); and assessing (61.0% versus 12.2%).

For two of these categories, the teacher interaction styles converge with patterns of findings for Dimension 5: Compared with control teachers, the intervention teachers—including computers—did more reading aloud and more assessing, which is consistent with the listening to connected text and assessment findings for literacy activities.

In more segments, control group teachers used recitation (68.6% control versus 22.0% intervention) and other styles of interaction (71.1% versus 47.5%).

Table 3.13 Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Dimension 6: Teacher's Interaction Style, by Study Condition

	Intervention		Control			
Teacher's interaction style	**Number**	**Percent**	**Number**	**Percent**	$\chi^2(df = 1)$	***p* value**
Telling or giving information	399	96.6	380	87.8	22.70	.00
Recitation	91	22.0	297	68.6	184.54	.00
Discussion	5	1.2	5	1.2	0.01	.94
Modeling, coaching, or scaffolding	264	63.9	247	57.0	4.18	.04
Reading aloud	333	80.6	112	25.9	254.26	.00
Assessment	252	61.0	53	12.2	218.14	.00
Other	196	47.5	308	71.1	49.19	.00
Total 5-minute segments	413		433			

Note: Percentages do not sum to 100 because teachers could be observed using more than one style in a 5-minute segment. Analysis includes segments from both classroom observations for each teacher.
Source: Classroom observations conducted by study team.

Dimension 7: Expected student response. Statistically significant differences were found between the intervention and control groups in all expected student responses except manipulating (Table 3.14). In more segments, more students in the intervention group than in the control group were expected to read (68.3% intervention versus 52.4% control), listen (98.1% versus 90.5%), write (80.4% versus 54.7%), and use another response (63.0% versus 14.1%).

These differences may again reflect *Thinking Reader* elements and findings for other dimensions. That intervention students were expected to read and listen in more segments than control students is consistent with the intervention and with the findings on listening to connected text and working individually. The expectation for writing is consistent with the program prompts to write responses to computerized strategy questions and quizzes and with the assessment findings for literacy activities.

For more segments, control group students were expected to talk (79.0% control versus 29.5% intervention). Intervention group students might have had fewer opportunities to talk if they spent many observation segments working individually on the computer.

Table 3.14 Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Dimension 7: Expected Student Response, by Study Condition

	Intervention		Control			
Student response	**Number**	**Percent**	**Number**	**Percent**	$\chi^2(df = 1)$	***p* value**
Reading	282	68.3	227	52.4	22.17	.00
Talking	122	29.5	342	79.0	208.67	.00
Listening	405	98.1	392	90.5	21.98	.00
Writing	332	80.4	237	54.7	63.17	.00
Manipulating	0	0.00	18	4.2	17.54	.00
Other	260	62.9	61	14.1	214.37	.00
Total 5-minute segments	413		433			

Note: Percentages do not sum to 100 because more than one type of expected student response could be observed in a 5-minute segment. Analysis includes segments from both classroom observations for each teacher.
Source: Classroom observations conducted by study team.

Study Limitations

Caution should be exercised in interpreting the observation data for the following reasons:

- Only a subset of classrooms was observed. These classrooms might not have been representative of every classroom in the sample.
- No videotaped lessons of *Thinking Reader* classrooms were available for observer training. Observers were trained to use the protocol for both intervention and control classrooms by observing face-to-face classroom activities that did not involve sustained computer use. Observers were trained to look for the range of codes that applied specifically to *Thinking Reader* activities, but the study team did not formally assess inter-rater reliability among observers for the intervention condition.
- The CIERA measure captured whether particular practices occurred during observation segments, but the CIERA codes do not reflect the quality of instruction. Observers documented behavior but did not evaluate its quality. This is especially true for *Thinking Reader* classrooms; the data do not reveal much about the specific use of the software and do not allow for conclusions on whether students or teachers made “the best” use of the software.
- When observers called to schedule their visits to classrooms, they asked to see “typical” *Thinking Reader* instruction and “typical” reading instruction in control classes; however, observers could not confirm that activities on those days were indeed typical.
- As specified by the CIERA measure, the codes at Dimension 7 represent *expected* student responses to teacher interaction (Dimension 6), not the *actual* observed student response.

The observations were designed to capture more than general or prevailing classroom activities. Codes were recorded for every applicable literacy activity during a segment even if just one student was engaged in the activity, which could occur when students are self-paced and working individually—such as during *Thinking Reader* computer instruction.

The CIERA measure attempts to capture the activities and interactions of students in a classroom (even if these vary), and observers were encouraged to walk around the classroom to capture possible differences in instruction. However, the data are limited by the inability to capture what every student is doing all the time during a classroom observation.

Summary

The program usage data gathered by the electronic worklogs indicate that the intensity of the intervention was weaker and more diffuse than recommended. Students used the *Thinking Reader* program for fewer minutes per week on average than the recommended 110–165 minutes—60 minutes (Book 1), 56 minutes (Book 2), and 42 minutes (Book 3). The average number of weeks per book—8.3 weeks (Book 1), 7.1 weeks (Book 2), and 1.7 weeks (Book 3)—also differed from the recommended 4–6 weeks.

Two classroom observations were conducted during the course of the school year for a subset of classrooms. Data from these indicate that intervention teachers did not follow the recommended three-phase instructional routine in 80% of the observed lessons. The observations found statistically significant differences between the intervention and control conditions on 47 of 57 measured classroom variables, indicating that the use of *Thinking Reader* altered the instruction observed in the intervention and control groups.

Chapter 4.
Impact Results and Sensitivity Analysis

This chapter describes the results of the impact analysis and the sensitivity analyses conducted to determine how much the impact estimates depended on the assumptions made. Members from the AIR evaluation study team conducted all analyses; the software developer and distributor had no role in this phase of the study.

Impact Results

The results presented here are based on an intent-to-treat analysis, using the original randomized and eligible sample. The intent-to-treat approach of “as randomized, as analyzed” uses the complete dataset and avoids bias that might be caused by removing cases that do not reflect program implementation as intended.

The tables display the standard errors and *p* values for each impact estimate. The standard error indicates the magnitude of the uncertainty about the true mean of each impact, given the number of schools, teachers, and students in the analysis. The *p* value indicates the chance of obtaining an impact as large as the estimated impact if no true impact existed. Results are considered statistically significant if the observed *p* value is lower than .025, indicating a less than 5% chance of obtaining the estimated impact if no true program effect existed. Multilevel models acknowledging clustered data structures were used to estimate the intervention’s effect (see Appendix C3 for a detailed description of the three-level impact model).

Table 4.1 displays the regression-adjusted group means, by study condition, for reading vocabulary and reading achievement—the two primary research questions about the effect of *Thinking Reader* on student achievement. The intervention and control group students attained similar gains from pretest to posttest on the two primary outcomes as demonstrated by the lack of statistically significant differences between the groups, the inclusion of zero in each of the 95% confidence intervals, and the small effect sizes.

Although the 32 schools were located in three states and varied in demographic characteristics, none of the parameters capturing the variation in treatment effects across schools was statistically significant. The estimate for the treatment variance for the Gates-MacGinitie Reading Tests (GMRT) was 12.16 (*p* value of .09) for the reading vocabulary subtest and 17.50 (*p* value of .30) for the reading comprehension subtest. These results indicate that intervention effects did not vary enough from the average intervention effect across schools for the difference to reach statistical significance. The multilevel model results for the GMRT reading vocabulary subtest and the GMRT reading comprehension subtest are presented in Table C3.1 in Appendix C3.

Table 4.1 Reading Achievement Posttest Scores, by study Condition

Primary outcomes: Gates-MacGinitie Reading Tests extended scale scores	Intervention (*n* = 1,286)		Control (*n* = 1,121)		Estimated impact (standard error)	*p* value	95% confidence interval[a]	Effect size[b]
	Number	Mean	Number	Mean				
Reading vocabulary	1,156	515.75	991	516.99	–1.24 (1.31)	.35	–4.17 to 1.69	–0.04
Reading comprehension	1,154	507.42	986	506.52	0.90 (1.72)	.61	–2.95 to 4.75	0.03

Note: This table presents regression-adjusted means.
[a]The 95% confidence interval is adjusted for multiple comparisons and uses the critical value of $z = 2.24$.
[b]Standardized difference is computed by using the following pooled standard deviation of posttest: reading vocabulary, 34.86; reading comprehension, 33.70.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Similarly, no statistically significant differences were found between groups for the Metacognitive Awareness of Reading Strategies Inventory (MARSI) and Motivation for Reading Questionnaire (MRQ) measures, which address the two ancillary research questions about the effect of *Thinking Reader* on students' approaches to reading. Table 4.2 displays the regression-adjusted group means by condition for the two measures. Again, the intervention effect did not vary significantly between conditions. The estimated treatment variance is 0.01 (*p* value of .33) for the MARSI and 0.00 (*p* value of .14) for the MRQ. The multilevel model results for the MARSI and MRQ are shown in Table C3.2 in Appendix C3.

Table 4.2 Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Posttest Scores, by Study Condition

	Intervention (*n* = 1,286)		Control (*n* = 1,121)		Estimated impact (standard error)	*p* value	95% confidence interval	Effect size[a]
Ancillary outcomes	Number	Mean	Number	Mean				
Reading strategies: MARSI (1–5 scale)	1,181	3.09	1,020	3.09	–0.00 (0.04)	.99	–0.08 to 0.08	–0.00
Reading motivation: MRQ (1–4 scale)	1,183	2.76	1,025	2.77	–0.01 (0.03)	.62	–0.06 to 0.04	–0.03

Note: This table presents regression-adjusted means. MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
[a]Standardized difference is computed by using the following pooled standard deviation of posttest: MARSI, 0.73; MRQ, 0.50.
Source: MARSI and MRQ surveys administered by study team.

Results of Sensitivity Analyses

Several sensitivity analyses were conducted to determine how much the impact estimates presented in Tables 4.1 and 4.2 depended on the assumptions made. The direction and magnitude of the treatment effects and the overall conclusions did not change under the following conditions: using multiple imputation of missing data, modeling treatment as a fixed effect, using a two-level impact model, reducing the sample to one classroom per teacher, and removing students whose teachers left the study at the beginning of the implementation period. See Appendix D for detailed information.

Chapter 5.
Exploratory Analysis

This chapter presents the exploratory analyses that investigate whether the impact of the *Thinking Reader* intervention varied for different subgroups of students. The subgroups of students we examined were formed on the basis of baseline achievement and motivation to read measures. We answer the following four exploratory research questions:

1. Does the effect of *Thinking Reader* on students' reading vocabulary vary according to their baseline reading vocabulary scores?

2. Does the effect of *Thinking Reader* on students' reading comprehension vary according to their baseline reading comprehension scores?

3. Does the effect of *Thinking Reader* on students' reading vocabulary vary according to their baseline reading motivation scores?

4. Does the effect of *Thinking Reader* on students' reading comprehension vary according to their baseline reading motivation scores?

The outcomes of interest for these exploratory research questions are the vocabulary and comprehension subtests of the Gates-MacGinitie Reading Tests (GMRT) that served as the achievement measures for the primary research questions.

Exploratory Questions 1 and 2 investigate whether the impacts on the primary outcomes (vocabulary and comprehension) varied for subgroups of students formed on the basis of baseline measures of these outcomes. While a subgroup analysis of this sort is important to explore for a wide variety of interventions, it is particularly appropriate for *Thinking Reader* because the program intends to provide differentiated support to students with different skill levels. Helping teachers monitor progress and facilitating teachers' individualization of instruction based on student skills are distinct features of the *Thinking Reader* program. As described in Chapter 3, teachers can customize the amount of program support given to each student. Exploratory Research Questions 1 and 2 are especially important empirical questions because the literature lacks evidence to indicate whether an intervention like *Thinking Reader* might be more or less beneficial for students with lower or higher baseline achievement scores.

Exploratory Research Questions 3 and 4 examine whether impacts on the primary outcomes varied for subgroups of students formed on the basis of a baseline measure of reading motivation, as measured with the Motivation for Reading Questionnaire (MRQ). In theory, an interactive software program might be motivating for adolescent readers, but strong evidence does not exist to indicate whether an intervention like *Thinking Reader* might be more or less beneficial for students with lower or higher self-reported baseline motivation levels.

For the exploratory analyses, we partitioned the baseline achievement scores into subgroups. The sample was divided into three groups representing the lowest, middle, and highest achieving tertiles. The sample of students with baseline achievement scores was split into tertiles (as opposed to quartiles or quintiles) to avoid having cells with very small sample sizes.[25] We calculated the power of these analyses and minimum detectable effect size, and these calculations are presented in Appendix E.

In order to discover any non-linear effect of the treatment across the baseline achievement score distribution, we used subgroups rather than the continuous score distribution for these analyses. By partitioning the pretest scores into tertiles and generating interaction between the tertiles and the intervention indicator, we allowed the effect of the intervention to be conditional on the baseline achievement scores and different in each tertile. This specification is more flexible than using the continuous form of the pretest score and could produce more relevant findings in light of the main impact results.[26] For example, the intervention could be ineffective for the majority of students (and could drive the overall results) but the subgroup results may be able to detect a non-zero effect on a smaller subset of students.

When the intervention indicator is interacted with the continuous form of the baseline achievement scores, this specification tests, on average, whether the effect of the intervention varies linearly as a function of the baseline achievement scores. Thus, the intervention could have a non-zero effect in certain parts of the baseline scores distribution but not in others, and those different effects would not be detected.

To ensure that the subgroups created using baseline achievement are distinct from the subgroups created using baseline motivation, we explored the correlation between these baseline measures. These correlations are positive and statistically significant, although not large in magnitude. The correlation between the pretest vocabulary and pretest MRQ scores is .105, and the correlation between pretest comprehension and pretest MRQ scores is .156. The magnitudes of these correlations indicate that, at baseline, students who performed higher on achievement measures are not necessarily the same students who reported higher motivation to read. This means that Questions 3 and 4 are not simply variations of Questions 1 and 2.

Analytic Approach

Similar to the main impact models and because of the study research design and hierarchical data structure (students nested within teachers and teachers nested within schools), a three-level multilevel model using listwise deletion samples was used to estimate the exploratory research questions. This section describes in more detail the analyses used to answer each of the exploratory research questions.

Exploratory Research Questions 1 and 2. To address research Questions 1 and 2, we generated dummy indicators capturing whether each student's baseline achievement score was in the lowest, middle, or highest tertiles of the pretest score distribution. For each outcome, tertiles were generated using information from all of the students with pretest score data.[27] The

[25] The original power analysis was not calculated to accommodate dividing the sample into subgroups.

[26] When interaction effects are tested with continuous baseline achievement scores, this tests whether, on average, the effect of the intervention varies linearly as a function of the baseline achievement scores. Thus, the intervention could have a non-zero effect in certain parts of the baseline scores distribution but not in others, and those differential effects would not be detected.

[27] The stata command used to generate the tertiles *xtile* includes within each tertile the upper bound values of the tertile (intervals are defined as semi-closed). Therefore, Tertiles 1 and 2 include the upper bound values that

indicators distinguishing students in the lowest and middle tertiles were included in the Level 1 equation, using the dummy indicator capturing students in the highest tertile as the reference group. To determine whether impacts on the primary outcomes varied for subgroups of students formed on the basis of baseline measures of these outcomes, cross-level interaction products between the lowest tertile and the treatment indicator and between the middle tertile and the treatment indicator were included in Level 2. The multilevel equations are presented in Appendix E. Table 5.1 presents descriptive statistics of the baseline reading achievement pretest scores by tertile.

Table 5.1 Reading Achievement Pretest Scores, by Baseline Achievement Tertile[28]

Baseline achievement tertiles	N	Percent	Mean	Standard deviation	Minimum	Maximum
Reading vocabulary						
Tertile 1	808	33.84	467.55	17.93	367	486
Tertile 2	822	34.42	501.64	7.87	490	514
Tertile 3	758	31.74	539.61	22.34	517	653
Reading comprehension						
Tertile 1	879	36.81	468.52	16.4	396	488
Tertile 2	725	30.36	501.55	6.83	491	512
Tertile 3	784	32.83	537.07	19.24	516	652

Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Exploratory Research Questions 3 and 4. To address Exploratory Research Questions 3 and 4, we created indicators capturing whether each student's baseline motivation to read score was in the lowest, middle, or highest tertiles of the distribution. The indicators distinguishing students in the lowest and middle tertiles were included in the Level 1 equation, using the dummy capturing students in the highest tertile as the reference group. To explore whether impacts on the primary outcomes varied for subgroups of students formed on the basis of the baseline motivation to read measure, cross-level interaction products between the lowest tertile and the treatment indicator and between the middle tertile and the treatment indicator are included in Level 2. Table 5.2 shows the descriptive statistics of the baseline motivation to read measure by baseline reading motivation tertile.

Table 5.2 Motivation to Read Pretest Scores, by Baseline Reading Motivation Tertile[29]

Baseline motivation tertiles	N	Percent	Mean	Standard deviation	Minimum	Maximum
Tertile 1	799	33.47	2.31	0.32	1.19	2.67
Tertile 2	808	33.85	2.89	0.11	2.67	3.08
Tertile 3	780	32.68	3.33	0.19	3.08	3.96

Source: Motivation for Reading Questionnaire survey administered by study team.

represent the cumulative percentages of 33% and 66%, respectively. For example in the case of the reading vocabulary pretest, the cumulative percentage of 29.9% obtained scores of 483 or lower. The cumulative percentage up to the subsequent score of 486 is 33.84%. The upper bound of Tertile 1 includes the score of 486. In the section that describes the sensitivity analysis, we also defined the upper bound of the tertiles' intervals as open.

[28] The descriptive statistics presented in this table are based on all of the students with pretest achievement scores.

[29] The descriptive statistics presented in this table are based on all of the students with pretest motivation to read scores.

Exploratory Impact Results

In this section we present two sets of results. The first results show whether the *Thinking Reader* impacts on each tertile are statistically different from each other. The second results reveal whether tertile impacts are statistically different from zero.

Exploring whether tertile impacts differ from each other. Table 5.3 presents the estimated interaction effect coefficients for all four exploratory research questions, computed using the multilevel models described in Appendix E. This table also displays the standard error, *p* value, and 95% confidence interval for each of the interaction terms.

Results indicate that 11 of the 12 contrasts are not statistically significant (see Table 5.3). The only significant contrast suggests that *Thinking Reader* had a positive effect for Tertile 1 (lowest achieving group) relative to the effect on Tertile 2 (middle achieving group) equal to 5.77, *p* = .03. This significant interaction reveals that the subgroup impacts of Tertiles 1 and 2 are different from each other and captures the fact that the direction of the effect of *Thinking Reader* is positive in Tertile 1, and negative in Tertile 2, as described in the next section.

Table 5.3 Interaction Estimates Obtained From Multilevel Models

Research question	Primary outcomes: Gates-MacGinitie Reading Tests extended scale scores	Contrast	Estimated interaction effect (standard error)	*p* value	95% confidence interval[a]
Subgroups of students formed on the basis of baseline achievement					
Q1	Reading vocabulary	Tertile 1 vs. Tertile 3	3.19 (2.57)	.21	–1.85 to 8.23
		Tertile 2 vs. Tertile 3	–0.32 (2.42)	.90	–5.06 to 4.42
		Tertile 1 vs. Tertile 2	3.51 (2.51)[b]	.16	–1.41 to 8.43
Q2	Reading comprehension	Tertile 1 vs. Tertile 3	2.69 (2.69)	.32	–2.58 to 7.96
		Tertile 2 vs. Tertile 3	–3.07 (2.65)	.25	–8.26 to 2.12
		Tertile 1 vs. Tertile 2	5.77 (2.69)[b]	.03	0.50 to 11.04
Subgroups of students formed on the basis of baseline motivation					
Q3	Reading vocabulary	Tertile 1 vs. Tertile 3	–1.15 (2.13)	.59	–5.32 to 3.02
		Tertile 2 vs. Tertile 3	0.43 (2.06)	.84	–3.61 to 4.47
		Tertile 1 vs. Tertile 2	–1.58 (2.09)[b]	.45	–5.68 to 2.52
Q4	Reading comprehension	Tertile 1 vs. Tertile 3	–2.37 (2.50)	.35	–7.27 to 2.53
		Tertile 2 vs. Tertile 3	–2.04 (2.42)	.40	–6.78 to 2.70
		Tertile 1 vs. Tertile 2	–0.33 (2.45)[b]	.89	–5.13 to 4.47

[a] The 95% confidence interval is not adjusted for multiple comparisons and uses the critical value of 1.96.
[b] The standard error for this contrast was obtained by changing the reference group from Tertile 3 to Tertile 2 in the multilevel model.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Exploring whether tertile impacts differ from zero. To interpret the interaction coefficients presented in Table 5.3, we computed the estimated treatment impact for each tertile and tested whether the tertile impacts were statistically different from zero. Table 5.4 presents the adjusted posttest means by study condition and for each tertile separately. Table 5.4 also includes the estimated impact, its standard error, *p* value, and 95% confidence interval. The estimated impact presented in Column 8 represents the regression of posttest score on the treatment indicator at different tertiles.[30] The results in Table 5.4 reveal that the direction of the effect of *Thinking Reader* on Tertile 1 (lowest achieving group) was positive but not statistically significant. The coefficients for treatment are 2.58 (*SE* = 1.91) for reading vocabulary and 2.15 (*SE* = 2.22) for reading comprehension. In achievement Tertiles 2 and 3, the direction of the estimated *Thinking Reader* effect was negative, but again, neither was statistically significant. The Tertiles 2 and 3 effects are –0.93 (*SE* = 1.79) and –0.61 (*SE* = 1.85) for reading vocabulary, and –3.61 (*SE* = 2.22) and –0.54 (*SE* = 2.19) for reading comprehension. For reading comprehension, the positive and negative effects of *Thinking Reader* on Tertiles 1 and 2 produced the statistically significant interaction described in Table 5.3.

For the tertiles formed on the basis of baseline motivation to read scores, the estimated effects of the *Thinking Reader* intervention were in a negative direction and not statistically different from zero.

Table 5.4 Reading Achievement Adjusted Posttest Scores, by Study Condition and Subgroups of Students Formed on the Basis of Baseline Achievement and Motivation to Read Scores

Research question	Primary outcomes: Gates-MacGinitie Reading Tests extended scale scores	Tertile	Intervention (*n* = 1,286)		Control (*n* = 1,121)		Estimated impact (standard error)[a]	*t* value	*p* value	95% confidence interval[b]
			Number	Mean	Number	Mean				
Subgroups of students formed on the basis of baseline achievement										
Q1	Reading vocabulary	Tertile 1	425	461.38	383	458.8	2.58 (1.91)	1.35	.18	−1.16 to 6.32
		Tertile 2	460	483.99	362	484.92	−0.93 (1.79)	−0.52	.60	−4.44 to 2.58
		Tertile 3	391	516.05	367	516.66	−0.61 (1.85)	−0.33	.74	−4.24 to 3.02
Q2	Reading comprehension	Tertile 1	489	459.29	390	457.14	2.15 (2.22)	0.97	.33	−2.20 to 6.50
		Tertile 2	370	474.83	355	478.44	−3.61 (2.22)	−1.63	.10	−7.96 to 0.74
		Tertile 3	421	505.91	363	506.45	−0.54 (2.19)	−0.25	.81	−4.83 to 3.75
Subgroups of students formed on the basis of baseline motivation to read scores										
Q3	Reading vocabulary	Tertile 1	438	512.50	357	514.78	−2.28 (1.62)	−1.41	.16	−5.46 to 0.90
		Tertile 2	418	515.10	387	515.80	−0.70 (1.60)	−0.44	.66	−3.84 to 2.44
		Tertile 3	412	515.89	363	517.02	−1.13 (1.61)	−0.70	.48	−4.29 to 2.03
Q4	Reading comprehension	Tertile 1	440	502.15	355	503.73	−1.58 (2.06)	−0.77	.44	−5.62 to 2.46
		Tertile 2	420	505.30	386	506.55	−1.25 (2.04)	−0.61	.54	−5.25 to 2.75
		Tertile 3	414	507.43	362	506.64	0.79 (2.04)	0.39	.70	−3.21 to 4.79

Note: The table presents regression-adjusted means.

[a] Standard errors were calculated using this formula: $SE_b = [VAR(treat_coeff) + 2TertileZ*COV(treat_coeff, TertileZ_coeff) + TertileZ^2 *VAR (tertileZ_coeff)]^{1/2}$where TertileZ takes the value of 1 for each of the three tertiles. The variance components for the coefficients used in this formula were obtained from the gamvc.dat file generated by HLM 6.8.

[b] The 95% confidence interval is not adjusted for multiple comparisons and uses the critical value of 1.96.

Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

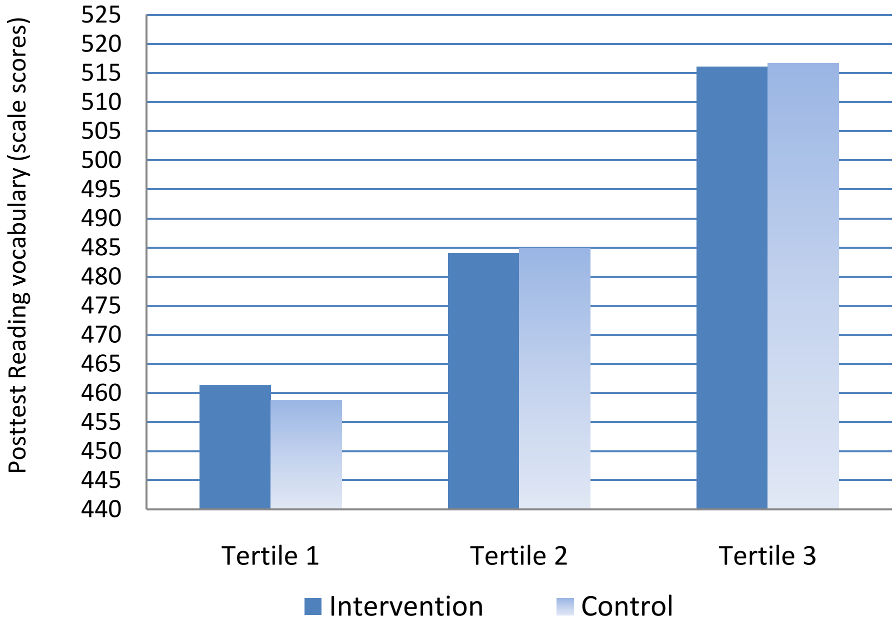
The results presented in Table 5.4 are illustrated graphically in Figures 5.1, 5.2, 5.3, and 5.4 for Exploratory Research Questions 1, 2, 3, and 4, respectively.

[30] These effects were obtained by combining the estimates of the intercept, the treatment coefficient, the tertile's coefficient, and the coefficients of the two interaction products. Results presented in Tables 5.3 and 5.4 for each outcome are obtained from the same multilevel model presented in Appendix E.

Figure 5.1 illustrates that in Tertile 1, intervention students scored 2.58 points higher on the reading vocabulary posttest than control students, while in Tertiles 2 and 3, control students scored less than a point higher than intervention students. But again, none of these differences were statistically significant.

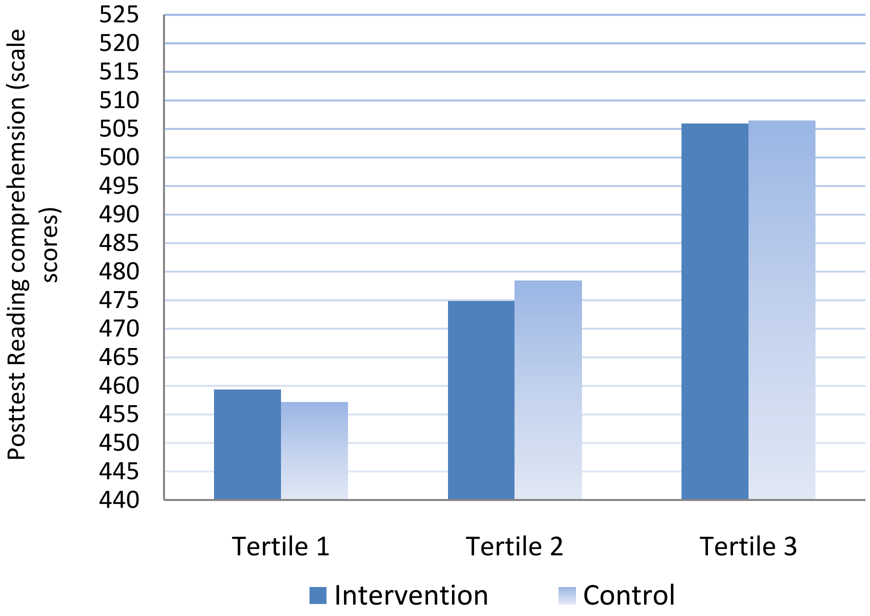
Similarly for reading comprehension, Figure 5.2 illustrates that in Tertile 1, intervention students scored 2.15 points higher on the reading comprehension posttest than control students, while in Tertiles 2 and 3, control students scored 3.61 and 0.54 points higher than intervention students, respectively. Although these differences are not statistically significant, the difference in the direction of the treatment effect in comprehension for Tertile 1 (positive, in favor of intervention) and Tertile 2 (negative, in favor of control) yielded the statistically significant interaction described in Table 5.3 (see Table 5.3, row 6).

Figure 5.1 Adjusted Posttest Reading Vocabulary Score Means, by Study Condition and Achievement Tertile (Exploratory Research Question 1)



Source: Gates-MacGinitie Reading Tests vocabulary subtest administered by study team.

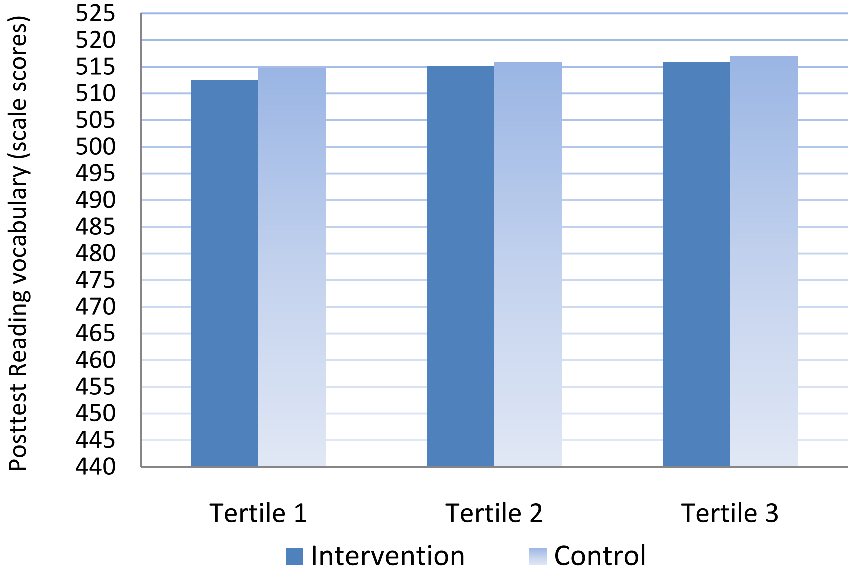
Figure 5.2 Adjusted Posttest Reading Comprehension Score Means, by Study Condition and Achievement Tertile (Exploratory Research Question 2)



Source: Gates-MacGinitie Reading Tests comprehension subtest administered by study team.

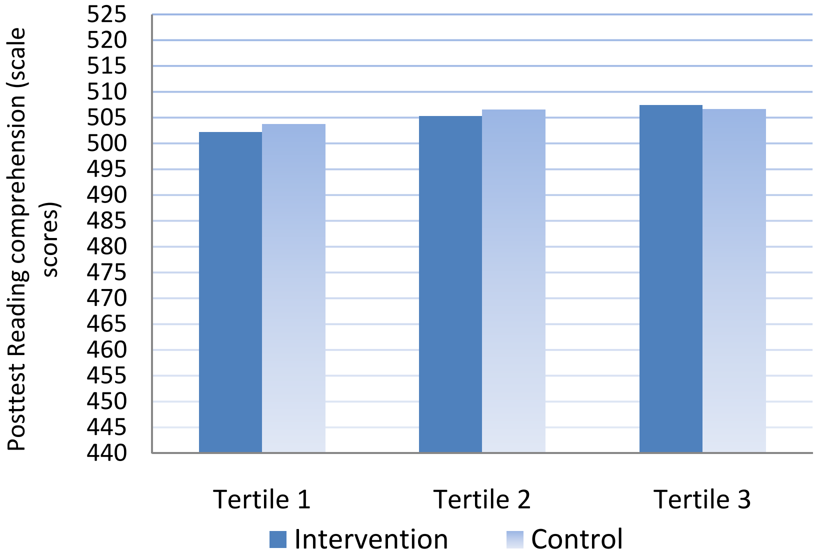
Figures 5.3 and 5.4 present the adjusted posttest means for Exploratory Research Questions 3 and 4. Differences presented in these graphs are not statistically significant.

Figure 5.3 Adjusted Posttest Reading Vocabulary Score Means, by Study Condition and Achievement Tertile (Exploratory Research Question 3)



Source: Gates-MacGinitie Reading Tests vocabulary subtest administered by study team.

Figure 5.4 Adjusted Posttest Reading Comprehension Score Means, by Study Condition and Achievement Tertile (Exploratory Research Question 4)



Source: Gates-MacGinitie Reading Tests comprehension subtest administered by study team.

The full multilevel model results for Exploratory Research Questions 1, 2, 3, and 4 are presented in Appendix E (see Tables E3.1 and E3.2).

Summary of the Results

In this chapter, we explored whether the *Thinking Reader* impacts were statistically different across subgroups defined by baseline reading comprehension and vocabulary scores and by a baseline motivation to read measure. We also looked at whether impacts for individual tertiles were statistically different from zero.

Results from the multilevel analyses testing the exploratory research questions revealed that 11 of 12 interactions tested across the four exploratory research questions were not significantly different from each other. None of the interaction terms formed between the intervention indicator and the baseline reading vocabulary achievement tertiles, and the baseline motivation to read tertiles was statistically significant. For the reading comprehension outcome, we found one statistically significant interaction (5.77, *p* = .03). This interaction revealed that the effect of the intervention was different for the lowest achieving group (Tertile 1) compared to the middle achieving group (Tertile 2) and results from the finding that in Tertile 1, intervention students performed 2.15 points higher than control students and in Tertile 2, control students outperformed intervention students by 3.61 points. However, the statistically significant interaction is difficult to interpret because it was found only for reading comprehension and for the contrast that compares the effect of the intervention between Tertiles 1 and 2. Because of the large number of post hoc analyses involved in the exploratory analyses, the statistically significant interaction may be due to chance.

Furthermore, when testing whether the impact effect of the intervention was different from zero in each tertile, we found that the *Thinking Reader* program had no statistically significant effects on any of the tertiles formed on the basis of pretest measures.

These results reveal a lack of evidence supporting the hypothesis that the *Thinking Reader* program might have a differential impact across different subgroups of students formed on the basis of baseline achievement and motivation to read.

Results of Exploratory Sensitivity Analyses

We performed two sensitivity analyses to explore the robustness of the results to changes in the size and composition of the tertiles. For the sensitivity analyses, the tertiles were defined on the basis of the study sample (Sensitivity Analysis 1) and on an external benchmark used by the publisher of the Gates-MacGinitie Reading Tests (Sensitivity Analysis 2) (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2007).

Results for Sensitivity Analysis 1 reveal similar findings to the exploratory impact results. The sign and magnitude of the interaction terms are very similar, although the magnitude of the point estimates obtained in Sensitivity Analysis 1 tend to be larger than the exploratory impact results. In Sensitivity Analysis 2, the exploratory impact results are no longer replicated. The difference in the distribution between the study sample and the sample used for test norming helps to explain why the results of the second sensitivity analysis are different from those found in the main exploratory analysis. Detailed information about these analyses is presented in Appendices E4 and E5.

Chapter 6.
Summary of Key Findings

This study indicates that Grade 6 students with teachers randomly assigned to use *Thinking Reader* for one school year performed no better on primary achievement outcomes of reading vocabulary and reading comprehension than students whose teachers followed the standard reading curriculum (a business-as-usual approach). The two groups also showed no statistically significant differences on two ancillary self-report surveys that measured the use of reading comprehension strategies and the motivation to read. The study results were robust to various assumptions that were made; the direction and magnitude of the intervention effects and overall conclusions did not change when sensitivity analyses were conducted.

With regard to the exploratory research questions detailed in Chapter 5, the *Thinking Reader* program had no statistically significant effects on any of the subgroups formed on the basis of baseline achievement and motivation to read measures. In other words, these results confirmed the impact findings. Exploratory findings suggest that no strong evidence is available to support the hypotheses that the *Thinking Reader* program might have differential impact effects on students from different achievement and motivation to read subgroups. Eleven of 12 interactions tested across the four exploratory research questions were not statistically significant. For the reading comprehension outcome, we found one statistically significant interaction (5.77, $p = .03$). This interaction results from the fact that in the lowest tertile, intervention students performed 2.15 points higher than control students, and in the middle tertile, control students outperformed intervention students by 3.61 points. However, this statistically significant interaction is difficult to interpret because this finding was not replicated on any other contrast or outcome. Because of the large number of post hoc analyses involved in the exploratory analyses, this statistically significant interaction may be due to chance.

The classroom observation data—conducted at two separate times on a subset of classrooms—showed statistically significant differences between the intervention and control classrooms on 47 of 57 measured classroom variables, indicating that when *Thinking Reader* was in use, the nature of instruction differed between the intervention and control groups during the period of classroom observation. However, the classroom observation data did not reveal more about why the software showed no effects.

The software was not always used as intended by its developer. Students used the *Thinking Reader* program for fewer minutes per week, on average, than the recommended 110–165 minutes—60 minutes for Book 1, 56 minutes for Book 2, and 42 minutes for Book 3. The number of weeks that students spent per book—on average 8.3 for Book 1, 7.1 for Book 2, and 1.7 for Book 3—differed from the recommended 4–6 weeks. Teachers made limited adjustments to students' levels of support in the books, with at least 1 out of 4 teachers not making any adjustments for each book. When teachers made adjustments, they appeared to be calibrated to some extent with students' average quiz scores. Data on students' specific navigation through and use of other specific program features were not available for analysis in this study. We also did not collect detailed information on the quality and depth of teacher implementation.

This study was the first randomized controlled trial of *Thinking Reader*. It used randomization to create comparable groups at baseline and maintained the integrity of the randomization through the end of the study. The intent-to-treat analytical approach, which analyzes participants on the basis of how they are randomly assigned, yielded unbiased estimates. The lack of statistically significant, positive achievement effects contrasts with the findings of Dalton et al. (2002) in their quasi-experimental study of *Thinking Reader* and of other studies that have found statistically significant, positive effects with interventions that used reciprocal teaching or other strategy-based methods for targeting comprehension (e.g., National Institute of Child Health and Human Development, 2000; Rosenshine & Meister, 1994; Rosenshine et al., 1996).

The study represents only one of the ways in which the program can be implemented in schools. Results of the study do not infer that the same results would be produced under other conditions. The results reported here apply to implementation of *Thinking Reader* software as a partial substitute for the regular curriculum used in a whole-group setting at Grade 6. Results also apply to a condition in which teachers were instructed to use the software for three books for 110–165 minutes per week during a 4- to 6-week period when the books were being read. The use of a volunteer sample limits the study findings to the schools, teachers, and students in Connecticut, Massachusetts, and Rhode Island that participated in the study. Results should not be generalized beyond this sample.

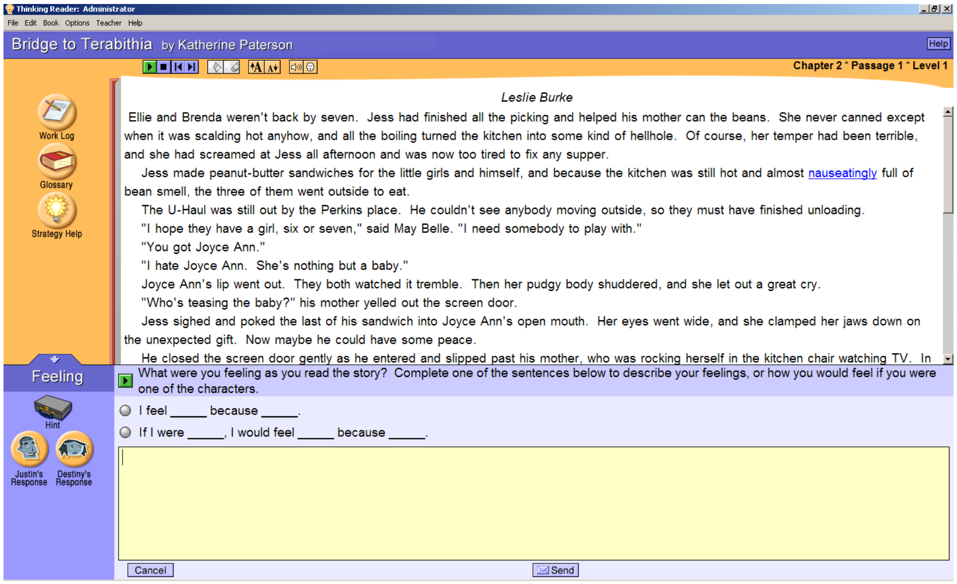
Appendix A. Examples From the *Thinking Reader* Program

The materials in this appendix provide screenshot examples from the *Thinking Reader* program.

The *Thinking Reader* software provides five levels of scaffolding support for students. The screenshot examples illustrate how the support for one passage (*Bridge to Terabithia*, Chapter 2, passage 1) (Tom Snyder Productions, 2006a) differs from level to level. As students move from Level 1 to Level 4, they receive less guidance in their strategy questions, less context in the coach-provided think aloud, and fewer hints. By Level 5, students receive no coaching and choose their own strategy to apply to the passage. Exhibits A.1–A.7 present screenshots for each of the five levels.

Exhibit A.1 Level 1 Screenshot, Close-Ended Feeling Strategy Prompt

A hint and two coach think-alouds are provided.



Thinking Reader: Administrator
File Edit Book Options Teacher Help
Bridge to Terabithia by Katherine Paterson
Help
Chapter 2 * Passage 1 * Level 1
Work Log
Glossary
Strategy Help
Feeling
Hint
Justin's Response
Destiny's Response
Cancel
Send

Leslie Burke

Ellie and Brenda weren't back by seven. Jess had finished all the picking and helped his mother can the beans. She never canned except when it was scalding hot anyhow, and all the boiling turned the kitchen into some kind of hellhole. Of course, her temper had been terrible, and she had screamed at Jess all afternoon and was now too tired to fix any supper.

Jess made peanut-butter sandwiches for the little girls and himself, and because the kitchen was still hot and almost nauseatingly full of bean smell, the three of them went outside to eat.

The U-Haul was still out by the Perkins place. He couldn't see anybody moving outside, so they must have finished unloading.

"I hope they have a girl, six or seven," said May Belle. "I need somebody to play with."

"You got Joyce Ann."

"I hate Joyce Ann. She's nothing but a baby."

Joyce Ann's lip went out. They both watched it tremble. Then her pudgy body shuddered, and she let out a great cry.

"Who's teasing the baby?" his mother yelled out the screen door.

Jess sighed and poked the last of his sandwich into Joyce Ann's open mouth. Her eyes went wide, and she clamped her jaws down on the unexpected gift. Now maybe he could have some peace.

He closed the screen door gently as he entered and slipped past his mother, who was rocking herself in the kitchen chair watching TV. In

What were you feeling as you read the story? Complete one of the sentences below to describe your feelings, or how you would feel if you were one of the characters.

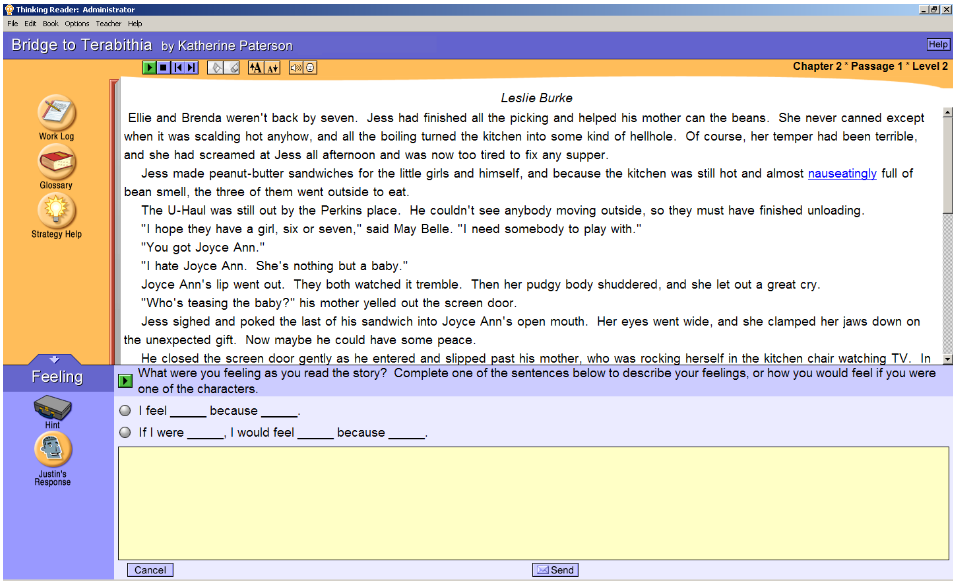
○ I feel _____ because _____.

○ If I were _____, I would feel _____ because _____.

Source: Tom Snyder Productions (2006a).

Exhibit A.2 Level 2 Screenshot, Close-Ended Feeling Strategy Prompt

A hint and one coach think-aloud are provided.


Thinking Reader: Administrator
File Edit Book Options Teacher Help
Bridge to Terabithia by Katherine Paterson
Help
Chapter 2 * Passage 1 * Level 2
Work Log
Glossary
Strategy Help
Feeling
Hint
Justin's Response
Cancel
Send

Leslie Burke

Ellie and Brenda weren't back by seven. Jess had finished all the picking and helped his mother can the beans. She never canned except when it was scalding hot anyhow, and all the boiling turned the kitchen into some kind of hellhole. Of course, her temper had been terrible, and she had screamed at Jess all afternoon and was now too tired to fix any supper.

Jess made peanut-butter sandwiches for the little girls and himself, and because the kitchen was still hot and almost nauseatingly full of bean smell, the three of them went outside to eat.

The U-Haul was still out by the Perkins place. He couldn't see anybody moving outside, so they must have finished unloading.

"I hope they have a girl, six or seven," said May Belle. "I need somebody to play with."

"You got Joyce Ann."

"I hate Joyce Ann. She's nothing but a baby."

Joyce Ann's lip went out. They both watched it tremble. Then her pudgy body shuddered, and she let out a great cry.

"Who's teasing the baby?" his mother yelled out the screen door.

Jess sighed and poked the last of his sandwich into Joyce Ann's open mouth. Her eyes went wide, and she clamped her jaws down on the unexpected gift. Now maybe he could have some peace.

He closed the screen door gently as he entered and slipped past his mother, who was rocking herself in the kitchen chair watching TV. In

What were you feeling as you read the story? Complete one of the sentences below to describe your feelings, or how you would feel if you were one of the characters.

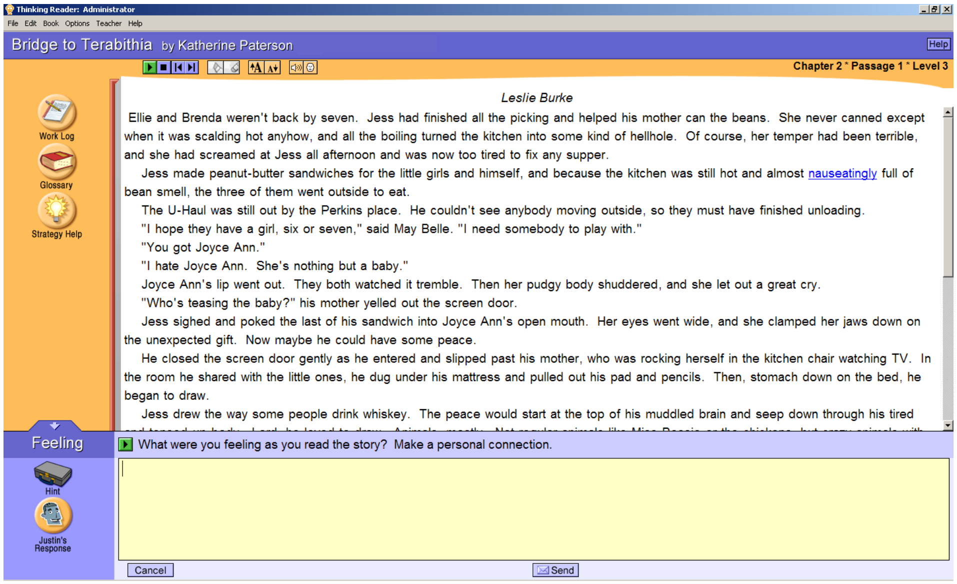
○ I feel _____ because _____.

○ If I were _____, I would feel _____ because _____.

Source: Tom Snyder Productions (2006a).

Exhibit A.3 Level 3 Screenshots, Open-Ended Feeling Strategy Prompt and Context-Specific Think-Aloud Illustration

A hint and one coach think-aloud (context-specific) are provided.


Thinking Reader: Administrator
File Edit Book Options Teacher Help
Bridge to Terabithia by Katherine Paterson
Help
Chapter 2 * Passage 1 * Level 3
Work Log
Glossary
Strategy Help
Feeling
Hint
Justin's Response
Cancel
Send

Leslie Burke

Ellie and Brenda weren't back by seven. Jess had finished all the picking and helped his mother can the beans. She never canned except when it was scalding hot anyhow, and all the boiling turned the kitchen into some kind of hellhole. Of course, her temper had been terrible, and she had screamed at Jess all afternoon and was now too tired to fix any supper.

Jess made peanut-butter sandwiches for the little girls and himself, and because the kitchen was still hot and almost nauseatingly full of bean smell, the three of them went outside to eat.

The U-Haul was still out by the Perkins place. He couldn't see anybody moving outside, so they must have finished unloading.

"I hope they have a girl, six or seven," said May Belle. "I need somebody to play with."

"You got Joyce Ann."

"I hate Joyce Ann. She's nothing but a baby."

Joyce Ann's lip went out. They both watched it tremble. Then her pudgy body shuddered, and she let out a great cry.

"Who's teasing the baby?" his mother yelled out the screen door.

Jess sighed and poked the last of his sandwich into Joyce Ann's open mouth. Her eyes went wide, and she clamped her jaws down on the unexpected gift. Now maybe he could have some peace.

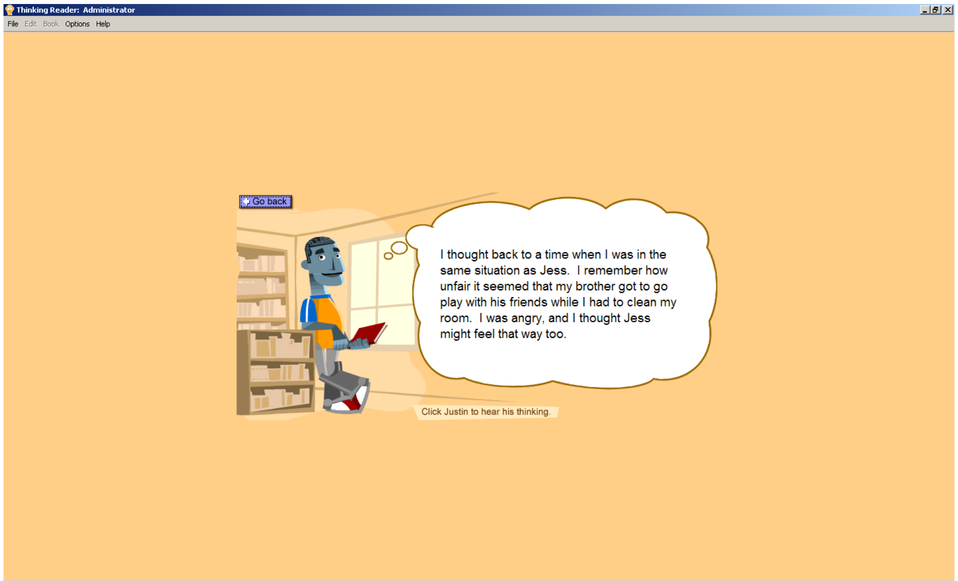
He closed the screen door gently as he entered and slipped past his mother, who was rocking herself in the kitchen chair watching TV. In the room he shared with the little ones, he dug under his mattress and pulled out his pad and pencils. Then, stomach down on the bed, he began to draw.

Jess drew the way some people drink whiskey. The peace would start at the top of his muddled brain and seep down through his tired

What were you feeling as you read the story? Make a personal connection.

Source: Tom Snyder Productions (2006a).

Exhibit A.4 Level 3 Screenshots, Open-Ended Feeling Strategy Prompt and Context-Specific Think-Aloud Illustration *(continued)*

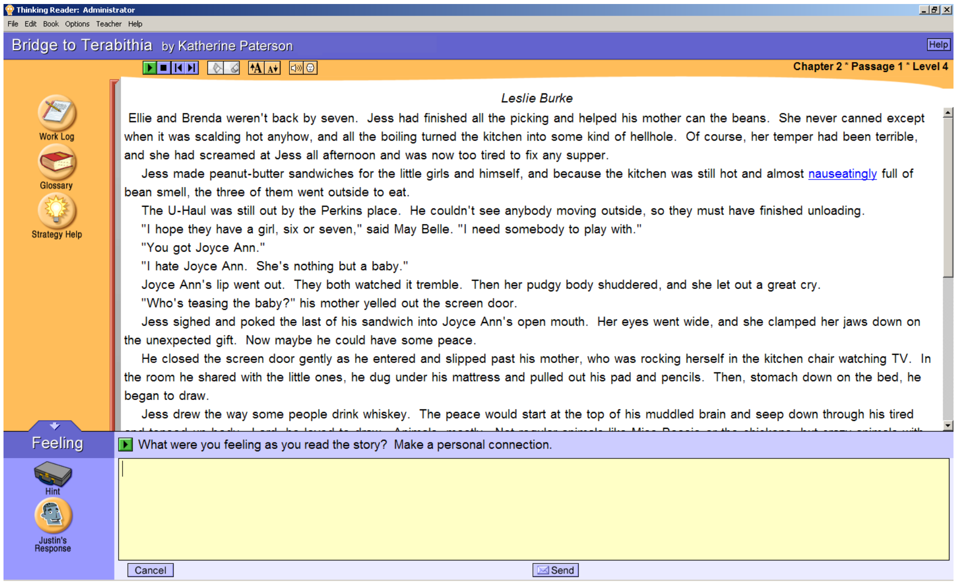


I thought back to a time when I was in the same situation as Jess. I remember how unfair it seemed that my brother got to go play with his friends while I had to clean my room. I was angry, and I thought Jess might feel that way too.

Click Justin to hear his thinking.

Source: Tom Snyder Productions (2006a).

Exhibit A.5 Level 4 Screenshots, Open-Ended Strategy Prompt and General Think-Aloud Illustration


Thinking Reader: Administrator
File Edit Book Options Teacher Help
Bridge to Terabithia by Katherine Paterson
Help
Chapter 2 * Passage 1 * Level 4
Work Log
Glossary
Strategy Help
Feeling
Hint
Justin's Response
Cancel
Send

Leslie Burke

Ellie and Brenda weren't back by seven. Jess had finished all the picking and helped his mother can the beans. She never canned except when it was scalding hot anyhow, and all the boiling turned the kitchen into some kind of hellhole. Of course, her temper had been terrible, and she had screamed at Jess all afternoon and was now too tired to fix any supper.

Jess made peanut-butter sandwiches for the little girls and himself, and because the kitchen was still hot and almost nauseatingly full of bean smell, the three of them went outside to eat.

The U-Haul was still out by the Perkins place. He couldn't see anybody moving outside, so they must have finished unloading.

"I hope they have a girl, six or seven," said May Belle. "I need somebody to play with."

"You got Joyce Ann."

"I hate Joyce Ann. She's nothing but a baby."

Joyce Ann's lip went out. They both watched it tremble. Then her pudgy body shuddered, and she let out a great cry.

"Who's teasing the baby?" his mother yelled out the screen door.

Jess sighed and poked the last of his sandwich into Joyce Ann's open mouth. Her eyes went wide, and she clamped her jaws down on the unexpected gift. Now maybe he could have some peace.

He closed the screen door gently as he entered and slipped past his mother, who was rocking herself in the kitchen chair watching TV. In the room he shared with the little ones, he dug under his mattress and pulled out his pad and pencils. Then, stomach down on the bed, he began to draw.

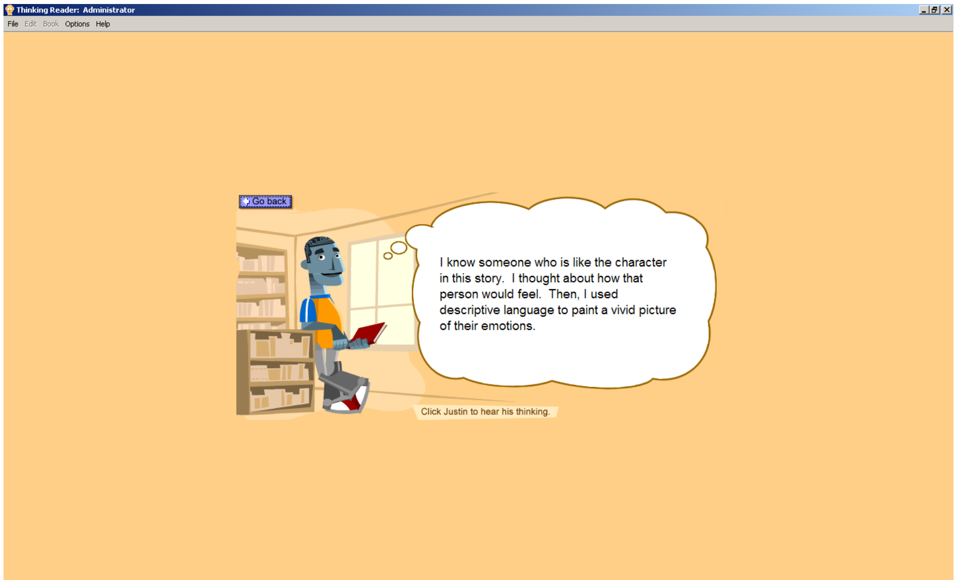
Jess drew the way some people drink whiskey. The peace would start at the top of his muddled brain and seep down through his tired

What were you feeling as you read the story? Make a personal connection.

A hint and one coach think-aloud (context-general) are provided.

Source: Tom Snyder Productions (2006a).

Exhibit A.6 Level 4 Screenshots, Open-Ended Strategy Prompt and General Think-Aloud Illustration *(continued)*

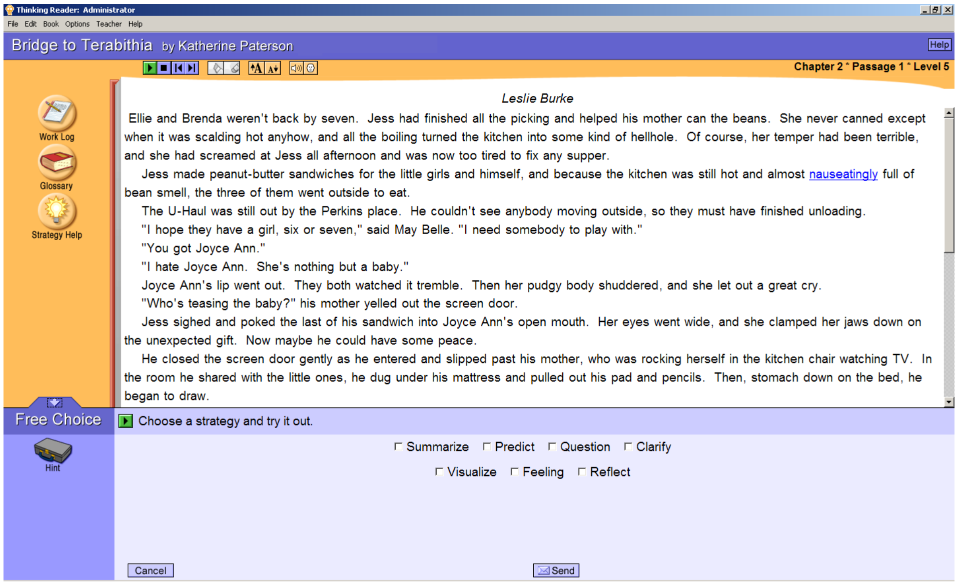


I know someone who is like the character in this story. I thought about how that person would feel. Then, I used descriptive language to paint a vivid picture of their emotions.

Click Justin to hear his thinking.

Source: Tom Snyder Productions (2006a).

Exhibit A.7 Level 5 Screenshot, Free-Choice Strategy Prompt


Thinking Reader: Administrator
File Edit Book Options Teacher Help
Help
Work Log
Glossary
Strategy Help
Free Choice
Hint
Cancel
Send

Bridge to Terabithia by Katherine Paterson

Chapter 2 * Passage 1 * Level 5

Leslie Burke

Ellie and Brenda weren't back by seven. Jess had finished all the picking and helped his mother can the beans. She never canned except when it was scalding hot anyhow, and all the boiling turned the kitchen into some kind of hellhole. Of course, her temper had been terrible, and she had screamed at Jess all afternoon and was now too tired to fix any supper.

Jess made peanut-butter sandwiches for the little girls and himself, and because the kitchen was still hot and almost nauseatingly full of bean smell, the three of them went outside to eat.

The U-Haul was still out by the Perkins place. He couldn't see anybody moving outside, so they must have finished unloading.

"I hope they have a girl, six or seven," said May Belle. "I need somebody to play with."

"You got Joyce Ann."

"I hate Joyce Ann. She's nothing but a baby."

Joyce Ann's lip went out. They both watched it tremble. Then her pudgy body shuddered, and she let out a great cry.

"Who's teasing the baby?" his mother yelled out the screen door.

Jess sighed and poked the last of his sandwich into Joyce Ann's open mouth. Her eyes went wide, and she clamped her jaws down on the unexpected gift. Now maybe he could have some peace.

He closed the screen door gently as he entered and slipped past his mother, who was rocking herself in the kitchen chair watching TV. In the room he shared with the little ones, he dug under his mattress and pulled out his pad and pencils. Then, stomach down on the bed, he began to draw.

Choose a strategy and try it out.

□ Summarize □ Predict □ Question □ Clarify

□ Visualize □ Feeling □ Reflect

One hint is provided

Source: Tom Snyder Productions (2006a).

Appendix B. Data Collection

Appendix B provides information about data collection, including the assent forms (Appendix B1), and the student and teacher measures (Appendix B2)—including the actual measures for the Metacognitive Awareness of Reading Strategies Inventory (MARSI), the Motivation for Reading Questionnaire (MRQ), and the teacher background questionnaire. The information on the Center for the Improvement of Early Reading Achievement (CIERA) Classroom Observation Scheme (Taylor & Pearson, 2000; Taylor, 2004) and an example of CIERA narrative notes and codes are presented in Appendices B3 and B4. The power analysis is detailed in Appendix B5.

B1. Informed Assent Procedure

Student assent to complete Gates-MacGinitie Reading Tests, Metacognitive Awareness of Reading Strategies Inventory, and Motivation for Reading Questionnaire. To ensure that students were aware that taking the Gates-MacGinitie Reading Tests (GMRT), MARSI, and MRQ was voluntary, an assent form was distributed to each student and read aloud by an administrator before pre- and posttests were administered. The form is shown in Box B1.1.

Box B1.1 Student Assent Form

We are interested in learning more about how kids understand the books, articles, and other things they read in school. We would like to ask you to take a reading test and two short questionnaires. In the questionnaires, we ask you about how you feel about reading activities and how you try to understand what you are reading.

Taking the test and surveys is voluntary. A student does not have to answer questions that he/she does not want to answer. Students can decide not to participate at any time without penalty. Results will be used only for research purposes and all results are kept strictly confidential. We will give each of you a study identification number in place of your names. The reports prepared for this study will summarize findings and will not link responses with a specific school or student. We will not provide information that identifies you or your school to anyone outside the study team, except as required by law.

There are no known risks in participating in this study. Your participation in the study will be helpful in understanding how the *Thinking Reader* software program has an impact on students' reading comprehension, motivation, and use of reading strategies.

If you would like more information about answering these questions, call Kathryn Drummond at AIR toll free, at 1-866-236-4285, or e-mail kdrummond@air.org. For questions about your rights as a participant, please contact the IRB Chair at IRBChair@air.org, or toll free at 1-800-634-0797.

Teacher assent to complete background questionnaire. The cover page of the questionnaire informed teachers that answering the background questionnaire was voluntary. The text of the statement is shown in Box B1.2.

Box B1.2 Teacher Questionnaire Assent Statement

Taking this background survey is voluntary. A teacher does not have to answer questions that he/she does not want to answer. Teachers can decide not to participate at any time without penalty. Results will be used only for research purposes and all results are kept strictly confidential. We will give each of you a study identification number in place of your names. The reports prepared for this study will summarize findings and will not link responses with a specific school or student. We will not provide information that identifies you or your school to anyone outside the study team, except as required by law.

There are no known risks in participating in this study. Your participation in the study will be helpful in understanding how the *Thinking Reader* software program has an impact on students' reading comprehension, motivation, and use of reading strategies.

If you would like more information about answering these questions, call Kathryn Drummond at AIR toll free, at 1-866-236-4285, or e-mail kdrummond@air.org. For questions about your rights as a participant, please contact the IRB Chair at IRBChair@air.org, or toll free at 1-800-634-0797.

B2. Student and Teacher Measures

Additional information about the Gates-MacGinitie Reading Tests. Level 6, Form S of the GMRT (used for the pretest) was vertically scaled using the Rasch model and transformed into extended scale scores. Level 6, Form T (used for the posttest) was equated to Form S through equipercentile equating (Riverside Publishing, personal communication, October 9, 2008). Each test form consisted of a vocabulary subtest and a comprehension subtest. The alternate forms reliabilities between Forms S and T are 0.87 for the vocabulary subtest and 0.82 for the comprehension subtest (MacGinitie, MacGinitie, Maria, & Dreyer, 2002). Kuder-Richardson Formula 20 values were used to assess internal consistency. For each subtest (vocabulary and comprehension), for both forms (S and T), and for fall and spring, the values ranged from 0.90 to 0.92 (MacGinitie et al., 2002).

Riverside, the publisher of the GMRT, scored the student answer sheets and provided the derived scores that were used in the impact analyses. Riverside's scoring department marks answers as incorrect if they are left blank, are illegible, or contain multiple responses.

Additional information about the Metacognitive Awareness of Reading Strategies Inventory. The 30-item MARSI is shown in Box B2.1. MARSI scores are computed by averaging the 30 items; scale scores range from 1 ("I never or almost never do this") to 5 ("I always or almost always do this"). Cronbach's alpha for the MARSI was 0.94 at the pretest and 0.93 at the posttest (similar to Cronbach's alphas obtained for Grade 6 students in the original validity study, Mokhtari & Reichard, 2002). A minimum of 20 MARSI items were required to compute a valid score (67% of the 30-item scale). Nineteen students at the pretest (0.79%) and fewer than four students at the posttest (0.13%) completed fewer than 20 MARSI items.

Box B2.1 The Metacognitive Awareness of Reading Strategies Inventory (MARSI)

Directions: Listed below are statements about what people do when they read academic or school-related materials, such as textbooks or library books.

Five numbers follow each statement (1, 2, 3, 4, 5), and each number means the following:

- **1** means "I **never or almost never** do this."
- **2** means "I do this **only occasionally**."
- **3** means "I **sometimes** do this" (50% of the time).
- **4** means "I **usually** do this."
- **5** means "I **always or almost always** do this."

After reading each statement, circle the number (1, 2, 3, 4, or 5) that applies to you using the scale provided. Please note that there are no right or wrong answers to the statements in this inventory.

- I have a purpose in mind when I read.
- I take notes while reading to help me understand what I read.
- I think about what I know to help me understand what I read.
- I preview the text to see what it's about before reading it.
- When text becomes difficult, I read aloud to help me understand what I read.
- I summarize what I read to reflect on important information in the text.
- I think about whether the content of the text fits my reading purpose.
- I read slowly but carefully to be sure I understand what I'm reading.
- I discuss what I read with others to check my understanding.
- I skim the text first by noting characteristics like length and organization.
- I try to get back on track when I lose concentration.
- I underline or circle information in the text to help me remember it.
- I adjust my reading speed according to what I'm reading.
- I decide what to read closely and what to ignore.
- I use reference material such as a dictionary to help me understand what I read.
- When the text becomes difficult, I pay closer attention to what I'm reading.
- I use tables, figures, and pictures in the text to increase my understanding.
- I stop from time to time and think about what I'm reading.
- I use context clues to help me better understand what I'm reading.
- I paraphrase (restate ideas in my own words) to better understand what I read.
- I try to picture or visualize information to help me remember what I read.
- I use typographical aids like boldface and italics to identify key information.
- I critically analyze and evaluate the information presented in the text.
- I go back and forth in the text to find relationships among ideas in it.
- I check my understanding when I come across conflicting information.
- I try to guess what the material is about when I read.
- When the text becomes difficult, I reread to increase my understanding.
- I ask myself questions I like to have answered in the text.
- I check to see whether my guesses about the text are right or wrong.
- I try to guess the meaning of unknown words or phrases.

Source: Mokhtari & Reichard (2002).

Additional information about the Motivation for Reading Questionnaire. A composite score based on 28 items of the MRQ was used as an overall, robust measure of reading motivation; internal consistency was 0.85 (Guthrie, Wigfield, Metsala, & Cox, 1999). Previous research found that the scales loaded onto three factors, generally mapping onto intrinsic aspects of motivation; extrinsic aspects of motivation; and competition, work avoidance, and lack of involvement.

However, the factor analysis (principal components extraction, varimax rotation) revealed that items loaded onto a single factor. Cronbach's alpha for the entire 52-item scale was high: 0.93 at the pretest and 0.94 at the posttest. To address missing data at the item level, a minimum of two-thirds of the items (in this case, 35 of 52 MRQ items) were required to compute a valid score. Twenty students at the pretest (0.83%) and zero students at the posttest completed fewer than 35 MRQ items.

The MRQ items are rated on a four-point scale, from 1 (low motivation) to 4 (high motivation). They are shown in Box B2.2. Five items are negatively worded and were reversed before computing the composite score.

Box B2.2 The Motivation for Reading Questionnaire (MRQ)

We are interested in your reading. The sentences in this questionnaire describe how some students feel about reading. Read each sentence and decide whether it describes a person who is like you or different from you. There are no right or wrong answers. We only want to know how you feel about reading. For many of the statements, you should think about the kinds of things you read in your class.

- If the statement is **very different from you**, circle a 1.
- If the statement is **a little different from you**, circle a 2.
- If the statement is **a little like you**, circle a 3.
- If the statement is **a lot like you**, circle a 4.

Remember, when you give your answers you should think about the things you are reading in your class. **There are no right or wrong answers. We just are interested in YOUR ideas about reading. To give your answer, circle ONE number on each line. The answer numbers are right next to each statement.**

Let's turn the page and start. **Please read each of the statements carefully, and then circle your answer.**

- I visit the library often with my family
- I like hard, challenging books.
- I know that I will do well in reading next year.
- I do as little schoolwork as possible in reading. (REVERSED)
- If the teacher discusses something interesting, I might read more about it.
- I read because I have to.
- I like it when the questions in books make me think.
- I read about my hobbies to learn more about them.
- I am a good reader.
- I read stories about fantasy and make-believe.
- I often read to my brother, sister, friend, or relative.
- I like being the only one who knows an answer in something we read.
- I read to learn new information about topics that interest me.
- My friends sometimes tell me I am a good reader.
- I learn more from reading than most students in the class.

- I like to read about new things.
- I like hearing the teacher say I read well.
- I like being the best at reading.
- I look forward to finding out my reading grade.
- I sometimes read to my mother or father.
- My friends and I like to trade things to read.
- It is important for me to see my name on a list of good readers.
- I don’t like reading something when the words are too difficult. (REVERSED)
- I make pictures in my mind when I read.
- I always do my reading work exactly as the teacher wants it.
- I usually learn difficult things by reading.
- I don’t like vocabulary questions. (REVERSED)
- Complicated stories are no fun to read. (REVERSED)
- I am happy when someone recognizes my reading.
- I feel like I make friends with people in good books.
- My mother or father often tells me what a good job I am doing in reading.
- Finishing every reading assignment is very important to me.
- I like mysteries.
- I talk to my friends about what I am reading.
- I like to get compliments for my reading.
- Grades are a good way to see how well you are doing in reading.
- I like to help my friends with their schoolwork in reading.
- I read to improve my grades.
- My mother or father asks me about my reading grade.
- I enjoy a long, involved story or fiction book.
- I like to tell my family about what I am reading.
- I try to get more answers right than my friends.
- If the project is interesting, I can read difficult material.
- I enjoy reading books about people in different countries.
- I read a lot of adventure stories.
- I always try to finish my reading on time.
- If a book is interesting, I don’t care how hard it is to read.
- I like to finish my reading before other students.
- I am willing to work hard to read better than my friends.
- I don’t like it when there are too many people in the story. (REVERSED)
- It is very important to me to be a good reader.
- In comparison to other activities I do, it is very important to me to be a good reader.

Note: Five items are negatively worded and were reversed before computing the composite score.
Source: Wigfield & Guthrie (1997).

Teacher background questionnaire. Box B2.3 displays the teacher questionnaire, and Table B2.1 shows the frequencies of the teacher background questionnaire.

Box B2.3 Teacher Questionnaire

Education and professional certification

What academic degree(s) do you hold?
Mark all that apply.

○ No degree
○ Associate degree
○ Bachelor's degree
○ Master's degree
○ Doctorate (e.g., Ph.D.)
○ First professional degree (e.g., M.D., L.L.B., J.D., D.D.S.)
○ Education specialist/professional diploma based on at least 1 year of work (e.g., credential, 6-year certificate)

Which of the following describes the teaching certificate you currently hold?
Mark all that apply.

○ Regular or standard certificate or advanced professional certificate
○ Probationary certificate (issued after satisfying all requirements except the completion of a probationary period)
○ Provisional or other type of certificate given to persons who are still participating in what the state calls an "alternative certification program"
○ Temporary certificate (requires some additional college coursework, student teaching, and/or passage of a test before regular certification can be obtained)
○ Waiver or emergency certificate (issued to persons with insufficient teacher preparation who must complete a regular certification program in order to continue teaching)
○ I have received National Board Certification
○ I am currently working toward National Board Certification
○ I do not have any of the above certifications

Are you endorsed or certified in any of the areas below?
Mark all that apply.

○ Elementary Education
○ English
○ Foreign Language
○ Early Childhood Education
○ Language Arts
○ Language Therapy
○ Special Education
○ Reading Specialist
○ Speech Therapy
○ English as a Second Language (ESL) or English for Speakers of Other Languages (ESOL) or English Language Learners (ELL) or Limited English Proficiency (LEP)
○ Other (please specify): ____

Counting this year, how many years have you taught as an elementary or secondary teacher? *Include any full-time teaching assignments, part-time teaching assignments, and long-term substitute assignments. If less than 4 months' total experience, enter "0."*

Number of years: ______________________________

Counting this year, how many years have you taught as sixth-grade teacher? *Include any full-time teaching assignments, part-time teaching assignments, and long-term substitute assignments in sixth grade. If less than 4 months' total experience, enter "0."*

Number of years: ______________________________

Current Classrooms

Please think about all the English language arts sections you are teaching this year when answering the items in this section.

How do your current classes compare to previous classes you have taught?
Mark only one bubble.

○ Fewer students
○ More students
○ About the same number of students

How does the number of ELL students compare to the number of ELL students you have taught in previous classes?
Mark only one bubble.

○ Fewer students ○ More students ○ About the same number of students

How does this number of special education students (with IEPs) compare to the number of special education students in your previous classes?
Mark only one bubble.

○ Fewer students ○ More students ○ About the same number of students

What is the sixth-grade core reading program or anthology?

____

____

Table B2.1 Frequencies on Teacher Background Characteristics for Categories With Responses[31], by Study Condition

	Intervention (*n* = 48)		Control (*n* = 42)			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	**χ^2 (df)[a]**	***p* value**
Academic degree						
Associate	≤3	≤6.3	≤3	≤7.1		
Bachelor’s	48	100	42	100		
Master’s	37	77.1	27	64.3	χ^2 (1) = 1.79	.18
Education specialist/professional diploma/first professional	6	12.5	≤3	≤7.1		
Teaching certificate						
Regular/standard/advanced professional	45	93.8	41	97.6	χ^2 (1) = 0.79	.37
Probationary/provisional	4	8.3	≤3	≤7.1		
National board/working toward national board	≤3	≤6.3	≤3	≤7.1		
Certification or endorsement area						
Elementary education	46	95.8	38	90.5	χ^2 (1) = 1.03	.31
Early childhood education	≤3	≤6.3	≤3	≤7.1		
Special education/English language learner	5	10.4	8	19.0		
English language arts/reading specialist	9	18.8	4	9.5		
Language therapy	≤3	≤6.3	≤3	≤7.1		
Other	6	12.5	9	21.4	χ^2 (1) = 1.29	.26

Note: Results are based on 90 teachers. Percentages do not always sum to 100 because teachers could mark more than one answer per question. The calculation of the statistics does not account for the clustering of teachers by school.

[a] Numbers in parentheses are degrees of freedom for chi-squared tests. Chi-squared statistics are not presented for characteristics when the frequency is below 15 or for bachelor’s degree (100% of teachers held this degree).

Source: Teacher background questionnaire administered by study team.

[31] Additional response categories were offered to teachers, but those not presented in this table had response frequencies of zero. In addition, some categories with low frequencies were combined for presentation in this table.

B3. Adapted Center for the Improvement of Early Reading Achievement (CIERA) Classroom Observation Scheme Codes and Definitions

Table B3.1 Adapted Center for the Improvement of Early Reading Achievement Classroom Observation Scheme Codes and Definitions

Component	Code	Definition
Dimension 1: Who (Who or what in the classroom is providing instruction/working with students?)		
Who		
Classroom teacher	c	Classroom teacher.
Computer	m	Computer/software.
Specialist	s	Reading teacher, Title 1 teacher, reading resource teacher, special education teacher, speech and language teacher, English language learner teacher, bilingual teacher, etc.
Aide	a	Paraprofessional, instructional aide, parent volunteer.
No one	n	No one is in the room, or no one is directly working with the students (e.g., students are working in their seats independently and no one is circulating).
Not applicable	9	No instruction is occurring.
Dimension 2: Groupings (What instructional groupings do you see?)		
Grouping		
Whole class/large group	w	All of the students present in the classroom or lab (except for one or two individuals who might be working with someone else).
Small group	s	Students are working in groups of three or more students. If there are more than 10 students in a group, call this large group.
Pairs	p	Students are working in pairs.
Individual	i	Students are working independently.
Individual w/teacher	it	Student is working with the teacher (e.g., teacher–student conference).
Other	o	Some other grouping practice is in place.
Not applicable	9	No instruction is taking place.
Dimension 3: Major Focus (What major academic area is being covered?)		
Major focus		
Reading	r	Reading, reading comprehension, writing in response to reading (where this is the major purpose for the writing), literature study, reading vocabulary, journal/worklog writing.
Composition/ writing process	w	Writing for the purpose of expressing or communicating ideas (but not writing in which the major purpose is to respond to reading); learning how to write; writer's workshop, creative writing, and report writing.
Other language	l	Aspects of language arts other than the above (e.g. grammar, mechanics, oral expression, spelling, handwriting) not associated with reading text.
Other (not literacy)	o	Focus is academic but not in literacy.
Not academic	9	None of the above seems to apply; focus is not academic.
Dimension 4: Materials		
Material		
Textbook	tn ti	School textbook (e.g., basal reader, anthology, social studies book). Distinguish between narrative (tn) and informational text (ti).
Thinking Reader novel—hard copy version	trb	*Thinking Reader* novels—*Roll of Thunder, Hear My Cry, Bridge to Terabithia, Tuck Everlasting, Bud, Not Buddy, Dragonwings, My Brother Sam Is Dead, Esperanza Rising, A Wrinkle in Time,* and *The Giver.*
Thinking Reader computer—digital text	trc	*Thinking Reader* software program.
Other narrative text	n	Narrative text (e.g., narrative book, historical fiction, novel, poem, other trade book). **Please note: Do not code "n" for *Thinking Reader* novels.**

Other informational text	i	Informational book, reference book (encyclopedia, dictionary, etc.), newspapers, and magazines.
Computer	c	Computers/laptops that are in the classroom or computer lab. **Please note: Do not code** *Thinking Reader* **software program here.**
Video/television/audio	Av	Videos, television, DVDs, audiotapes, CDs, and listening center.
Overhead projector	op	Overhead projector, opaque projector, LCD projector, and smartboard
Student writing	w	Student writing (more than words or disconnected sentences) is being finished or in progress. Writing should be a paragraph or more of connected sentences.
Worksheet	s	Worksheet, workbook page, sheets for one-word or one-sentence answers, blank sheet of paper, or graphic organizer (e.g., chart, map, Venn diagram). **This does not include printed prompts for writing.**
Board/chart/posters	b	Board or chart is being used (e.g., chalkboard, whiteboard, pocket chart, hanging chart, poster, chart paper).
Other	o	Something other than the above is being used—for example, flashcards, Post-it notes, highlighters (either markers or the computer highlighter), or manipulatives.
Not applicable	9	None of the above seems to apply.
Dimension 5: Literacy Activity		
Activity		
Reading connected text	r	Students are engaged in reading text. This includes silent reading, reading text on a computer, choral reading (even if not all students are participating), simultaneous oral reading, oral turn-taking reading, and repeated oral readings.
Listening to connected text	l	Students are engaged in listening to text that the teacher or the computer is reading. If the teacher or the computer is reading to students, code as 1, even if the students are to be following along silently.
Vocabulary	v	Students are engaged in discussing/working on word meaning(s), including using the glossary to look up word meanings in the dictionary or on the computer. This may include discussions of cognates, synonyms, antonyms, homonyms, homophones, homographs, classifying words, etc.
Comprehension: Lower level	ml	Students are engaged in *talking* or *writing* about the meaning of text that is at a lower level of text interpretation. That is, students are asked to identify meaning that is **explicitly stated** in the text. The writing may be a journal entry about the text requiring a lower level of text interpretation or may be a fill-in-the-blank worksheet that is on explicit text meaning.
Comprehension: Higher level	mh	Students are involved in *talking* or *writing* about the meaning of text that is **engaging them in higher level thinking.** This talking or writing about the text requires a higher level of text interpretation or goes beyond the text: generalization, application, evaluation, or aesthetic response. A student must go beyond a yes or no answer (e.g., in the case of an opinion or aesthetic response).
Comprehension: Identification	ci	The teacher/computer and/or students are engaged in ***naming, defining,*** or ***pointing out*** a comprehension activity. Comprehension activities may include identifying the main idea and important details, determining cause and effect, distinguishing fact from opinion or reality from fantasy, identifying the author's purpose or bias, sequencing, classifying, comparing, making predictions or connections, drawing conclusions or inferring, clarifying, summarizing, asking questions, or visualizing. **This differs from ml and mh in that the specific comprehension activity is identified in an explicit manner** (not simply done or practiced, as when a teacher asks students to make predictions in a text, without identifying the comprehension activity).

Comprehension: Metacognition	cm	The teacher and/or students are engaged in reviewing ***how***, ***when***, or ***why*** one might engage in a comprehension activity. Comprehension activities may include identifying the main idea and important details, determining cause and effect, distinguishing fact from opinion or reality from fantasy, identifying the author's purpose or bias, sequencing, classifying, comparing, making predictions or connections, drawing conclusions or inferring, clarifying, summarizing, asking questions, or visualizing. **This differs from ci in that it considers *how* one engages in the activity** (e.g., how to identify the important details), ***why* one might choose to engage in the activity** (e.g., "We might visualize, or make pictures in our mind of what's happening, to make sure we have a clear idea of what's happening in the story"), or ***when* one would find this activity most useful** (e.g., "Distinguishing fact from opinion is important as you read about history, because everyone who writes about historical events has a particular perspective. So, it will be important to notice when people are writing their opinions and when they are writing facts about the mission").
Text elements	te	Attending to various elements of ***written*** text, including: mechanical/visual features, (punctuation, font, and mechanics); text level or genre structures, such as sequence, topic sentences, macro-level text grammar (plot, temporal sequence, cause-and-effect, problems and solutions, genre); and literary devices (i.e., devices that authors use to convey meaning, nuance, and attitudes toward characters or other aspects of a text). Examples include foreshadowing, metaphor, symbolism, literary or historical allusions, point of view, tone, mood, and theme.
Language development	ld	If the instruction is ***oral,*** code here; ***if in response to an element from a text, code as text elements.*** Teacher or computer is helping students attend to figurative language, such as metaphors, clichés, idioms; word-level linguistic structures, such as contractions and morphology; sentence-level linguistic structures, such as verb tense, subject-verb agreement, and parts of speech; pronunciation and articulation; or translating for the purposes of developing students' language skills.
Writing	w	Students are engaged in a writing activity. This includes students taking notes in response to a teacher's lecture or while reading a text. Students may be copying notes.
Word work	ww	Word identification: Students are focusing on identifying words. For example, the teacher is supplying a word when students get stuck during reading, or the teacher is reviewing words prior to reading. Word recognition strategy: Students are focusing on using of one or more strategies to figure out words while reading, typically prompted by the teacher or the computer.
Spelling	sp	Students are focusing on how to spell word(s).
Quiz/assessment	q	Students are taking a quiz or assessment.
Worklog related	wl	Teacher and/or students are engaged in an activity related to the *Thinking Reader* worklog. Such activities include a teacher commenting on student worklogs, or students' reviewing or reading their worklogs.
Other	o	Literacy focus other than one of the above.
Not applicable	9	None of the above seems to apply (i.e., no literacy activities are occurring in the classroom).
Dimension 6: Teacher's Interaction Style		
Interaction style		
Telling/giving information	t	Telling or giving students information or explaining how to do something. This may include paraphrasing text or translating to convey information.

Recitation	r	The teacher or the computer is engaging the students in answering questions or responding (question–answer–question–answer). The purpose primarily appears to be getting the students to answer the questions asked rather than engaging them in a discussion. When the dynamic is recitation, but the teacher is requesting elaborated responses from students, Level 6 would be coded as r and d. Also, recitation is coded if the teacher asks students to report their responses to a particular prompt to one or two neighbors. Teacher is clearly in control and managing turns.
Discussion	d	Students are engaged in a discussion or conversation that is largely facilitated by the teacher. Students may respond to each other, but with the teacher's mediation. Exchange may be teacher–student–student–student, rather than teacher–student–teacher–student. A midpoint between conversation and recitation, the teacher is in control and but not always managing turns.
Modeling or coaching/scaffolding	mcs	*Modeling:* The teacher or computer is explicitly **showing, demonstrating, or thinking aloud** the steps of how to do something or how to do a process as opposed to simply explaining it (e.g., a teacher or computer models fluent reading after modeling word-by-word reading, and talks about the difference). When modeling is coded at Dimension 6, listening should be coded at Dimension 7. *Coaching/scaffolding:* The teacher or computer is coded as **prompting/providing support,** which will transfer to other situations as students are attempting to perform a strategy or activity or to answer a question. The teacher or computer's apparent purpose is to foster independence to get a more complete action or to help students elaborate on an answer (rather than to simply get a student to answer a question).
Reading aloud	ra	The teacher is reading aloud to the students.
Assessment	a	The teacher is engaging in questioning, explaining, and providing directions for the purpose of assessing student performance. Typically this would involve recordkeeping or assigning students an earned grade. This also includes the teacher providing feedback on students' worklogs.
Other	o	Interaction style other than what is listed above. Listening or watching without giving feedback would be coded as o.
Not applicable	9	If students are not working with the teacher or computer (i.e., no direct teacher–student interaction).
Dimension 7: Expected Student Response		
Student response		
Reading	r	Students are to be reading, either silently or orally, from a book or text on the computer. Includes reading individually, in pairs, choral reading as a group, or simultaneous oral reading.
Talking	t	Talking is coded whenever students are expected to respond orally (but are not reading) whether individually, one after another, to each other, or as a group (e.g., choral response).
Listening	l	Typically listening is coded when the teacher is operating in some sort of presentational mode—telling students information, modeling, or reading aloud to the students (at Dimension 6). Do not code if students are reading, orally responding.
Writing	w	Students are to be writing or typing words (including a spelling task), sentences, or paragraphs.
Manipulating	m	Students are to be manipulating, using their hands (other than writing).
Other	o	Some form of responding not listed above code.
Not applicable	9	If students are not required to respond at all.

a. The original CIERA classroom observation scheme (Taylor & Pearson, 2000; Taylor, 2004) refers to the coding areas as "levels." This report calls them "dimensions" to avoid confusion with the discussion of levels of support in the *Thinking Reader* program.

Source: Adapted from Taylor & Pearson, 2000; Taylor, 2004

B4. Example of Narrative Notes and Codes for an Observation Segment

Box B4.1 Example of Narrative Notes and Codes for an Observation Segment From an Intervention Group Classroom

Segment Start Time: 9:39 a.m.

Total Number of Students in Class: 15	Total Number of Students On Task: 15

Eight students are reading with the headphones on, using the blue highlighter read-aloud tool.

Four students are responding to strategy prompts: Two students are completing a prompt using the Summarize strategy, one student is completing a prompt using the Prediction strategy, and one student is completing a prompt using the Feeling strategy.

One student is using the peer coach, Justin, to help in completing the provided prompt.

One student is using the dictionary.

Two students are reading the teacher's comments in their worklogs.

The teacher is reading and responding to the students' worklogs.

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6	Dimension 7
Who	Grouping	Major Focus	Materials	Activity	Interaction Style	Student Response
m, c	i	r	trc	l	ra	l
				mh	t, mcs	l, w, r
				v	t, mcs	l, r
				wl	a	o, r

Source: Classroom observations conducted by study team.

B5. Power Analyses

This section presents power analyses for the *Thinking Reader* study design based on baseline (pretest) counts. The precision of impact estimates is assessed by computing the minimum detectable effect size (MDES) based on a multisite cluster randomized design in which teachers are randomly assigned within schools. Teachers were randomly assigned to intervention and control conditions, so clustering was assumed to take place at the teacher level, although a teacher might have taught more than one classroom.[32]

The following assumptions were made in the power calculations:

- Power: 80%.
- *Statistical significance: p* < 0.05, two-tailed test, lowered to 0.0125 due to multiple comparisons (simple Bonferroni corrections to adjust for four outcomes: reading vocabulary,

[32] Teachers who teach multiple classrooms were given more weight in the analysis because they represented larger clusters. On average, cluster size was larger in middle schools than in elementary schools.

reading comprehension, use of reading comprehension strategies, and motivation for reading).[33]

- *Intraclass correlation* (ICC): 0.15 between teachers and 0.15 between schools.[34]
- *Covariate R-square:* Three different values for simplicity assumed to be the same for student, class, and school levels.
  - 0.50
  - 0.60
  - 0.70
- *School sample size:* 32 schools.
- *Teacher and student sample size:* Using baseline counts, each school averaged 2.8 teachers. The power calculations assumed an average of three teachers per school. Based on pretest information, each classroom averaged 18.73 students. The power calculations assumed an average of 15 students per classroom at posttest, allowing for approximately 20% attrition.
- *Balance:* Using baseline counts, the design is slightly imbalanced, with approximately 53% of the teachers and 54% of the students belonging to the intervention condition. This small imbalance was incorporated in the power calculations as 55% to 45% imbalance.

Here, the statistical formula in which schools are treated as random effects is used (see equations below that define elements of the equations and make additional explicit assumptions):

Var (Pooled Impact) =

$$\frac{2\sigma^2\rho_1(1-c_3)(1-R^2)}{s}+\frac{2\sigma^2\rho_2(1-R^2)}{(.45s)k}+\frac{2\sigma^2(1-\rho_1-\rho_2)(1-R^2)}{(.45s)kn}$$

MDES = M (α, β, *df*) * $\sqrt{\frac{var(impact)}{\sigma}}$

where

s is the total number of schools in the study sample.

k is the number of classrooms in each school.

n is the average number of students per classroom (including 20% attrition).

σ^2 is the variance of the outcome measure (assumed to be 1).

ρ_1 is the unconditional intraclass correlation between schools (without covariates).

[33] When the power analyses were calculated, the intention was to correct for the four outcomes; however, in the final analyses, corrections were made for only the two primary outcomes: reading vocabulary and reading comprehension.

[34] Minimal guidance is available for choosing an ICC during power analysis for a cluster randomized trial. Schochet (2008) notes that depending on homogeneity of schools, ICCs typically range from 0.10 to 0.20 when dealing with standardized test scores. Following this lead, a 0.15 ICC was used.

ρ_2 is the unconditional intraclass correlation between teachers (without covariates).

R^2 is the proportion of the variance components at the school, classroom, and student levels, as explained by the student-level pretests.

C_3 is the correlation between the intervention and control group classroom means within a school. Using $C_{3\,=}$ 0.97 implies, with a rho of 0.15, an effect size variance of 0.01 (or a standard deviation of 0.10).

M (α, β, *df*) is the multiplier that translates the standard error into a minimum detectable effect estimate. It is equal to the *t* critical value for α (the significance level of the intended statistical test) plus the *t* critical value for β (the likelihood of detecting significant effects given a true effect of a particular size)—that is, the power of the test (multiplier = 3.51, *df* = 30).

With these assumptions, the MDES estimates are 0.24, 0.22, and 0.19, for *R*-squared values of 0.50, 0.60, and 0.70, respectively.

Appendix C. Missing Data, Baseline Equivalence of the Analytic Sample, and the Impact Model

This appendix contains information about missing data (Appendix C1) and baseline equivalence of the analytic sample (Appendix C2) and presents the impact model (Appendix C3).

C1. Examining Patterns of Missing Data

Table C1.1 presents the number and percentage of missing data in the Gates-MacGinitie Reading Tests (GMRT) outcomes, by study condition. It shows that the percentage of eligible students missing pretest reading vocabulary scores is 0.78% in the intervention group and 0.80% in the control group, and that the percentage of missing pretest reading comprehension scores is 0.47% in the intervention group and 1.16% in the control group. The percentage of students missing posttest reading vocabulary scores is 9.80% in the intervention group and 11.15% in the control group, and the percentage missing reading comprehension scores is 10.19% in the intervention group and 11.33% in the control group. Differences between the intervention and control conditions in percentages of students missing pretest or posttest scores are not statistically significant.

Table C1.1 Students Missing Reading Achievement Data, by Study Condition

Primary outcomes: Gates-MacGinitie Reading Tests extended scale scores	Intervention (*n* = 1,286)		Control (*n* = 1,121)			
	Number	Percent	Number	Percent	χ^2 (*df* = 1)	*p* value
Pretest measures						
Reading vocabulary	10	0. 8	9	0.8	0.00	.94
Reading comprehension	6	0.5	13	1.2	3.67	.06
Posttest measures						
Reading vocabulary	126	9.8	125	11.2	1.17	.28
Reading comprehension	131	10.2	127	11. 3	0.82	.37

Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table C1.2 presents the number and percentage of missing data for the ancillary outcomes, by study condition. At the pretest, the percentage of students missing Metacognitive Awareness of Reading Strategies Inventory (MARSI) scores is 0.62% in the intervention group and 0.98% in the control group, and the percentage missing Motivation for Reading Questionnaire (MRQ) scores is 0.86% in the intervention group and 0.80% in the control group. At the posttest, the percentages of students missing MARSI scores is 7.62% in the intervention group and 8.21% in the control group, and the percentage missing MRQ scores are 7.31% in the intervention group and 7.85% in the control group. None of the differences between the intervention and control conditions are statistically significant.

Table C1.2 Students Missing Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Scores, by Study Condition

	Intervention (*n* = 1,286)		Control (*n* = 1,121)			
Ancillary outcomes	**Number**	**Percent**	**Number**	**Percent**	**χ2 (*df* = 1)**	***p* value**
Pretest						
Reading strategies: MARSI	8	0.6	11	0.9	0.99	.32
Reading motivation: MRQ	11	0.9	9	0.8	0.02	.89
Posttest						
Reading strategies: MARSI	98	7.6	92	8.2	0.28	.60
Reading motivation: MRQ	94	7.3	88	7.9	0.25	.62

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
Source: MARSI and MRQ surveys administered by study team.

Because missing levels are below 1.17% for each pretest, the discussion here focuses on levels of missing data at posttest. The primary outcomes of reading vocabulary and reading comprehension have greater levels of missing posttest data (10.43% and 10.72%, respectively) than the ancillary outcomes of MARSI and MRQ (7.89% and 7.56%, respectively), which may be due in part to the availability of testing accommodations. MARSI and MRQ were read aloud to all students. However, accommodations for the GMRT reading vocabulary and reading comprehension subtests were provided on an individual basis and limited by staff availability. Sometimes accommodations could not be provided at posttesting even when they had been provided at pretesting.

Student attrition, or students transferring out of the study schools during the year, accounts for 65.34% of missing reading vocabulary posttests, 63.57% of missing reading comprehension posttests, 86.32% of missing MARSI posttests, and 90.11% of missing MRQ posttests.

Student-, teacher-, and school-level characteristics were used to examine patterns of missing posttests for the student population as a whole and separately for both conditions. Of the 269 students missing at least one of the four posttests, 93.31% were missing reading vocabulary, 95.91% were missing reading comprehension, 70.63% were missing MARSI, and 67.66% were missing MRQ. Accordingly, “missing” is defined as missing any of the posttest measures.

Results at the student level show systematic differences between students missing outcome measures and those with complete data. For example, Table C1.3 shows that 31.23% of all students missing a posttest had an individualized education program (IEP), compared with only 7.67% of students without a missing posttest. This difference is statistically significant and means that students with an IEP were more likely to have a missing posttest than non-IEP students. This relationship holds true whether examining all students combined or the intervention and control conditions separately (see Tables C1.1 and C1.2).

As shown in Tables C1.3–C1.5, other statistically significant student-level characteristics related to missing posttest data are age (students missing data were older) and GMRT pretest scores (students missing data had lower scores). The relationships between missing posttest data and reading motivation (students missing data had lower MRQ pretest scores) and gender (students missing data were more likely to be male) are statistically significant overall, but not for both intervention and control conditions separately (see Tables C1.3–C1.5).

In the overall sample, no statistically significant teacher-level characteristics were related to missing posttest data (see Table C1.3). However within the intervention group, students missing data were statistically significantly less likely to have a teacher with a master's degree or higher (see Table C1.4). At the school level, statistically significant variables related to higher levels of missing posttest data included having 75% or more of students eligible for free or reduced-price lunch, having 50% or more Black students, and having less than 50% White students (see Table C1.3).

Table C1.3 Patterns of Missing Posttests, by Student, Teacher, and School Characteristics: Intervention and Control Groups Combined

	Not missing any posttest (*n* = 2,138)		**Missing at least one posttest (*n* = 269)**			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	**Test statistic[a]**	***p* value**
Student level						
Individualized education program	164	7. 7	84	31.2	χ^2 (1) = 143.46	.00
English language learner	209	9. 8	28	10.4	χ^2 (1) = 0.11	.74
Female	1,099	51.4	112	41.6	χ^2 (1) = 9.12	.00
Race/ethnicity[b]					χ^2 (4) = 4.19	.38
Black	268	12. 7	22	14.3		
Asian	154	7.3	7	4.6		
Hispanic	583	27.6	42	27.3		
White	783	37.0	65	42.2		
Other race/ethnicity	327	15.5	18	11.7		
Mean age (standard deviation)	2,134	11.54 (0.51)	258	11.73 (0.60)	*t* = –4.78 (.04)	.00
Pretests						
Mean reading vocabulary score (standard deviation)	2,131	503.72 (33.81)	257	489.22 (30.33)	*t* = 6.57 (2.21)	.00
Mean reading comprehension score (standard deviation)	2,129	502.80 (32.24)	259	486.72 (30.21)	*t* = 7.63 (2.11)	.00
MARSI score (standard deviation)	2,124	3.17 (0.68)	264	3.16 (0.77)	*t* = 0.09 (.04)	.93
MRQ score (standard deviation)	2,124	2.85 (0.47)	263	2.77 (0.52)	*t* = 2.42 (.03)	.02
Teacher level						
Master's degree or higher	1,579	73.85	189	70.26	χ^2 (1) = 1.58	.21
Mean years teaching (standard deviation)	2,138	13.12 (8.74)	269	13.13 (8.42)	*t* = –0.01 (.56)	.99
Mean years teaching Grade 6 (standard deviation)	2,138	7.44 (5.88)	269	7.11 (5.40)	*t* = 0.88 (.38)	.38
School level						
Middle school	1,391	65.06	165	61.34	χ^2 (1) = 1.45	.23
Mean enrollment (standard deviation)	2,138	617 (227.55)	269	632.96 (227.80)	*t* = –1.08 (14.72)	.28
High poverty[c]	738	34.52	120	44.61	χ^2 (1) = 10.61	.00
High English language learner[d]	770	36.01	112	41.64	χ^2 (1) = 3.25	.07

	Not missing any posttest (*n* = 2,138)		**Missing at least one posttest (*n* = 269)**			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	**Test statistic[a]**	***p* value**
Race/ethnicity						
More than 50% Black, non-Hispanic[e]	110	5.14	29	10.78	χ^2 (1) = 13.95	.00
More than 50% Hispanic[e]	425	19.88	67	24.91	χ^2 (1) = 3.72	.05
More than 50% White, non-Hispanic[e]	1,052	49.20	110	40.89	χ^2 (1) = 6.61	.01

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.

[a] Numbers in parentheses are standard errors (for *t*-statistics) or degrees of freedom (chi-squared). We used Levene's test (1960) to assess the equality of variance assumption across the two groups.

[b] 2,115 students were not missing race/ethnicity data and not missing any posttest; 154 were missing at least one.

[c] Dummy indicator equals 1 if the percentage of students eligible for free or reduced-price lunch is 75% or higher, and 0 otherwise.

[d] Dummy indicator equals 1 if the percentage of English language learner students (those classified as English as a second language or limited English proficient) is 10% or higher, and 0 otherwise.

[e] Dummy indicator equals 1 if the percentage of students of the given race/ethnicity is 50% or more, and 0 otherwise.

Source: Student rosters completed by study teachers; student self-report section on Gates-MacGinitie Reading Tests (GMRT) administered by study team; GMRT vocabulary and comprehension subtests administered by study team; MARSI and MRQ surveys administered by study team; teacher background questionnaire administered by study team; Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b).

Table C1.4 Patterns of Missing Posttests, by Student and Teacher Characteristics: Intervention Group Only

	Not missing any posttest (*n* = 1,149)		**Missing at least one posttest (*n* = 137)**			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	**Test statistic[a]**	***p* value**
Student level						
Individualized education program	120	10.4	44	32.1	χ^2 (1) = 51.67	.00
English language learner	80	6.9	9	6.6	χ^2 (1) = 0.03	.86
Female	587	51.1	62	45.3	χ^2 (1) = 1.67	.20
Race/ethnicity[b]					χ^2 (4) = 6.58	.16
Black	138	12.1	12	15.4		
Asian	70	6.2	4	5.1		
Hispanic	324	28.5	22	28.2		
White	420	36.9	35	44.9		
Other race/ethnicity	186	16.3	5	6.4		
Mean age (standard deviation)	1,147	11.55 (0.52)	132	11.70 (.60)	*t* = –2.80 (.05)	.01
Pretests						
Mean reading vocabulary score (standard deviation)	1,147	503.73 (32.52)	129	487.93 (30.18)	*t* = 5.27 (3.00)	.00
Mean reading comprehension score (standard deviation)	1,148	501.82 (32.15)	132	485.99 (29.14)	*t* = 5.41 (2.93)	.00
MARSI score (standard deviation)	1,143	3.18 (0.65)	135	3.20 (0.77)	*t* = –0.30 (.07)	.77
MRQ score (standard deviation)	1,142	2.84 (0.46)	133	2.76 (0.51)	*t* = 1.75 (.05)	.08

Characteristic	Not missing any posttest (*n* = 1,149)		Missing at least one posttest (*n* = 137)		Test statistic[a]	*p* value
	Number	Percent	Number	Percent		
Teacher level						
Master’s degree or higher	905	78.76	95	69.34	χ^2 (1) = 6.28	.01
Mean years teaching (standard deviation)	1,149	12.95 (8.66)	137	12.74 (8.24)	*t* = 0.27 (.78)	.79
Mean years teaching grade 6 (standard deviation)	1,149	7.81 (6.11)	137	7.26 (5.76)	*t* = 0.99 (.55)	.32

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
[a] Numbers in parentheses are standard errors (for *t*-statistics) or degrees of freedom (for chi-squared). We used Levene’s test (1960) to assess the equality of variance assumption across the two groups.
[b] 1,138 students were not missing race/ethnicity data and not missing any posttest; 78 were missing at least one posttest.
Source. Student rosters completed by study teachers; student self-report section on Gates-MacGinitie Reading Tests (GMRT) administered by study team; GMRT vocabulary and comprehension subtests administered by study team; MARSI and MRQ surveys administered by study team; teacher background questionnaire administered by study team.

Table C1.5 Posttest Missing Patterns, by Student and Teacher Characteristics: Control Group Only

Characteristic	Not missing any posttest (*n* = 989)		Missing at least one posttest (*n* = 132)		Test statistic[a]	*p* value
	Number	Percent	Number	Percent		
Student level						
Individualized education program	44	4.4	40	30.3	χ^2 (1) = 112.30	.00
English language learner	129	13.0	19	14.4	χ^2 (1) = 0.19	.67
Female	512	51.8	50	37.9	χ^2 (1) = 8.99	.00
Race/ethnicity[b]					χ^2 (4) = 2.29	.68
Black	130	13.3	10	13.2		
Asian	84	8.6	3	3.9		
Hispanic	259	26.5	20	26.3		
White	363	37.2	30	39.5		
Other race/ethnicity	141	14.4	13	17.1		
Mean age (standard deviation)	987	11.53 (0.50)	126	11.75 (0.60)	*t* = –4.00 (.06)	.00
Pretests						
Mean reading vocabulary score (standard deviation)	984	503.70 (35.27)	128	490.52 (30.54)	*t* = 4.04 (3.27)	.00
Mean reading comprehension score (standard deviation)	981	503.95 (32.32)	127	487.48 (31.39)	*t* = 5.42 (3.04)	.00
MARSI score (standard deviation)	981	3.16 (0.70)	129	3.13 (0.77)	*t* = 0.42 (.07)	.67
MRQ score (standard deviation)	982	2.86 (0.47)	130	2.77 (0.53)	*t* = 1.70 (.05)	.09

Characteristic	Not missing any posttest (*n* = 989) Number	Not missing any posttest (*n* = 989) Percent	Missing at least one posttest (*n* = 132) Number	Missing at least one posttest (*n* = 132) Percent	Test statistic[a]	*p* value
Teacher level						
Master’s degree or higher	674	68.10	94	71.21	χ^2 (1) = 0.51	.48
Mean years teaching (standard deviation)	989	13.33 (8.82)	132	13.54 (8.60)	*t* = –0.25 (.82)	.80
Mean years teaching Grade 6 (standard deviation)	989	7.01 (5.57)	132	6.94 (5.01)	*t* = 0.13 (.51)	.90

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
a. Numbers in parentheses are standard errors (for *t*-statistics) or degrees of freedom (for chi-squared). We used Levene’s test (1960) to assess the equality of variance assumption across the two groups.
b. 997 students were not missing race/ethnicity data and not missing any posttest; 76 were missing at least one posttest is 76.
Source: Student rosters completed by study teachers; student self-report section on Gates-MacGinitie Reading Tests (GMRT) administered by study team; GMRT vocabulary and comprehension subtests administered by study team; MARSI and MRQ surveys administered by study team; teacher background questionnaire administered by study team.

C2. Baseline Equivalence of the Analytic Sample

Tables C2.1–C2.3 present the baseline characteristics of the analytic sample (students without any missing posttests, *N* = 2,138). Similar to the patterns found in examining the entire study sample, the differences in percentages of English language learner students and students with an individualized education program (IEP) were the only statistically significant differences found between the intervention and control conditions. Again, the intervention group had a larger percentage of students with IEPs, and the control group had a larger percentage of English language learner students (see Table C2.3). The test statistics included in the tables do not account for the clustering of students nested within teachers or schools.

Table C2.1 Reading Achievement Pretest Scores of Students Without Missing Posttest Data, by Study Condition

Primary outcomes: Gates-MacGinitie Reading Tests extended scale scores	Intervention (*n* = 1,149) Number	Intervention (*n* = 1,149) Mean	Intervention (*n* = 1,149) Standard deviation	Control (*n* = 989) Number	Control (*n* = 989) Mean	Control (*n* = 989) Standard deviation	*t*-statistic[a]	*p* value
Reading vocabulary	1,147	503.74	32.52	984	503.70	35.27	–0.02 (1.47)	.98
Reading comprehension	1,148	501.82	32.15	981	503.95	32.32	1.52 (1.40)	.13

[a] Numbers in parentheses are standard errors of the differences between the two means for each *t*-statistic. We used Levene’s test (1960) to assess the equality of variance assumption across the two groups.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table C2.2 Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Pretest Scores of Students Without Missing Posttest Data, by Study Condition

Ancillary outcomes	Intervention (*n* = 1,149) Number	Intervention (*n* = 1,149) Mean	Intervention (*n* = 1,149) Standard deviation	Control (*n* = 989) Number	Control (*n* = 989) Mean	Control (*n* = 989) Standard deviation	t-statistic[a]	*p* value
Reading strategies: MARSI	1,143	3.18	0.65	981	3.16	0.70	–0.80 (.03)	.42
Reading motivation: MRQ	1,142	2.84	0.46	982	2.86	0.47	0.76 (.02)	.45

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
[a] Numbers in parentheses are standard errors of the differences between the two means for each *t*-statistic. We used Levene’s test (1960) to assess the equality of variance assumption across the two groups.
Source: MARSI and MRQ surveys administered by study team.

Table C2.3 Characteristics of Students Without Missing Posttest Data, by Study Condition

	Intervention (n = 1,149)		Control (n = 989)			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	χ^2 (*df*)[a]	***p* value**
Female	587	51.1	512	51.8	χ^2 (1) = 0.10	.75
Race/ethnicity[b]					χ^2 (4) = 6.88	.14
Black	138	12.1	130	13.3		
Asian	70	6.2	84	8.6		
Hispanic	324	28.5	259	26.5		
White	420	36.9	363	37.2		
Other race/ethnicity	186	16.3	141	14.4		
Individualized education program	120	10.4	44	4. 5	χ^2 (1) = 26.97	.00
English language learner	80	6.9	129	13.0	χ^2 (1) = 22.28	.00

Note: No data are missing for female, individualized education program, or English language learner variables.

[a] Numbers in parentheses are degrees of freedom for chi-squared tests.

[b] Race/ethnicity was missing in 23 cases (11 in the intervention group and 12 in the control group).

Source: Student self-report section on Gates-MacGinitie Reading Tests administered by study team; student rosters completed by study teachers.

C3. Three-Level Impact Model

The multilevel model accounts for student and teacher sources of variability in the outcomes. Because the sample consists of a series of schools in which both intervention and control conditions are implemented, each block or school can be viewed as a mini-experiment. To account for the variability attributable to the randomization within different schools, a random-effects approach was used to model the schools as Level 3 clusters in the multilevel model. Thus, any heterogeneity in the treatment impact across schools is modeled as a random effect.

This three-level specification explores whether the treatment effect varies across the 32 schools (and if so, by how much) or remains relatively stable and homogeneous across schools. Heterogeneity of treatment effect across schools would not be surprising considering that schools can vary in terms of background characteristics of students and teachers, quality of implementation, and many other factors that can reduce or magnify the effects of the program (Seltzer, 2004).

Model specification. This section shows the three models used in the analysis.

Level 1 (between-student, within-teacher and school model)

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\,(\text{Pretest}_{ijk} - \text{Pretest}...) + \pi_{2jk}\,(\text{IEP}_{ijk} - \text{IEP}...) + \pi_{3jk}\,(\text{ELL}_{ijk} - \text{ELL}...) + \varepsilon_{ijk}$,

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Level 2 (between-teacher, within-school model)

$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,(\text{Trt}_{jk} - \text{Trt.}_{k}) + \beta_{02k}\,(\text{Master}_{jk} - \text{Master.}_{k}) + \beta_{03k}\,(\text{Exp\_low}_{jk} - \text{Exp\_low.}_{k})$

$+ \beta_{04k}\,(\text{Exp\_Med}_{jk} - \text{Exp\_Med.}_{k}) + U_{0jk}$, where $U_{0jk} \sim N(0, \tau_{\pi 00})$

$\pi_{1jk} = \beta_{10k}$

$\pi_{2jk} = \beta_{20k}$

$\pi_{3jk} = \beta_{30k}$

Level 3 (between-school model)

$\beta_{00k} = \gamma_{000} + \gamma_{001}$ (Hlunch$_k$ – Hlunch.) + γ_{002} (Midsize$_k$ – Midsize.)

+ γ_{003} (Lsize$_k$ – Lsize.) + V_{00k}

$\beta_{01k} = \gamma_{010} + V_{01k}$

$\beta_{02k} = \gamma_{020}$

$\beta_{03k} = \gamma_{030}$

$\beta_{04k} = \gamma_{040}$

$\beta_{10k} = \gamma_{100}$

$\beta_{20k} = \gamma_{200}$

$\beta_{30k} = \gamma_{300}$

where $\begin{pmatrix} V_{00k} \\ V_{01k} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta 00} & \tau_{\beta 01} \\ \tau_{\beta 10} & \tau_{\beta 11} \end{pmatrix} \right]$

Y_{ijk} represents the posttest outcome score (reading vocabulary, reading comprehension, Metacognitive Awareness of Reading Strategies Inventory [MARSI], or Motivation for Reading Questionnaire [MRQ]) for student *i* in teacher *j*'s class in school *k*.

In the Level 1 model, the student outcome score is modeled as a function of the following student characteristics: pretest score, English language learner (ELL) status, and special education status. These covariates improve precision of the impact estimate and adjust for the English language learner and special education differences between conditions shown in Table 2.7 (see Bloom, Richburg-Hayes, & Black, 2005).

The Individualized Education Program (IEP) variable[35] takes a value of 1 for students with special education status and 0 otherwise. Likewise, the ELL covariate takes a value of 1 for students who are ELLs and 0 otherwise. Thus, π_{2jk} and π_{3jk} capture the difference between special education students and non-special education students and ELL and non-ELL students in teacher *j*'s class in school *k*, respectively.

Student characteristics are centered on the grand mean so that π_{0jk} represents the adjusted mean score for teacher *j* in school *k*. Analogous to the analysis of covariance model, this centering method adjusts for preexisting differences in student characteristics.

The term ε_{ijk} is a random error term that is assumed to be normally distributed with mean 0 and constant variance σ^2.

In the Level 2 model, the adjusted mean outcome score π_{0jk} for teacher *j* in school *k* is modeled as varying randomly across teachers. The coefficient of the treatment indicator β_{01k} (intervention group teachers take a value of 1, and control group teachers take a value of 0) is the key

[35] In the equations in the report, IEP refers to students receiving special education and related services.

parameter of interest and captures the expected difference in achievement between intervention and control conditions in school *k*, holding constant the other covariates of the model. This is the intent-to-treat estimate.

To increase the precision of Level 2 estimates and control for imbalances in teacher characteristics (see Table 2.8), the Level 2 model adjusts for the teacher's education level and years of teaching experience. The *Master* dummy variable takes a value of 1 for teachers with a master's degree or higher (doctorate, education specialist, or professional degree) and 0 otherwise.[36] Similarly, the covariates *Exp_low* and *Exp_Med* capture whether the teacher has fewer than 4 years or between 4 and 20 years of teaching experience, respectively, when compared with teachers with more than 20 years of experience.[37]

The Level 2 covariates are centered within each school. Therefore, the treatment estimate captures the pooled within-school relationship between the intervention and achievement outcomes. By doing group mean centering, intervention and control teachers within the same school can be compared with each other, thus mirroring the way randomization was implemented.

U_{0jk} is a random effect associated with teacher *j*'s classroom in school *k* and is assumed to be normally distributed with mean 0 and variance $\tau_{\pi 00}$.

In the Level 3 model, the average outcome β_{00k} and the intervention effect in each school β_{01k} are modeled as random effects.

In the equation for the average outcome in each school, the parameter γ_{000} captures the average achievement mean for the population of schools, and the covariate *Hlunch* is a dummy variable equal to 1 if the percentage of students eligible for free or reduced-price lunch is higher than 74% and 0 otherwise.[38] Furthermore, the covariates *Msize* and *Lsize* are dummy variables that take the value of 1 if student enrollment at the school falls between 440 and 575 or if enrollment is more than 575 students, respectively, when compared with schools with enrollment sizes fewer than 440 students.[39] The Level 3 covariates are grand mean centered.

V_{00k} is a random effect assumed to be normally distributed with variance $\tau_{\beta 00}$. In the equation for the intervention effect in each school, the parameter represents the average treatment effect for the population of schools after holding constant the covariates of the model. Similarly, V_{01k} is a random effect assumed to be normally distributed with mean 0 and variance $\tau_{\beta 11}$, whereby the latter parameter captures the extent to which school intervention effects vary around γ_{010}. More

[36] Everyone coded 0 for the master's degree covariate has a bachelor's degree as the highest degree.

[37] Teaching experience was added in the impact model as two dummy variables. The experience variables were based on the breakdowns shown in the *Digest of Education Statistics* (Snyder, Dillow, & Hoffman, 2009) and on the empirical distribution of this variable in the sample of teachers. The final cut points and percentages of teachers in each category were fewer than 4 years (12.4%), 4–20 years (65%), and more than 20 years (22.6%).

[38] To distinguish between lower and higher poverty levels among the schools in our sample, the definition in *The 2009 Condition of Education* (Planty Hussar, Snyder, Kena, Kewalramani, et al., 2009) was used, with the highest poverty schools defined as public schools that had more than 74% of their students eligible for free or reduced-price lunch. Based on that definition, 16 of 32 schools in the sample were classified as among the highest poverty schools.

[39] These cut points were generated on the basis of the empirical distribution of variable in the sample. In other words, one-third of schools lie within each category.

precisely, the significance test addresses a key substantive issue in multisite evaluations—that is, whether the magnitude of the intervention varies across schools. The Level 3 random effects also covary, with covariance $\tau_{\beta 01}$.

Computing effect sizes. Effect sizes were computed using Hedges's *g* formula:

$$HG = \frac{\hat{\beta}_{01}}{\sqrt{\frac{(n_t - 1)\hat{S}_t^2 + (n_c - 1)\hat{S}_c^2}{(n_t + n_c - 2)}}}$$

where,

$\hat{\beta}$ is the estimated intervention effect obtained from the three-level impact model.

n_t is the number of students in the intervention group.

n_c is the number of students in the control group.

$\hat{S}_t^2$ is the posttest unadjusted student-level standard deviations for the intervention group.

$\hat{S}_c^2$ is the posttest unadjusted student-level standard deviations for the control group.

The multilevel model results for the reading vocabulary subtest and reading comprehension subtest of the Gates-MacGinitie Reading Tests, addressing the two primary research questions, are presented in Table C3.1. The multilevel model results for the MARSI and MRQ, addressing the two ancillary research questions, are shown in Table C3.2. (For both, see the highlighted row labeled "Thinking Reader," indicating that posttest differences between the intervention and control groups were not reliably different from 0.)

Table C3.1 Impact Results on Reading Achievement Based on Listwise Deletion of Missing Data

	GMRT: Vocabulary subtest			**GMRT: Comprehension subtest**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	517.00	0.66	.00	506.52	0.93	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–5.22	1.30	.00	–5.22	1.82	.01
Mid-sized: Enrollment 440–575 (vs. less than 440)	–4.30	1.62	.01	–3.16	2.30	.18
Large-sized: Enrollment greater than 575 (vs. less than 440)	–5.00	1.57	.00	–4.47	2.23	.06
Teacher-level covariates						
Thinking Reader	–1.24	1.31	.35	0.90	1.72	.61
Years of experience 4–20 (vs. more than 20)	2.60	2.10	.22	–1.56	2.81	.58
Years of experience less than 4 (vs. more than 20)	6.29	3.09	.05	1.75	4.13	.67
Master's degree or higher (vs. bachelor's)	0.41	1.68	.81	–2.20	2.24	.33
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–8.27	1.66	.00	–7.86	1.98	.00
English language learner (vs. non-English language learner)	–3.71	1.55	.02	–1.04	1.83	.57
Pretest score	0.80	0.01	.00	0.71	0.02	.00
Random effect	**Variance**	**χ^2(*df*)**	***p* value**	**Variance**	**χ^2(*df*)**	***p* value**
Level 1	368.46			504.21		
Level 2	8.98	42.37 (24)	.01	23.57	68.44 (24)	.00
Level 3	2.87	42.36 (28)	.04	8.08	45.22 (28)	.02
Thinking Reader achievement effect	12.16	42.06 (31)	.09	17.50	34.61 (31)	.30
N	2,147			2,140		

Note: GMRT is Gates-MacGinitie Reading Tests.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table C3.2 Impact Results on Reading Strategies and Reading Motivation Based on Listwise Deletion of Missing Data

	Reading strategies: MARSI			Reading motivation: MRQ		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.09	0.03	.00	2.77	0.01	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	0.06	0.05	.22	0.05	0.03	.08
Mid-sized: Enrollment 440–575 (vs. less than 440)	0.00	0.06	.95	0.00	0.03	.90
Large-sized: Enrollment greater than 575 (vs. less than 440)	0.00	0.06	1.00	–0.01	0.03	.58
Teacher-level covariates						
Thinking Reader	–0.00	0.04	.99	–0.01	0.03	.62
Years of experience 4–20 (vs. more than 20)	–0.07	0.07	.29	–0.02	0.04	.62
Years of experience less than 4 (vs. more than 20)	0.03	0.10	.73	0.03	0.06	.62
Master's degree or higher (vs. bachelor’s)	0.08	0.05	.15	0.03	0.03	.33
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–0.06	0.05	.19	–0.02	0.03	.44
English language learner (vs. non-English language learner)	–0.01	0.05	.80	–0.05	0.03	.07
Pretest score	0.56	0.02	.00	0.66	0.02	.00
Random effect	**Variance**	**$\chi^2(df)$**	***p* value**	**Variance**	**$\chi^2(df)$**	***p* value**
Level 1	0.35			0.14		
Level 2	0.01	59.93 (25)	.00	0.00	50.89 (25)	.00
Level 3	0.01	65.74 (28)	.00	0.00	53.70 (28)	.00
Thinking Reader achievement effect	0.01	34.21 (31)	.32	0.00	39.41 (31)	.14
N	2,201			2,208		

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; MARSI and MRQ surveys administered by study team.

Appendix D. Sensitivity Analyses

Appendix D reports on the analyses conducted to determine the sensitivity of the impact estimates to the assumptions made. These include looking at findings for schools with even and uneven number of teachers (Appendix D1), using two multiple imputation models (Appendix D2), an analysis using a fixed-effect treatment coefficient (Appendix D3), an analysis using a two-level model (Appendix D4), an analysis using a reduced sample of one class per teacher (Appendix D5), and an analysis on a reduced sample excluding students whose teachers left the sample (Appendix D6).

D1. Sensitivity Analysis to Different Randomization Procedures Across Schools With Even and Uneven Numbers of Teachers

Sensitivity Analysis 1: Excluding schools with uneven number of teachers. In this sensitivity analysis, the benchmark impact models were run with only the 23 schools that had an even number of classrooms (dropping 9 schools with an odd number of classrooms).

Table D1.1 Impact Results on Reading Achievement Based on Subsample of 23 Schools With Even Number of Teachers

	GMRT: vocabulary subtest			GMRT: comprehension subtest		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	516.49	0.68	.00	505.67	1.20	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–3.03	1.33	.04	–5.08	2.38	.05
Mid-sized: Enrollment 440–575 (vs. less than 440)	–2.09	1.88	.28	–3.40	3.14	.29
Large-sized: Enrollment greater than 575 (vs. less than 440)	–3.39	1.96	.10	–4.13	3.34	.23
Teacher-level covariates						
Thinking Reader	–0.85	1.39	.55	0.71	1.96	.72
Years of experience 4–20 (vs. more than 20)	2.85	2.47	.25	0.18	3.71	.96
Years of experience less than 4 (vs. more than 20)	3.68	3.88	.35	7.11	5.67	.22
Master's degree or higher (vs. bachelor’s)	0.90	1.77	.62	–0.22	2.59	.93
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–8.05	2.07	.00	–6.04	2.50	.02
English language learner (vs. non-English language learner)	–5.31	1.75	.00	–1.78	2.20	.42
Pretest score	0.79	0.02	.00	0.72	0.02	.00
Random effect	**Variance**	**χ^2(*df*)**	***p* value**	**Variance**	**χ^2(*df*)**	***p* value**
Level 1	344.82			489.76		
Level 2	4.55	18.23(11)	.08	25.23	38.75(11)	.00
Level 3	0.96	26.25(19)	.12	10.75	35.37(19)	.01
Thinking Reader achievement effect	8.83	31.12(22)	.09	6.77	20.82(22)	> .500
N	1,440			1,434		

Note: GMRT is Gates-MacGinitie Reading Tests.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table D1.2 Impact Results on Reading Strategies and Reading Motivation Based on Subsample of 23 Schools With an Even Number of Teachers

	Reading strategies: MARSI			**Reading motivation: MRQ**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.10	0.03	.00	2.77	0.01	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	0.01	0.05	.86	0.01	0.02	.68
Mid-sized: Enrollment 440–575 (vs. less than 440)	–0.02	0.07	.82	–0.06	0.04	.10
Large-sized: Enrollment greater than 575 (vs. less than 440)	–0.03	0.08	.74	–0.08	0.04	.05
Teacher-level covariates						
Thinking Reader	0.01	0.04	.84	–0.01	0.03	.77
Years of experience 4–20 (vs. more than 20)	–0.08	0.08	.34	–0.05	0.04	.22
Years of experience less than 4 (vs. more than 20)	–0.08	0.12	.50	–0.05	0.07	.45
Master's degree or higher (vs. bachelor’s)	0.03	0.06	.64	0.00	0.03	.95
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–0.07	0.06	.23	0.00	0.04	.92
English language learner (vs. non-English language learner)	–0.02	0.06	.72	–0.08	0.03	.01
Pretest score	0.54	0.02	.00	0.64	0.02	.00
Random effect	**Variance**	**χ^2(*df*)**	***p* value**	**Variance**	**χ^2(*df*)**	***p* value**
Level 1	0.35			0.14		
Level 2	0.01	19.85(12)	.07	0.00	21.48(12)	.04
Level 3	0.01	46.03(19)	.00	0.00	33.70(19)	.02
Thinking Reader achievement effect	0.01	27.66(22)	.19	0.01	42.91(22)	.01
N	1,486			1,492		

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; MARSI and MRQ surveys administered by study team.

Sensitivity Analysis 2: Controlling for indicators that capture whether the school has an even or uneven number of teachers. In this sensitivity analysis, we ran the impact models with all schools, adding a dummy covariate at the school-level (or Level 3) named “Odd_Teacher.” This dummy is coded 1 for the 9 schools that had an uneven number of participating teachers (each with a higher likelihood of being randomized to treatment than control) and 0 for the 23 schools that had an even number of participating teachers at the time of the randomization.

A sequence of three-level models (Models 1–3) was conducted. For each sequence, Level 1 represents the between-student and within-teacher and school model. The Level 2 represents the between-teacher and within-school model and the Level 3 represents the between-school model.

In Model 1, the equations for Levels 1 and 2 are exactly the same as the ones of the impact model. The Level 3 is different from the impact model. In Model 1, the Level 3 intercept equation includes only the new indicator Odd_Teacher. The following Model 2 is similar to Model 1, but in Model 2, the new indicator Odd_Teacher is also included in the equation of the treatment coefficient. This model tests whether the effect of the program is conditional on whether the school had an even or uneven number of participating teachers.

Finally, Model 3 includes the new indicator in the intercept equation as well as the other three school covariates also included at the impact model. The new Odd_Teacher dummy is removed from the coefficient of the treatment indicator because the interaction was not statistically significant in Model 2. The equations for these models are presented below.

Model 1. Including new “Odd_Teacher” indicator at the intercept of Level 3

Level 1 (between-student, within-teacher and school model)

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\,(\text{Pretest}_{ijk} - \text{Pretest}\ldots) + \pi_{2jk}\,(\text{IEP}_{ijk} - \text{IEP}\ldots) + \pi_{3jk}\,(\text{ELL}_{ijk} - \text{ELL}\ldots) + \varepsilon_{ijk}$,

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Level 2 (between-teacher, within-school model)

$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,(\text{Trt}_{jk} - \text{Trt}_{.k}) + \beta_{02k}\,(\text{Master}_{jk} - \text{Master}_{.k}) + \beta_{03k}\,(\text{Exp\_low}_{jk} - \text{Exp\_low}_{.k})$

$+ \beta_{04k}\,(\text{Exp\_Med}_{jk} - \text{Exp\_Med}_{.k}) + U_{0jk}$, where $U_{0jk} \sim N(0, \tau_{\pi 00})$

$\pi_{1jk} = \beta_{10k}$

$\pi_{2jk} = \beta_{20k}$

$\pi_{3jk} = \beta_{30k}$

Level 3 (between-school model)

$\beta_{00k} = \gamma_{000} + \gamma_{001}\,(\text{Odd\_Teacher}_{k} - \text{Odd\_Teacher.}) + V_{00k}$

$\beta_{01k} = \gamma_{010} + V_{01k}$

$\beta_{02k} = \gamma_{020}$

$\beta_{03k} = \gamma_{030}$

$\beta_{04k} = \gamma_{040}$

$\beta_{10k} = \gamma_{100}$

$\beta_{20k} = \gamma_{200}$

$\beta_{30k} = \gamma_{300}$

where $\begin{pmatrix} V_{00k} \\ V_{01k} \end{pmatrix} \sim N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta 00} & \tau_{\beta 01} \\ \tau_{\beta 10} & \tau_{\beta 11} \end{pmatrix}\right]$

Model 2. Including new "Odd_Teacher" indicator at the intercept and the coefficient of the treatment indicator at Level 3

Level 1 (between-student, within-teacher and school model)

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}$ (Pretest$_{ijk}$ – Pretest…) + π_{2jk} (IEP$_{ijk}$ – IEP…) + π_{3jk} (ELL$_{ijk}$ – ELL…) + ε_{ijk},

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Level 2 (between-teacher, within-school model)

$\pi_{0jk} = \beta_{00k} + \beta_{01k}$ (Trt$_{jk}$ – Trt.$_{k}$) + β_{02k} (Master$_{jk}$ – Master.$_{k}$) + β_{03k} (Exp_low$_{jk}$ – Exp_low.$_{k}$)

+ β_{04k} (Exp_Med$_{jk}$ – Exp_Med.$_{k}$) + U_{0jk}, where $U_{0jk} \sim N(0, \tau_{\pi 00})$

$\pi_{1jk} = \beta_{10k}$

$\pi_{2jk} = \beta_{20k}$

$\pi_{3jk} = \beta_{30k}$

Level 3 (between-school model)

$\beta_{00k} = \gamma_{000} + \gamma_{001}$ (Odd_Teacher$_k$ – Odd_Teacher.) + V_{00k}

$\beta_{01k} = \gamma_{010} + \gamma_{011}$ (Odd_Teacher$_k$ – Odd_Teacher.) + V_{01k}

$\beta_{02k} = \gamma_{020}$

$\beta_{03k} = \gamma_{030}$

$\beta_{04k} = \gamma_{040}$

$\beta_{10k} = \gamma_{100}$

$\beta_{20k} = \gamma_{200}$

$\beta_{30k} = \gamma_{300}$

where $\begin{pmatrix} V_{00k} \\ V_{01k} \end{pmatrix} \sim N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta 00} & \tau_{\beta 01} \\ \tau_{\beta 10} & \tau_{\beta 11} \end{pmatrix}\right]$

Model 3. Including new "Odd_Teacher" indicator at the intercept together with other school level covariates[40]

Level 1 (between-student, within-teacher and school model)

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}$ (Pretest$_{ijk}$ – Pretest…) + π_{2jk} (IEP$_{ijk}$ – IEP…) + π_{3jk} (ELL$_{ijk}$ – ELL…) + ε_{ijk},

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Level 2 (between-teacher, within-school model)

$\pi_{0jk} = \beta_{00k} + \beta_{01k}$ (Trt$_{jk}$ – Trt.$_{k}$) + β_{02k} (Master$_{jk}$ – Master.$_{k}$) + β_{03k} (Exp_low$_{jk}$ – Exp_low.$_{k}$)

+ β_{04k} (Exp_Med$_{jk}$ – Exp_Med.$_{k}$) + U_{0jk}, where $U_{0jk} \sim N(0, \tau_{\pi 00})$

$\pi_{1jk} = \beta_{10k}$

$\pi_{2jk} = \beta_{20k}$

$\pi_{3jk} = \beta_{30k}$

Level 3 (between-school model)

$\beta_{00k} = \gamma_{000} + \gamma_{001}$ (Odd_Teacher$_{k}$ – Odd_Teacher.) + γ_{002} (Hlunch$_{k}$ – Hlunch.) + γ_{003} (Midsize$_{k}$ – Midsize.) + γ_{004} (Lsize$_{k}$ – Lsize.) + V_{00k}

$\beta_{01k} = \gamma_{010} + V_{01k}$

$\beta_{02k} = \gamma_{020}$

$\beta_{03k} = \gamma_{030}$

$\beta_{04k} = \gamma_{040}$

$\beta_{10k} = \gamma_{100}$

$\beta_{20k} = \gamma_{200}$

$\beta_{30k} = \gamma_{300}$

where $$\begin{pmatrix} V_{00k} \\ V_{01k} \end{pmatrix} \sim N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta 00} & \tau_{\beta 01} \\ \tau_{\beta 10} & \tau_{\beta 11} \end{pmatrix}\right]$$

[40] We checked the extent to which the new dummy included at Level 3 was related to the other three indicators included in Model 3. All Spearman's correlation tests between new Odd_Teacher dummy and the other three school-level indicators were statistically non-significant.

Table D1.3 Impact Results for the Gates-MacGinitie Reading Tests Vocabulary Subtest Controlling for Indicators That Capture Whether the School Has an Even or Uneven Number of Teachers

	Model 1			**Model 2**			**Model 3**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	517.14	0.89	.00	517.13	0.89	.00	517.02	0.66	.00
School-level covariates									
Odd_Teacher	1.41	1.82	.45	1.88	1.92	.34	–0.24	1.40	.87
More than 74% of enrollment eligible for free or reduced-price lunch							–5.23	1.30	.00
Mid-sized: Enrollment 440–575 (vs. less than 440)							–4.40	1.73	.02
Large-sized: Enrollment greater than 575 (vs. less than 440)							–5.00	1.58	.00
Teacher-level covariates									
Thinking Reader	–1.14	1.27	.38	–1.09	1.26	.40	–1.30	1.31	.33
Years of experience 4–20 (vs. more than 20)	1.89	2.10	.37	1.55	2.13	.47	2.64	2.10	.21
Years of experience less than 4 (vs. more than 20)	5.51	3.08	.08	4.64	3.27	.16	6.28	3.09	.05
Master's degree or higher (vs. bachelor’s)	0.28	1.68	.87	0.12	1.69	.94	0.38	1.68	.82
Odd_Teacher				–2.23	2.84	.44			
Student-level covariates									
Individualized education program (vs. non-individualized education program)	–7.70	1.67	.00	–7.64	1.67	.00	–8.25	1.67	.00
English language learner (vs. non-English language learner)	–4.00	1.58	.01	–4.02	1.58	.01	–3.74	1.55	.02
Pretest score	0.81	0.01	.00	0.81	0.01	.00	0.80	0.01	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	368.71			368.69			368.52		
Level 2	10.10	41.04(22)	.01	10.05	41.56(22)	.01	8.81	40.20(22)	.01
Level 3	14.25	77.83(30)	.00	14.26	77.95(30)	.00	2.85	42.38(27)	.03
Thinking Reader achievement effect	8.74	39.61(31)	.14	8.00	38.88(30)	.13	12.48	42.30(31)	.09
N	2,147			2,147			2,147		

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Gates-MacGinitie Reading Tests vocabulary subtest administered by study team.

Table D1.4 Impact Results for the Gates-MacGinitie Reading Tests Comprehension Subtest Controlling for Indicators That Capture Whether the School Has an Even or Uneven Number of Teachers

	Model 1			Model 2			Model 3		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	506.66	1.09	.00	506.65	1.09	.00	506.59	0.94	.00
School-level covariates									
Odd_Teacher	0.19	2.18	.93	0.76	2.34	.75	–1.33	2.02	.51
More than 74% of enrollment eligible for free or reduced-price lunch							–5.33	1.84	.01
Mid-sized: Enrollment 440–575 (vs. less than 440)							–3.66	2.44	.15
Large-sized: Enrollment greater than 575 (vs. less than 440)							–4.56	2.24	.05
Teacher-level covariates									
Thinking Reader	0.94	1.75	.59	0.98	1.75	.58	0.79	1.73	.65
Years of experience 4–20 (vs. more than 20)	–2.33	2.78	.41	–2.70	2.84	.34	–1.58	2.80	.57
Years of experience less than 4 (vs. more than 20)	1.01	4.11	.81	0.04	4.36	.99	1.64	4.12	.69
Master's degree or higher (vs. bachelor's)	–2.08	2.23	.36	–2.26	2.25	.32	–2.11	2.24	.35
Odd_Teacher				–2.60	3.93	.51			
Student-level covariates									
Individualized education program (vs. non–individualized education program)	–7.43	1.97	.00	–7.38	1.97	.00	–7.80	1.98	.00
English language learner (vs. non–English language learner)	–1.75	1.84	.34	–1.78	1.84	.33	–1.08	1.83	.56
Pretest score	0.71	0.02	.00	0.71	0.02	.00	0.71	0.02	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	504.28			504.20			503.88		
Level 2	23.16	63.60(22)	.00	23.07	63.84(22)	.00	23.13	63.16(22)	.00
Level 3	19.18	64.89(30)	.00	19.24	65.17(30)	.00	8.39	44.61(27)	.02
Thinking Reader achievement effect	22.77	37.82(31)	.19	22.74	37.49(30)	.16	18.68	37.74(31)	.19
N	2,140			2,140			2,140		

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Gates-MacGinitie Reading Tests comprehension subtest administered by study team.

Table D1.5 Impact Results for Metacognitive Awareness of Reading Strategies Inventory Controlling for Indicators That Capture Whether the School Has an Even or Uneven Number of Teachers

	Model 1			**Model 2**			**Model 3**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.09	0.03	.00	3.09	0.03	.00	3.09	0.03	.00
School-level covariates									
Odd_Teacher	–0.04	0.05	.48	–0.03	0.06	.56	–0.03	0.06	.58
More than 74% of enrollment eligible for free or reduced-price lunch							0.06	0.05	.24
Mid-sized: Enrollment 440–575 (vs. less than 440)							–0.01	0.07	.90
Large-sized: Enrollment greater than 575 (vs. less than 440)							0.00	0.06	.96
Teacher-level covariates									
Thinking Reader	–0.01	0.04	.83	–0.01	0.04	.84	–0.01	0.04	.88
Years of experience 4–20 (vs. more than 20)	–0.07	0.07	.32	–0.07	0.07	.30	–0.07	0.07	.31
Years of experience less than 4 (vs. more than 20)	0.03	0.10	.75	0.02	0.11	.84	0.03	0.10	.74
Master's degree or higher (vs. bachelor's)	0.08	0.06	.13	0.08	0.06	.14	0.08	0.06	.13
Odd_Teacher				–0.03	0.09	.79			
Student-level covariates									
Individualized education program (vs. non-individualized education program)	–0.06	0.05	.17	–0.06	0.05	.17	–0.06	0.05	.19
English language learner (vs. non-English language learner)	–0.01	0.05	.83	–0.01	0.05	.83	–0.01	0.05	.76
Pretest score	0.56	0.02	.00	0.56	0.02	.00	0.55	0.02	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	0.35			0.35			0.35		
Level 2	0.01	59.15(23)	.00	0.01	59.59(23)	.00	0.01	58.95(23)	.00
Level 3	0.01	67.86(30)	.00	0.01	68.00(30)	.00	0.01	61.84(27)	.00
Thinking Reader achievement effect	0.01	37.59(31)	.19	0.01	37.40(30)	.17	0.01	37.18(31)	.21
N	2,201			2,201			2,201		

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Metacognitive Awareness of Reading Strategies Inventory administered by study team.

Table D1.6 Impact Results for Motivation for Reading Questionnaire Controlling for Indicators That Captures Whether the School Has an Even or Uneven Number of Teachers

	Model 1			**Model 2**			**Model 3**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	2.77	0.01	.00	2.77	0.01	.00	2.77	0.01	.00
School-level covariates									
Odd_Teacher	–0.03	0.03	.31	–0.02	0.03	.45	–0.02	0.03	.41
More than 74% of enrollment eligible for free or reduced-price lunch							0.05	0.03	.09
Mid-sized: Enrollment 440–575 (vs. less than 440)							–0.02	0.04	.67
Large-sized: Enrollment greater than 575 (vs. less than 440)							–0.02	0.03	.51
Teacher-level covariates									
Thinking Reader	–0.02	0.03	.51	–0.02	0.03	.52	–0.01	0.03	.61
Years of experience 4–20 (vs. more than 20)	–0.02	0.04	.64	–0.02	0.04	.58	–0.02	0.04	.64
Years of experience less than 4 (vs. more than 20)	0.02	0.06	.69	0.01	0.06	.82	0.03	0.06	.63
Master's degree or higher (vs. bachelor’s)	0.03	0.03	.31	0.03	0.03	.33	0.03	0.03	.30
Odd_Teacher				–0.03	0.06	.64			
Student-level covariates									
Individualized education program (vs. non-individualized education program)	–0.02	0.03	.39	–0.02	0.03	.40	–0.02	0.03	.46
English language learner (vs. non-English language learner)	–0.05	0.03	.09	–0.05	0.03	.09	–0.05	0.03	.07
Pretest score	0.67	0.02	.00	0.67	0.02	.00	0.66	0.02	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	0.14			0.14			0.14		
Level 2	0.00	50.99(23)	.00	0.00	51.17(23)	.00	0.00	50.91(23)	.00
Level 3	0.00	60.78(30)	.00	0.00	61.20(30)	.00	0.00	50.64(27)	.00
Thinking Reader achievement effect	0.01	42.95(31)	.08	0.01	42.83(30)	.06	0.01	42.27(31)	.09
N	2,208			2,208			2,208		

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Motivation for Reading Questionnaire administered by study team.

D2. Sensitivity Analysis: Impact Results Based on Multiple Imputation of Missing Data

Several analyses were conducted to determine how sensitive the impact estimates were to the assumptions made. The first analysis assessed the sensitivity of impact analyses run using listwise deletion as opposed to multiple imputation.

Introduction to multiple imputation. The results presented in Appendix C1 indicate statistically significant differences between students with complete and those with missing data.[41] On the basis of these patterns and other exploratory analyses explained below, this study assumes that missing patterns can be predicted by observed covariates. In other words, the probability of having missing data is assumed to be random after controlling for observed covariates.[42] In research on missing data, this type of data is commonly referred to as missing at random.[43] If the assumption of missing at random is true, many techniques can deal with missing data. To obtain accurate parameter estimates for the relationships of interest in the impact models, multiple imputation by chained equations was used to impute values for missing data for the four posttest and four pretest scores.[44]

Multiple imputation is a Monte Carlo technique in which missing values are replaced by $m > 1$ simulated versions. By convention, generating 5–10 datasets is sufficient for obtaining parameter estimates close to being fully efficient. In this study, m was set at 10. Each impact analysis was then run on each of the simulated and complete datasets, and results were later combined to generate estimates and confidence intervals that incorporate missing-data uncertainty (Schafer, 1999). The overall variance of the estimate is computed so that it accounts for the within- and between-imputation variance (Little & Rubin, 2002).

The multiple imputation by chained equations approach involves developing an imputation model and cycling through each of the variables with missing data and imputing them conditional on all the variables without missing data. The process starts with the variable having the fewest missing values and continues until no missing data remain. This process is then repeated multiple times using the new dataset until imputations stabilize (i.e., the order in which the variables are imputed no longer matters) and a single dataset with no missing data is created (Stuart, Azur, Frangakis, & Leaf, 2009). Finally, this entire process is repeated to create a number of imputed datasets without missing data.

Because the four posttest scores are included in the imputation model of the pretest score, and to avoid attenuation of the impact estimate, separate imputations were conducted for the intervention and control groups (Puma et al., 2009).

[41] The differences in percentages of students missing pretest or posttest scores between the intervention and control conditions were found to be not statistically significant.

[42] In other words, this assumption implies that missing on posttest does not depend on unobserved covariates after controlling for observed ones.

[43] The missing at random assumption, however, is not testable, nor is it possible to test whether levels of missing data on posttests depend on the values that are missing. For example, a student with low posttest scores may be more likely to be absent during posttesting.

[44] The imputation analysis was conducted using the Stata ice routine (imputation by chained equations).

Specifying the multiple imputation models. The four pretest and posttest scores and the race/ethnicity covariate were imputed. The race/ethnicity variable was imputed for use in the imputation models of the pretest and posttest scores. The percentage of cases imputed for race/ethnicity was 5.73% (or 138 of the total eligible sample of 2,407).

To make the missing at random assumption more reasonable and to keep the model as general as possible, the multiple imputation models included the covariates of the final analytical model plus other "auxiliary" covariates, or covariates that help to predict missing data but are not necessarily part of the impact model.

To build the imputation model and to choose the auxiliary variables, several exploratory analyses were conducted to identify the most important predictors of missing values on posttest scores. For these analyses, a missing indicator was generated to capture whether the student had any missing posttests; this indicator was used as the outcome in exploratory logistic regression models. Interaction terms were tested between all of the student covariates and between student covariates and pretest scores. All of the main effect and interaction terms that were found to be statistically significant predictors of the missing patterns on posttest scores were included in the imputation model.

Because missing values on posttest scores can be explained by characteristics at the student level and by the students' teacher or school membership, teacher and school covariates in the imputation model were further tested.[45]

To account for covariates from different levels of aggregation, models were explored that included different forms of the teacher and school covariates after controlling for the student predictors. Additionally, cluster indicators were tested at the teacher and school levels. Because the missing-data research does not explain how to account for multilevel data, two imputation models were specified to check the sensitivity of the results to different model specifications.

The first multiple imputation model, Model A, includes the student-level covariates individualized education program (IEP) status, English language learner (ELL) status, gender, race/ethnicity dummy variables, and age, as well as interactions between gender and race/ethnicity, ELL status and race/ethnicity, IEP status and race/ethnicity, and age and a pretest composite achievement score of reading vocabulary and reading comprehension called "reading total." Additional key student covariates are pretest scores (reading total, reading vocabulary, reading comprehension, Metacognitive Awareness of Reading Strategies Inventory [MARSI], and Motivation for Reading Questionnaire [MRQ]) and the corresponding posttest scores.[46] In Model A, teacher membership is accounted for by including teacher dummy variables (in cases with perfect colinearity, the variables were dropped) and some school covariates—such as whether the school is a middle or elementary school, school enrollment size dummy variables, percentage of students eligible for free or reduced-price lunch dummy variables, and racial/ethnic composition percentages.

[45] Additionally, the relative importance of some student covariates in predicting the missing patterns may vary from teacher to teacher or site to site.

[46] The strategy of including all of the dependent variables in the model when imputing any particular outcome is thought to protect against omitted variable bias in the imputation model (Puma et al., 2009).

The second imputation model, Model B, includes the same student covariates as Model A; some teacher covariates, such as an indicator whether the teacher has a master's degree and dummy variables for total teaching experience; and a continuous variable that captures total teaching experience in Grade 6. Also in Model B, school membership is accounted for by including school dummy variables.

Diagnostics. This section presents the results of diagnostic analyses for imputation Models A and B.

Graphic diagnostics. Figures D2.1–D2.3 compare the distributions of observed values, imputed values, and combined observed and imputed values through kernel density plots. The first line, "observed values only," represents the distribution of the posttest scores of the sample of students that did not have a posttest. The second line, "observed and imputed values," is the distribution of the outcomes for the complete dataset. And the third line, "imputed values," represents the distribution of the outcomes based on only the sample of students with missing data on posttest and whose outcome scores were imputed by the multiple imputation models. These plots were generated using 10 datasets, which multiplies the sample size by 10.

Diagnostic tests of imputed data are not absolute indicators of the success of the imputation but rather a means to confirm that the imputed data do not seem unreasonable. Differences between observed and imputed data do not necessarily mean that the specifications of the imputation model were incorrect. Substantive knowledge must be used in conjunction with the diagnostics when checking for potential misspecifications of the multiple imputation model (Stuart et al., 2009). Graphic diagnostics could flag a potential misspecification in the multiple imputation model if, for example, the distribution of "imputed values only" for a particular posttest is skewed greatly to the right (compared with the distribution of the "observed values only"), because students who were missing the posttest would not be expected to have higher imputed posttest values than students who had the posttest.

The figures below show that for reading comprehension, MARSI, and MRQ posttests, the imputed and observed distributions are similar (see Figures D2.1–D2.3), likely because the highest level of missing data was 11.4% for any of the posttests. Relative to the other posttests, a greater difference exists between the imputed and observed distributions for vocabulary posttest scores, with imputed values being more spread out and skewed farther to the left than the observed values (see Figure D2.1). However, the numeric diagnostics discussed below show that the multiple imputation models are appropriate and that similar diagnostic conclusions were found across all posttests for both treatment and control groups.

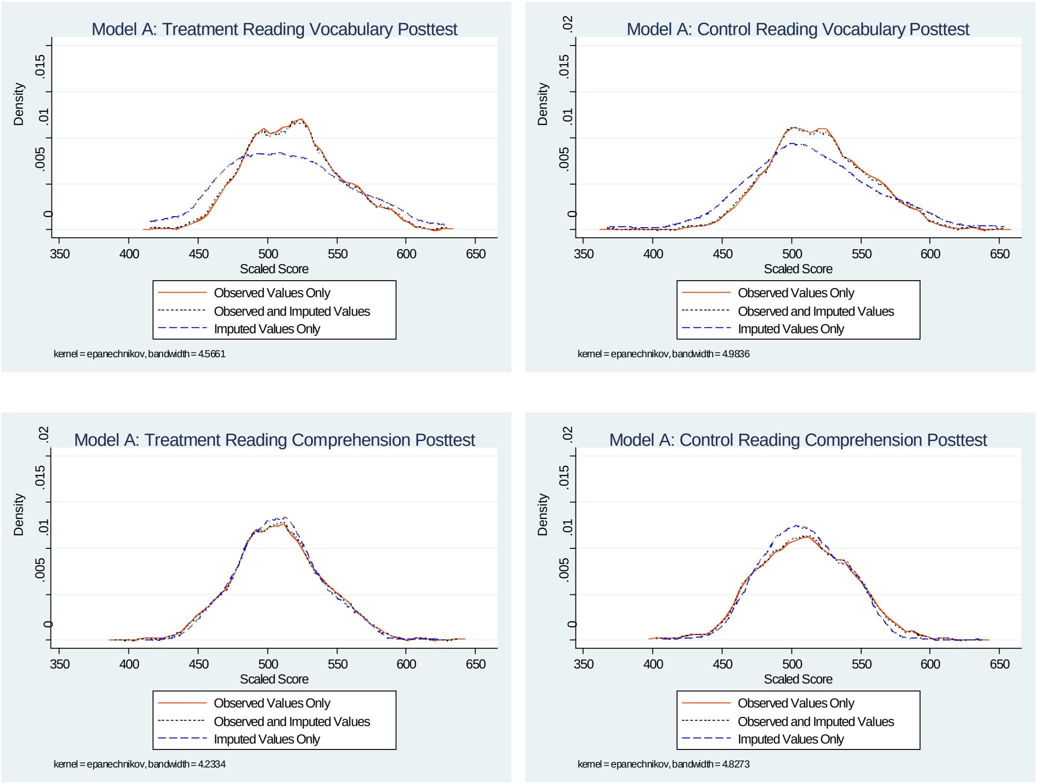
Graphic diagnostics were also conducted using pretest scores. Compared with posttests, greater variation was found between imputed and observed values. However, because only a small amount (1%) of data was missing on any of the pretests, the graphic diagnostics for these outcomes are not presented here.

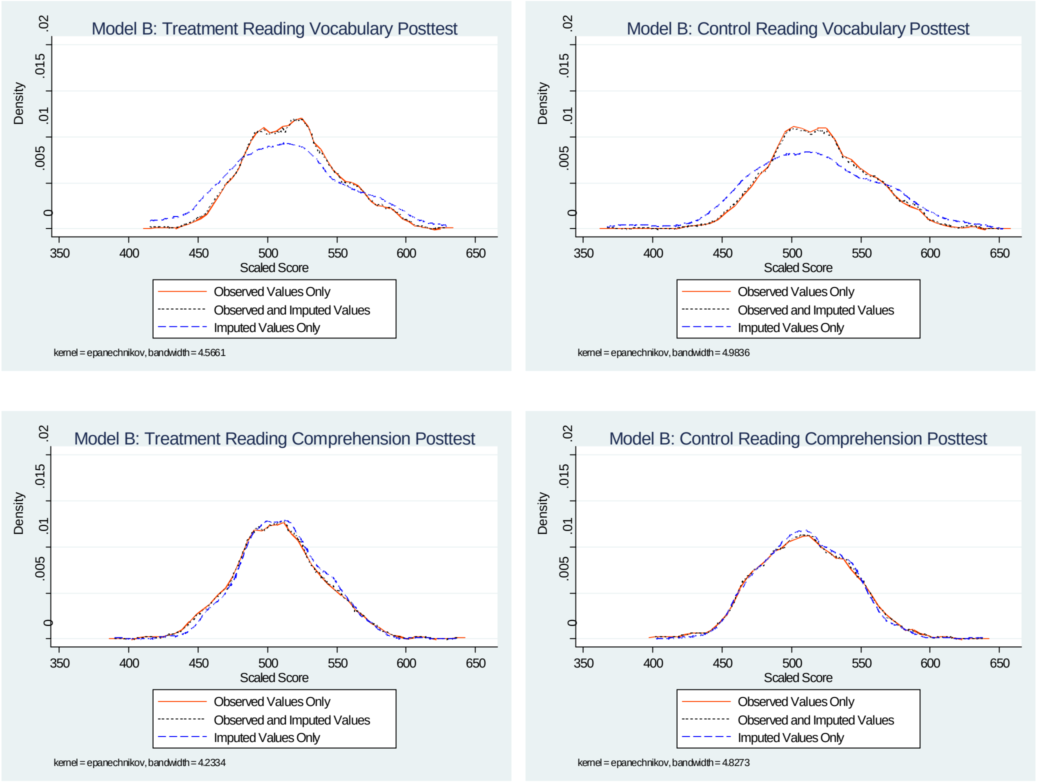
Numeric diagnostics. Numeric diagnostics analyze the means and standard deviations of observed values, imputed values, and combined observed and imputed values in search of differences that could indicate a problem with the imputation model.

Tables D2.1–D2.8 present the means and standard deviations for pretests and posttests and the ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values, by treatment and control conditions and for imputation Models A and B.

For the ratio, Stuart et al. (2009) suggest that an absolute value greater than 2 may indicate that the variable should be flagged for further investigation. The numeric diagnostics do not suggest misspecification of the multiple imputation models because none of the ratios from the imputed datasets approach this threshold, and differences in means and standard deviations are well within reason. Overall, numeric diagnostics parallel the graphic diagnostics.

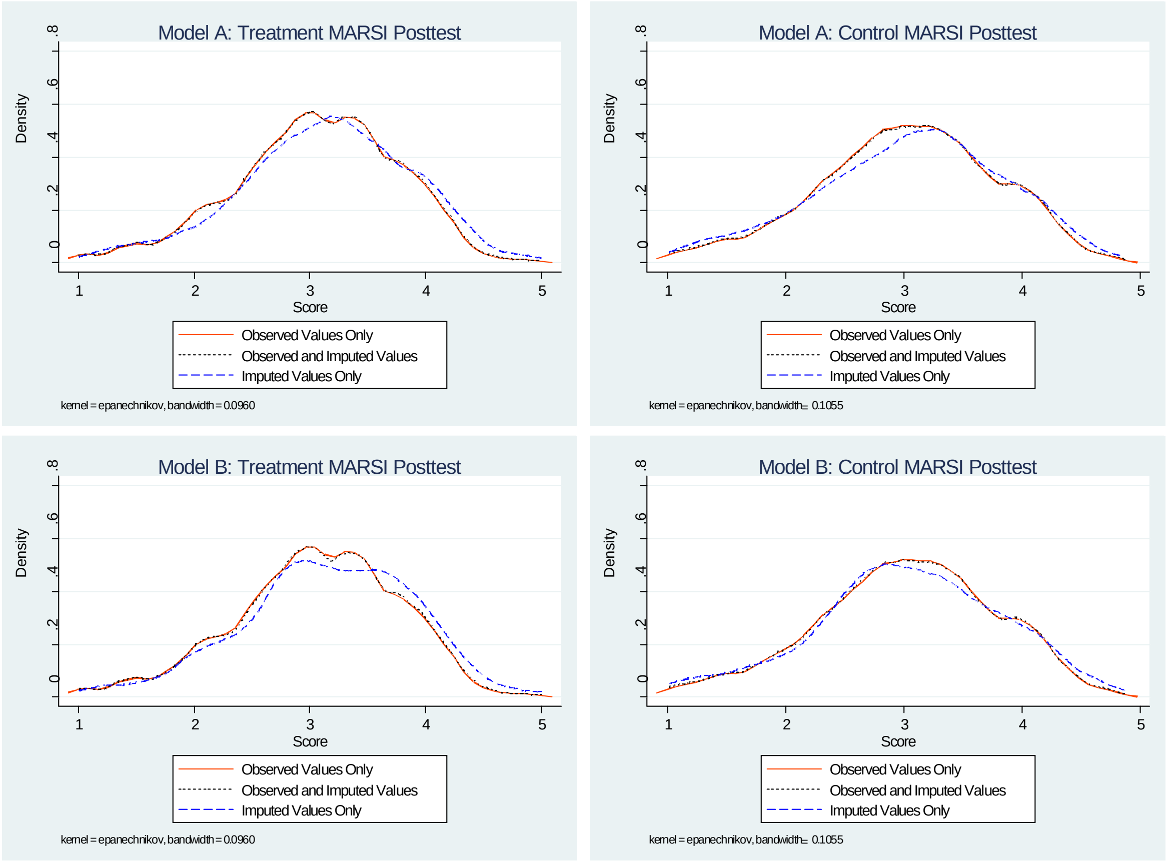
Figure D2.1 Multiple Imputation Using Models A and B: Graphic Diagnostics for Reading Achievement Outcomes





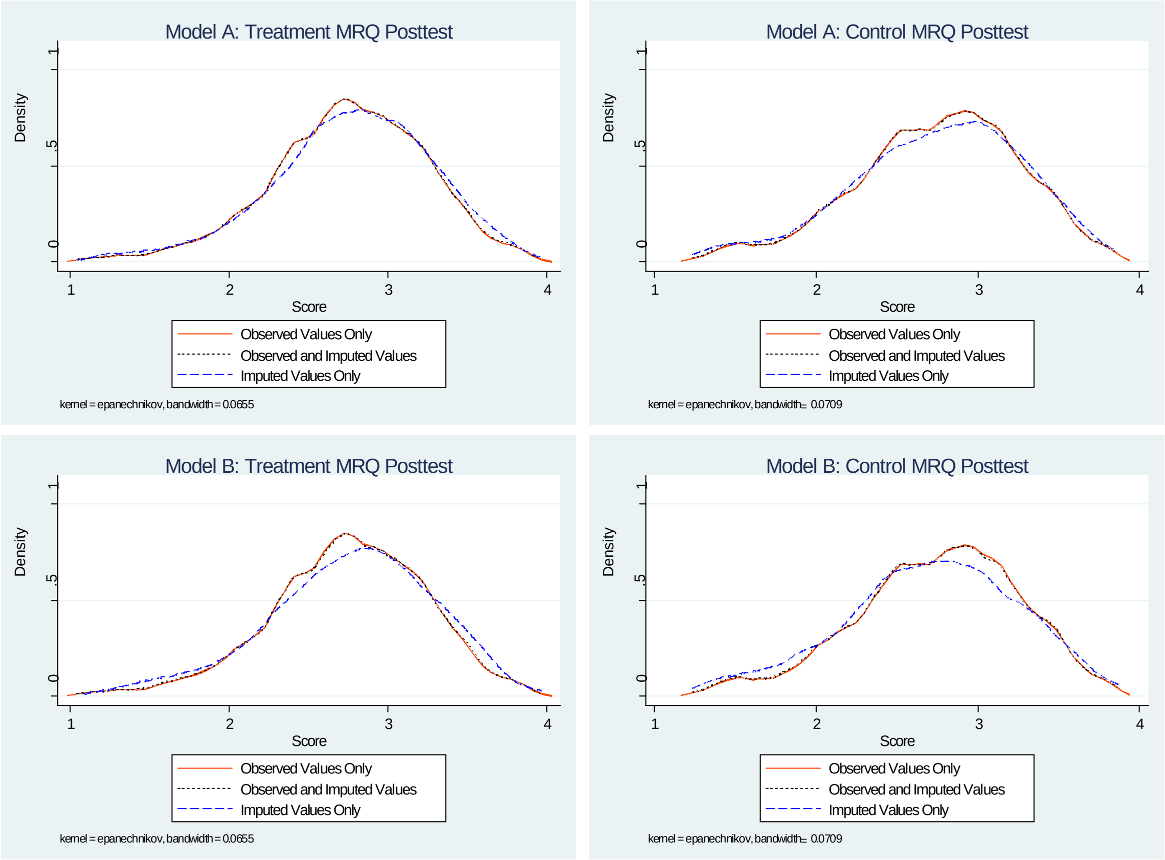
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Figure D2.2 Multiple Imputation Using Models A and B: Graphic Diagnostics for the Metacognitive Awareness of Reading Strategies Inventory



Source: Metacognitive Awareness of Reading Strategies Inventory survey administered by study team.

Figure D2.3 Multiple Imputation Using Models A and B: Graphic Diagnostics for the Motivation for Reading Questionnaire



Source: Motivation for Reading Questionnaire survey administered by study team.

Table D2.1 Multiple Imputation Using Models A and B: Numeric Diagnostics of Reading Achievement Pretests for Intervention Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	12,760	502.14	32.62	401	619		
Imputed values	100	509.66	39.41	412	619	0.23	1.21
Observed and imputed	12,860	502.20	32.68	401	619		
Reading comprehension							
Observed values	12,800	500.19	32.19	396	652		
Imputed values	60	493.32	32.96	396	574	–0.21	1.02
Observed and imputed	12,860	500.16	32.20	396	652		
Model B							
Reading vocabulary							
Observed values	12,760	502.14	32.62	401	619		
Imputed values	100	521.51	40.12	438	619	0.59	1.23
Observed and imputed	12,860	502.29	32.73	401	619		
Reading comprehension							
Observed values	12,800	500.19	32.19	396	652		
Imputed values	60	498.82	39.55	396	574	–0.04	1.23
Observed and imputed	12,860	500.18	32.23	396	652		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.
[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table D2.2 Multiple Imputation Using Models A and B: Numeric Diagnostics of Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Pretests for Intervention Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	12,780	3.18	0.67	1	5		
Imputed values	80	3.33	0.57	1.34	4.37	0.23	0.86
Observed and imputed	12,860	3.18	0.67	1	5		
Reading comprehension							
Observed values	12,750	2.83	0.47	1.25	3.96		
Imputed values	110	2.73	0.42	1.65	3.71	–0.21	0.90
Observed and imputed	12,860	2.83	0.47	1.25	3.96		
Model B							
Reading vocabulary							
Observed values	12,780	3.18	0.67	1	5		
Imputed values	80	3.38	0.59	1.5	4.9	0.30	0.89
Observed and imputed	12,860	3.18	0.67	1	5		
Reading comprehension							
Observed values	12,750	2.83	0.47	1.25	3.96		
Imputed values	110	2.73	0.44	1.45	3.65	–0.22	0.94
Observed and imputed	12,860	2.83	0.47	1.25	3.96		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.
[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.
Source: Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire surveys administered by study team.

Table D2.3 Multiple Imputation Using Models A and B: Numeric Diagnostics of Reading Achievement Pretests for Control Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	11,120	502.19	34.99	367	653		
Imputed values	90	505.19	56.74	367	653	0.09	1.62
Observed and imputed	11,210	502.21	35.21	367	653		
Reading comprehension							
Observed values	11,080	502.06	32.61	406	652		
Imputed values	130	511.91	42.46	409.5	652	0.30	1.30
Observed and imputed	11,210	502.17	32.76	406	652		
Model B							
Reading vocabulary							
Observed values	11,120	502.19	34.99	367	653		
Imputed values	90	516.63	55.73	378	653	0.41	1.59
Observed and imputed	11,210	502.30	35.22	367	653		
Reading comprehension							
Observed values	11,080	502.06	32.61	406	652		
Imputed values	130	514.96	45.05	423	652	0.40	1.38
Observed and imputed	11,210	502.21	32.81	406	652		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.
[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table D2.4 Multiple Imputation Using Models A and B: Numeric Diagnostics of Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Pretests for Control Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	11,100	3.15	0.71	1	5		
Imputed values	110	3.18	0.64	1.83	4.55	0.04	0.91
Observed and imputed	11,210	3.15	0.71	1	5		
Reading comprehension							
Observed values	11,120	2.85	0.48	1.19	3.94		
Imputed values	90	2.78	0.46	1.57	3.57	–0.14	0.95
Observed and imputed	11,210	2.85	0.48	1.19	3.94		
Model B							
Reading vocabulary							
Observed values	11,100	3.15	0.71	1	5		
Imputed values	110	3.18	0.62	1.00	4.55	0.04	0.87
Observed and imputed	11,210	3.15	0.71	1	5		
Reading comprehension							
Observed values	11,120	2.85	0.48	1.19	3.94		
Imputed values	90	2.86	0.42	1.57	3.67	0.03	0.88
Observed and imputed	11,210	2.85	0.48	1.19	3.94		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.

[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.

Source: Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire surveys administered by study team.

Table D2.5 Multiple Imputation Using Models A and B: Numeric Diagnostics of Reading Achievement Posttests for Intervention Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	11,600	518.07	33.61	415	629		
Imputed values	1,260	514.09	44.70	415	629	–0.12	1.33
Observed and imputed	12,860	517.68	34.87	415	629		
Reading comprehension							
Observed values	11,550	508.07	33.02	390	638		
Imputed values	1,310	507.98	30.27	412	629	0.00	0.92
Observed and imputed	12,860	508.06	32.75	390	638		
Model B							
Reading vocabulary							
Observed values	11,600	518.07	33.61	415	629		
Imputed values	1,260	514.56	42.91	415	629	–0.10	1.28
Observed and imputed	12,860	517.73	34.65	415	629		
Reading comprehension							
Observed values	11,550	508.07	33.02	390	638		
Imputed values	1,310	510.08	30.78	390	638	0.06	0.93
Observed and imputed	12,860	508.28	32.80	390	638		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.
[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table D2.6 Multiple Imputation Using Models A and B: Numeric Diagnostics of Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Posttests for Intervention Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	11,880	3.05	0.71	1	5		
Imputed values	980	3.13	0.75	1	5	0.11	1.06
Observed and imputed	12,860	3.06	0.71	1	5		
Reading comprehension							
Observed values	11,920	2.74	0.49	1.04	3.96		
Imputed values	940	2.76	0.51	1.04	3.96	0.04	1.04
Observed and imputed	12,860	2.75	0.50	1.04	3.96		
Model B							
Reading vocabulary							
Observed values	11,880	3.05	0.71	1	5		
Imputed values	980	3.14	0.75	1	5	0.13	1.06
Observed and imputed	12,860	3.06	0.71	1	5		
Reading comprehension							
Observed values	11,920	2.74	0.49	1.04	3.96		
Imputed values	940	2.75	0.53	1.08	3.96	0.02	1.07
Observed and imputed	12,860	2.74	0.50	1.04	3.96		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.
[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.
Source: Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire surveys administered by study team.

Table D2.7 Multiple Imputation Using Models A and B: Numeric Diagnostics of Reading Achievement Posttests for Control Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	9,960	519.42	36.23	367	653		
Imputed values	1,250	513.74	47.06	367	653	–0.16	1.30
Observed and imputed	11,210	518.79	37.63	367	653		
Reading comprehension							
Observed values	9,940	509.74	34.45	402	638		
Imputed values	1,270	508.48	30.16	407	638	–0.04	0.88
Observed and imputed	11,210	509.6	33.99	402	638		
Model B							
Reading vocabulary							
Observed values	9,960	519.42	36.23	367	653		
Imputed values	1,250	515.88	47.63	367	653	–0.10	1.31
Observed and imputed	11,210	519.03	37.69	367	653		
Reading comprehension							
Observed values	9,940	509.74	34.45	402	638		
Imputed values	1,270	509.92	32.22	402	638	0.01	0.94
Observed and imputed	11,210	509.76	34.21	402	638		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.
[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.
Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table D2.8 Multiple Imputation Using Models A and B: Numeric Diagnostics of Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire Posttests for Control Condition

Measure	Number	Mean	Standard deviation	Minimum	Maximum	Ratio of mean to standard deviation[a]	Standard deviation ratio[b]
Model A							
Reading vocabulary							
Observed values	10,290	3.02	0.76	1	4.87		
Imputed values	920	3.03	0.81	1	4.84	0.02	1.07
Observed and imputed	11,210	3.02	0.76	1	4.87		
Reading comprehension							
Observed values	10,330	2.75	0.51	1.23	3.87		
Imputed values	880	2.74	0.53	1.23	3.87	–0.01	1.05
Observed and imputed	11,210	2.75	0.51	1.23	3.87		
Model B							
Reading vocabulary							
Observed values	10,290	3.02	0.76	1	4.87		
Imputed values	920	3.02	0.80	1	4.87	0.00	1.06
Observed and imputed	11,210	3.02	0.76	1	4.87		
Reading comprehension							
Observed values	10,330	2.75	0.51	1.23	3.87		
Imputed values	880	2.70	0.55	1.23	3.87	–0.09	1.07
Observed and imputed	11,210	2.74	0.51	1.23	3.87		

[a] Ratio of the difference between the mean of imputed and observed values to the standard deviation of observed values.

[b] Ratio of the standard deviation of the imputed values to the standard deviation of the observed values.

Source: Metacognitive Awareness of Reading Strategies Inventory and Motivation for Reading Questionnaire surveys administered by study team.

Impact results using multiple imputation. Using multiple imputation, the overall conclusions about the impact of *Thinking Reader* are the same as those derived from the listwise deletion analysis. Analyses were not sensitive to the specification of the multiple imputation model used; conclusions about the impact of *Thinking Reader* are the same whether the 10 imputed datasets were generated using Model A (student and school covariates; teacher dummy variables) or Model B (student and teacher covariates; school dummy variables). The results based on multiple imputation Model A are presented in Tables D2.9 and D2.10, and those based on Model B are presented in Tables D2.11 and D2.12.

Table D2.9 Impact Results on Reading Achievement Based on Multiple Imputation Data From Model A

	GMRT: Vocabulary subtest			**GMRT: Comprehension subtest**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	518.08	0.75	.00	508.27	0.92	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–4.03	1.76	.03	–3.54	1.87	.07
Mid-sized: Enrollment 440–575 (vs. less than 440)	–3.58	1.95	.08	–2.98	2.20	.19
Large-sized: Enrollment greater than 575 (vs. less than 440)	–4.56	1.78	.02	–4.16	2.24	.07
Teacher-level covariates						
Thinking Reader	–1.47	1.82	.43	0.37	1.69	.83
Years of experience 4–20 (vs. more than 20)	2.95	2.53	.25	–0.43	2.72	.88
Years of experience less than 4 (vs. more than 20)	5.59	3.72	.14	1.91	4.04	.64
Master's degree or higher (vs. bachelor’s)	–0.20	2.04	.92	–2.75	2.23	.22
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–2.63	2.42	.29	–0.18	2.29	.94
English language learner (vs. non-English language learner)	–3.05	2.18	.17	–0.27	2.06	.90
Pretest score	0.75	0.02	.00	0.63	0.02	.00
Random effect	**Variance**	**χ^2(*df*)**	***p* value**	**Variance**	**χ^2(*df*)**	***p* value**
Level 1	584.66			618.44		
Level 2	4.14	35.34 (25)	.08	17.72	62.76 (25)	.00
Level 3	2.76	40.85 (28)	.06	6.48	42.73 (28)	.04
Thinking Reader achievement effect	28.46	54.24 (31)	.01	13.59	34.27 (31)	.31
N	2,407					

Note: Based on imputation Model A. GMRT is Gates-MacGinitie Reading Tests.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table D2.10 Impact Results on Reading Strategies and Reading Motivation Based on Multiple Imputation Data From Model A

	Reading strategies: MARSI			**Reading motivation: MRQ**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.09	0.03	.00	2.77	0.02	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	0.06	0.05	.23	0.06	0.03	.05
Mid-sized: Enrollment 440–575 (vs. less than 440)	–0.01	0.07	.92	–0.01	0.04	.85
Large-sized: Enrollment greater than 575 (vs. less than 440)	0.00	0.06	1.00	–0.02	0.03	.66
Teacher-level covariates						
Thinking Reader	–0.01	0.04	.91	–0.02	0.03	.57
Years of experience 4–20 (vs. more than 20)	–0.07	0.07	.34	–0.02	0.04	.55
Years of experience less than 4 (vs. more than 20)	0.04	0.11	.72	0.04	0.06	.57
Master's degree or higher (vs. bachelor’s)	0.09	0.06	.11	0.04	0.03	.27
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–0.05	0.05	.24	–0.01	0.03	.62
English language learner (vs. non-English language learner)	–0.01	0.05	.83	–0.05	0.03	.12
Pretest score	0.55	0.02	.00	0.66	0.02	.00
Random effect	**Variance**	**χ^2(*df*)**	***p* value**	**Variance**	**χ^2(*df*)**	***p* value**
Level 1	0.36			0.15		
Level 2	0.01	66.48 (25)	.00	0.00	55.04 (25)	.00
Level 3	0.01	63.93 (28)	.00	0.00	54.83 (28)	.00
Thinking Reader achievement effect	0.01	35.45 (31)	.27	0.01	42.12 (31)	.08
N	2,407			2,407		

Note: Based on imputation Model A. MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; MARSI and MRQ surveys administered by study team.

Table D2.11 Impact Results on Reading Achievement Based on Multiple Imputation Data From Model B

	GMRT: Vocabulary subtest			GMRT: Comprehension subtest		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	518.26	0.80	.00	508.44	0.93	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–4.11	1.55	.01	–3.57	1.80	.06
Mid-sized: Enrollment 440–575 (vs. less than 440)	–3.72	1.90	.06	–2.92	2.29	.21
Large-sized: Enrollment greater than 575 (vs. less than 440)	–4.97	2.17	.03	–4.01	2.28	.09
Teacher-level covariates						
Thinking Reader	–1.64	1.78	.36	0.29	1.67	.86
Years of experience 4–20 (vs. more than 20)	3.31	2.48	.19	–0.77	2.79	.78
Years of experience less than 4 (vs. more than 20)	5.82	3.66	.12	0.96	4.18	.82
Master's degree or higher (vs. bachelor's)	0.07	2.06	.97	–2.36	2.23	.29
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–2.81	2.61	.29	–0.02	2.37	.99
English language learner (vs. non-English language learner)	–3.60	2.28	.12	–0.21	2.03	.92
Pretest score	0.74	0.02	.00	0.63	0.02	.00
Random effect	**Variance**	**$\chi^2(df)$**	***p* value**	**Variance**	**$\chi^2(df)$**	***p* value**
Level 1	583.89			630.79		
Level 2	5.00	37.93 (25)	.05	18.70	60.86 (25)	.00
Level 3	3.21	41.47 (28)	.05	6.43	43.07 (28)	.03
Thinking Reader achievement effect	30.36	54.35 (31)	.01	13.27	34.18 (31)	.32
N	2,407			2,407		

Note: Based on imputation Model B. GMRT is Gates-MacGinitie Reading Tests.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table D2.12 Impact Results on Reading Strategies and Reading Motivation Based on Multiple Imputation Data From Model B

	Reading strategies: MARSI			Reading motivation: MRQ		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.09	0.03	.00	2.77	0.02	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	0.06	0.05	.26	0.05	0.03	.09
Mid-sized: Enrollment 440–575 (vs. less than 440)	0.01	0.07	.93	–0.00	0.04	.94
Large-sized: Enrollment greater than 575 (vs. less than 440)	0.01	0.06	.88	–0.01	0.04	.72
Teacher-level covariates						
Thinking Reader	–0.00	0.04	.98	–0.01	0.03	.67
Years of experience 4–20 (vs. more than 20)	–0.07	0.07	.29	–0.02	0.04	.60
Years of experience less than 4 (vs. more than 20)	0.03	0.10	.79	0.04	0.06	.54
Master's degree or higher (vs. bachelor’s)	0.09	0.06	.11	0.04	0.03	.27
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–0.06	0.04	.21	–0.02	0.03	.45
English language learner (vs. non-English language learner)	–0.01	0.05	.81	–0.05	0.03	.10
Pretest score	0.56	0.02	.00	0.66	0.02	.00
Random effect	**Variance**	**$\chi^2(df)$**	***p* value**	**Variance**	**$\chi^2(df)$**	***p* value**
Level 1	0.36			0.15		
Level 2	0.01	61.64 (25)	.00	0.00	48.89 (25)	.00
Level 3	0.01	66.51 (28)	.00	0.00	60.37 (28)	.00
Thinking Reader achievement effect	0.01	35.38 (31)	.27	0.01	43.08 (31)	.07
N	2,407			2,407		

Note: Based on imputation Model B. MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; MARSI and MRQ surveys administered by study team.

D3. Sensitivity Analysis: Treatment as a Fixed Effect

The second sensitivity analysis assessed the sensitivity of the impact estimate and standard error to modeling the treatment indicator as a fixed effect as opposed to a random effect. Here, findings from four three-level models (Models 0–4), which build sequentially to the impact model, are presented. The results of Model 4 show the sensitivity of the impact estimate and standard error to the decision to model the treatment as a fixed or random effect. In all models, the Level 1 covariates were grand mean centered, the Level 2 covariates were group mean centered, and the Level 3 predictors were grand mean centered.

Before including the treatment indicator, **Model 0,** commonly referred to as the fully unconditional model because no predictors are specified at either level, is presented. This model yields estimates of the initial variance components—that is, the variances of the random effects at the school level and the teacher level and the residual variance at the student level. The variance component estimates from this model are used to estimate the intraclass correlation coefficients of outcome responses at the school level and at the teacher level. The intraclass correlation captures the proportion of the variance in the outcome between teachers and between schools.

Level 1 model

$Y_{ijk} = \pi_{0jk} + \varepsilon_{ijk}$, where $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Level 2 model

$\pi_{0jk} = \beta_{00k} + U_{0jk}$, where $\varepsilon_{ijk} \sim N(0, \tau_{\pi 00})$

Level 3 model

$\beta_{00k} = \gamma_{000} + V_{0k}$, where $V_{0k} \sim N(0, \tau_{\beta 00})$

In **Model 1,** the coefficient of the treatment indicator is added as a fixed effect with no other covariates included.

Level 1 model

$Y_{ijk} = \pi_{0jk} + \varepsilon_{ijk}$

Level 2 model

$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,(\mathrm{Trt}_{jk} - \mathrm{Trt}_{.k}) + U_{0jk}$

Level 3 model

$\beta_{00k} = \gamma_{000} + V_{0k}$

$\beta_{01k} = \gamma_{010}$

Model 2 tests whether the coefficient of the treatment indicator varies systematically across schools. The variance parameter $\tau_{\beta 01}$ captures the extent to which the school treatment effects vary around γ_{010}.

Level 1 model

$Y_{ijk} = \pi_{0jk} + \varepsilon_{ijk}$

Level 2 model

$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,(\mathrm{Trt}_{jk} - \mathrm{Trt}._{k}) + U_{0jk}$

Level 3 model

$\beta_{00k} = \gamma_{000} + V_{00k}$

$\beta_{01k} = \gamma_{010} + V_{01k}$

where

$$\begin{pmatrix} V_{00k} \\ V_{01k} \end{pmatrix} \sim N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\beta 00} & \tau_{\beta 01} \\ \tau_{\beta 10} & \tau_{\beta 11} \end{pmatrix}\right]$$

Model 3 adds student and teacher covariates. A nonrandomly varying slope is assumed for these covariates. The treatment indicator is added as a fixed effect.

Level 1 model

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\,(\mathrm{Pretest}_{ijk} - \mathrm{Pretest}...) + \pi_{2jk}\,(\mathrm{IEP}_{ijk} - \mathrm{IEP}...) + \pi_{3jk}\,(\mathrm{ELL}_{ijk} - \mathrm{ELL}...) + \varepsilon_{ijk}$

Level 2 model

$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,(\mathrm{Trt}_{jk} - \mathrm{Trt}._{k}) + \beta_{02k}\,(\mathrm{Master}_{jk} - \mathrm{Master}._{k}) + \beta_{03k}\,(\mathrm{Exp\_low}_{jk} - \mathrm{Exp\_low}._{k})$

$+ \beta_{04k}\,(\mathrm{Exp\_Med}_{jk} - \mathrm{Exp\_Med}._{k}) + \mathrm{U}_{0jk}$

Level 3 model

$\beta_{00k} = \gamma_{000} + V_{00k}$

Model 4 adds school-level covariates. This model is very similar to the impact model presented in Chapter 4, but in this case, the coefficient of the treatment indicator is added as a fixed effect.

In light of the relatively small number of schools, a parsimonious Level 3 model was specified.[47]

Level 1 model

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\,(\text{Pretest}_{ijk} - \text{Pretest}_{...}) + \pi_{2jk}\,(\text{IEP}_{ijk} - \text{IEP}_{...}) + \pi_{3jk}\,(\text{ELL}_{ijk} - \text{ELL}_{...}) + \varepsilon_{ijk}$$

Level 2 model

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,(\text{Trt}_{jk} - \text{Trt}_{\cdot k}) + \beta_{02k}\,(\text{Master}_{jk} - \text{Master}_{\cdot k}) + \beta_{03k}\,(\text{Exp\_low}_{jk} - \text{Exp\_low}_{\cdot k})$$

$$+ \beta_{04k}\,(\text{Exp\_Med}_{jk} - \text{Exp\_Med}_{\cdot k}) + U_{0jk}$$

Level 3 model

$$\beta_{00k} = \gamma_{000} + \gamma_{001}\,(\text{Hlunch}_{k} - \text{Hlunch}_{.}) + \gamma_{002}\,(\text{Midsize}_{k} - \text{Midsize}_{.})$$

$$+ \gamma_{003}\,(\text{Lsize}_{k} - \text{Lsize}_{.}) + V_{00k}$$

The results of the analyses (Tables D3.1–D3.4) show that across all outcomes, the impact estimates and their standard errors were not sensitive to modeling the treatment as a fixed effect.

[47] Aside from the two Level 3 covariates included in the multilevel model, two others were tested: *elementary* (a dummy variable that takes the value of 1 for elementary schools and 0 for middle schools) and *state* (a dummy variable that captures whether the state is Connecticut or Rhode Island versus Massachusetts). The covariates in the multilevel model were selected after regressing the estimated Bayes residuals from Model 3 on all the potential Level 3 predictors.

Table D3.1 Treatment as a Fixed Effect: Results for Gates-MacGinitie Reading Tests Reading Vocabulary Subtest

	Model 0: Fully unconditional			**Model 1: Adding treatment indicator as fixed effect**			**Model 2: Adding treatment indicator as a random effect**			**Model 3: Adding student and teacher covariates and treatment indicator as fixed effect**			**Model 4: Model 3 + school covariates**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	515.48	2.73	.00	515.50	2.73	.00	515.50	2.72	.00	517.11	0.92	.00	517.07	0.64	.00
School-level covariates															
More than 74% of enrollment eligible for free or reduced-price lunch													–4.95	1.31	.00
Mid-sized: Enrollment 440–575 (vs. less than 440)													–4.90	1.63	.01
Large-sized: Enrollment greater than 575 (vs. less than 440)													–6.49	1.58	.00
Teacher-level covariates															
Thinking Reader				–3.06	2.34	.20	–2.15	2.62	.42	–1.06	1.21	.39	–1.06	1.23	.39
Years of experience 4–20 (vs. more than 20)										1.98	2.17	.36	2.11	2.19	.34
Years of experience less than 4 (vs. more than 20)										5.98	3.16	.06	6.18	3.20	.06
Master's degree or higher (vs. bachelor's)										–0.02	1.73	.99	0.20	1.75	.91
Student-level covariates															
Individualized education program (vs. non-individualized education program)										–7.70	1.68	.00	–8.18	1.68	.00
English language learner (vs. non-English language learner)										–3.95	1.58	.01	–3.57	1.55	.02
Pretest score										0.81	0.01	0.00	0.80	0.01	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	989.42			989.30			989.80			368.54			368.34		
Level 2	73.35	166.92 (59)	.00	70.39	165.16 (58)	.00	40.09	47.87 (27)	.01	12.96	111.65 (55)	.00	13.83	111.75 (55)	.00
Level 3	188.99	167.26 (31)	.00	190.17	171.50 (31)	.00	200.58	229.96 (31)	.00	14.94	73.79 (31)	.00	0.65	34.36 (28)	.19
Thinking Reader achievement effect	—			—			84.42	52.64 (31)	.01	—			—		
N	2,156			2,156			2,156			2,147			2,147		

— Indicates that the random effect was not estimated for that model.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Gates-MacGinitie Reading Tests vocabulary subtest administered by study team.

Table D3.2 Treatment as a Fixed Effect: Results for Gates-MacGinitie Reading Tests Reading Comprehension Subtest

	Model 0: Fully unconditional			Model 1: Adding treatment indicator as fixed effect			Model 2: Adding treatment indicator as a random effect			Model 3: Adding student and teacher covariates and treatment indicator as fixed effect			Model 4: Model 3 + school covariates		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	505.78	2.42	.00	505.79	2.42	.00	505.82	2.43	.00	506.67	1.07	.00	506.60	0.89	.00
School-level covariates															
More than 74% of enrollment eligible for free or reduced-price lunch													–5.27	1.80	.01
Mid-sized: Enrollment 440–575 (vs. less than 440)													–3.53	2.25	.13
Large-sized: Enrollment greater than 575 (vs. less than 440)													–5.16	2.19	.03
Teacher-level covariates															
Thinking Reader				–2.69	2.49	.28	–2.11	2.58	.42	0.78	1.64	.64	0.75	1.65	.65
Years of experience 4–20 (vs. more than 20)										–1.63	2.90	.58	–1.55	2.92	.60
Years of experience less than 4 (vs. more than 20)										1.24	4.25	.77	1.36	4.28	.75
Master's degree or higher (vs. bachelor's)										–2.64	2.32	.26	–2.39	2.33	.31
Student-level covariates															
Individualized education program (vs. non-individualized education program)										–7.34	1.98	.00	–7.74	1.99	.00
English language learner (vs. non-English language learner)										–1.59	1.84	.39	–0.77	1.82	.67
Pretest score										0.71	0.02	.00	0.71	0.02	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	947.35			947.33			947.52			504.74			504.52		
Level 2	90.17	188.70 (59)	.00	87.72	187.39 (58)	.00	76.83	84.16 (27)	.00	30.56	136.31 (55)	.00	31.34	136.65 (55)	.00
Level 3	133.61	115.07 (31)	.00	134.22	116.93 (31)	.00	139.20	126.46 (31)	.00	15.51	57.28 (31)	.00	2.97	38.49 (28)	.09
Thinking Reader achievement effect	—			—			29.69	32.55 (31)	.39	—			—		
N	2,149			2,149			2,149			2,140			2,140		

— Indicates that the random effect was not estimated for that model..

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Gates-MacGinitie Reading Tests comprehension subtest administered by study team.

Table D3.3 Treatment as a Fixed Effect: Results for Metacognitive Awareness of Reading Strategies Inventory

	Model 0: Fully unconditional			**Model 1: Adding treatment indicator as fixed effect**			**Model 2: Adding treatment indicator as a random effect**			**Model 3: Adding student and teacher covariates and treatment indicator as fixed effect**			**Model 4: Model 3 + school covariates**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.11	0.04	.00	3.11	0.04	.00	3.11	0.04	.00	3.08	0.03	.00	3.09	0.03	.00
School-level covariates															
More than 74% of enrollment eligible for free or reduced-price lunch													0.08	0.05	.14
Mid-sized: Enrollment 440–575 (vs. less than 440)													–0.02	0.07	.82
Large-sized: Enrollment greater than 575 (vs. less than 440)													–0.03	0.07	.64
Teacher-level covariates															
Thinking Reader				0.03	0.05	.58	–0.00	0.05	.99	0.02	0.04	.71	0.01	0.04	.72
Years of experience 4–20 (vs. more than 20)										–0.09	0.07	.21	–0.09	0.07	.21
Years of experience less than 4 (vs. more than 20)										0.01	0.10	.95	0.01	0.10	.94
Master's degree or higher (vs. bachelor's)										0.06	0.06	.27	0.06	0.06	.28
Student-level covariates															
Individualized education program (vs. non-individualized education program)										–0.06	0.05	.20	–0.06	0.05	.23
English language learner (vs. non-English language learner)										–0.01	0.05	.90	–0.01	0.05	.84
Pretest score										0.56	0.02	.00	0.56	0.02	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	0.49			0.49			0.49			0.35			0.35		
Level 2	0.02	127.23 (60)	.00	0.02	126.13 (59)	.00	0.02	60.89 (28)	.00	0.02	124.58 (56)	.00	0.02	124.44 (56)	.00
Level 3	0.03	97.76 (31)	.00	0.03	98.78 (31)	.00	0.03	111.82 (31)	.00	0.01	63.47 (31)	.00	0.01	56.86 (28)	.00
Thinking Reader achievement effect	—			—			0.01	33.49 (31)	.35	—			—		
N	2,217			2,217			2,217			2,201			2,201		

— Indicates that the random effect was not estimated for that model...

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Metacognitive Awareness of Reading Strategies Inventory survey administered by study team.

Table D3.4 Treatment as a Fixed Effect: Results for Motivation for Reading Questionnaire

	Model 0: Fully unconditional			Model 1: Adding treatment indicator as fixed effect			Model 2: Adding treatment indicator as a random effect			Model 3: Adding student and teacher covariates and treatment indicator as fixed effect			Model 4: Model 3 + school covariates		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	2.78	0.02	.00	2.78	0.02	.00	2.78	0.02	.00	2.76	0.02	.00	2.77	0.01	.00
School-level covariates															
More than 74% of enrollment eligible for free or reduced-price lunch													0.06	0.03	.05
Mid-sized: Enrollment 440–575 (vs. less than 440)													–0.02	0.04	.60
Large-sized: Enrollment greater than 575 (vs. less than 440)													–0.04	0.04	.22
Teacher-level covariates															
Thinking Reader				–0.02	0.03	.51	–0.04	0.03	.26	–0.00	0.02	.95	–0.00	0.02	.95
Years of experience 4–20 (vs. more than 20)										–0.03	0.04	.46	–0.03	0.04	.50
Years of experience less than 4 (vs. more than 20)										0.01	0.06	.87	0.01	0.06	.84
Master's degree or higher (vs. bachelor's)										0.03	0.03	.45	0.03	0.03	.47
Student-level covariates															
Individualized education program (vs. non-individualized education program)										–0.02	0.03	.55	–0.01	0.03	.62
English language learner (vs. non-English language learner)										–0.05	0.03	.10	–0.05	0.03	.08
Pretest score										0.67	0.02	.00	0.66	0.02	.00
Random effect	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**	**Variance**	**χ2 (df)**	***p* value**
Level 1	0.24			0.24			0.24			0.14			0.14		
Level 2	0.01	127.60 (60)	.00	0.01	128.50 (59)	.00	0.01	63.66 (28)	.00	0.01	109.56 (56)	.00	0.01	109.60 (56)	.00
Level 3	0.01	60.61 (31)	.00	0.01	60.24 (31)	.00	0.01	68.81 (31)	.00	0.00	51.58 (31)	.01	0.00	41.44 (28)	.05
Thinking Reader achievement effect	—			—			0.01	34.63 (31)	.30	—			—		
N	2,225			2,225			2,225			2,201			2,201		

— Indicates that the random effect was not estimated for that model.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; Motivation for Reading Questionnaire surveys administered by study team.

D4. Sensitivity Analysis: Two-Level Impact Model

The third sensitivity analysis examined how sensitive the findings were to a two-level model as opposed to a three-level model. A sequence of two-level models (Models 0–4) was conducted. For each sequence, Level 1 represents the between-student and within-teacher model and Level 2 represents the between-teacher model. The fully unconditional model (Model 0) is presented first. Following is an unadjusted model that includes only the treatment indicator (Model 1) and a model that adjusts the treatment indicator by student and teacher covariates (Model 2). In Models 3 and 4, the school effects are added as fixed effects by including 32–1 school dummy variables at the teacher level. Finally, Model 4 adds the interactions between the treatment indicator and the school dummy variables. These models are presented below.

Model 0 is the fully unconditional model.

Level 1 model

$Y_{ij} = \pi_{0j} + \varepsilon_{ij}$

Level 2 model

$\pi_{0j} = \beta_{00} + U_{0j}$

Model 1 adds the treatment indicator centered within each school. The subscript $k = 1, \ldots , 32$ corresponds to the 32 schools.

Level 1 model

$Y_{ij} = \pi_{0j} + \varepsilon_{ij}$

Level 2 model

$\pi_{0j} = \beta_{00} + \beta_{01}\,(\mathrm{Trt}_j - \mathrm{Trt}._k) + U_{0j}$

Model 2 adds student- and teacher-level predictors.

Level 1 model

$Y_{ij} = \pi_{0j} + \pi_{1j}\,(\mathrm{Pretest}_{ij} - \mathrm{Pretest}...) + \pi_{2j}\,(\mathrm{IEP}_{ij} - \mathrm{IEP}...) + \pi_{3j}\,(\mathrm{ELL}_{ij} - \mathrm{ELL}...) + \varepsilon_{ij}$

Level 2 model

$\pi_{0j} = \beta_{00} + \beta_{01}\,(\mathrm{Trt}_j - \mathrm{Trt}._k) + \beta_{02}\,(\mathrm{Master}_j - \mathrm{Master}._k) + \beta_{03}\,(\mathrm{Exp\_low}_j - \mathrm{Exp\_low}._k)$

$+ \beta_{04}\,(\mathrm{Exp\_Med}_j - \mathrm{Exp\_Med}._k) + \mathrm{U}_{0j}$

Model 3 includes the 32–1 school dummy variables. To generate the school dummy variables, effects coding was used.

Level 1 model

$Y_{ij} = \pi_{0j} + \pi_{1j}$ (Pretest$_{ij}$ – Pretest…) + π_{2j} (IEP$_{ij}$ – IEP...) + π_{3j} (ELL$_{ij}$ – ELL...) + ε_{ij}

Level 2 model

$$\pi_{0j} = \beta_{00} + \beta_{01}(Trt_j - Trt._k) + \beta_{02}(Master_j - master._k) + \beta_{03}(Exp\_low_j - Exp\_low._k) +$$
$$\beta_{04}(Exp\_Med_j - Exp\_Med._k) + \delta_k \sum_{k=1}^{31} school_k + U_{0j}$$

Finally, **Model 4** includes the interactions between the school dummy variables and the treatment indicator. These interactions aim to test whether the treatment effect varies from school to school.

Level 1 model

$Y_{ij} = \pi_{0j} + \pi_{1j}$ (Pretest$_{ij}$ – Pretest…) + π_{2j} (IEP$_{ij}$ – IEP...) + π_{3j} (ELL$_{ij}$ – ELL...) + ε_{ij}

Level 2 model

$$\pi_{0j} = \beta_{00} + \beta_{01}(Trt_j - Trt._k) + \beta_{02}(Master_j - master._k) + \beta_{03}(Exp\_low_j - Exp\_low._k) +$$
$$\beta_{04}(Exp\_Med_j - Exp\_Med._k) + \delta_k \sum_{k=1}^{31} school_k + \lambda_k \sum_{k=1}^{31} school_k \times Trt_j + U_{0j}$$

The treatment by school interactions is not identified in schools in which only two teachers were randomized (15 schools). In these cases, the interactions are confounded by the between-teacher error. In schools with more than two teachers (17 schools), these interactions can be identified but with low power.

Tables D4.1–D4.4 show the results of the four models. The direction and magnitude of the treatment effects and overall conclusions did not change using the two-level model compared with the three-level model presented in Chapter 4.

Table D4.1 Two-Level Impact Model: Results for Gates-MacGinitie Reading Tests Reading Vocabulary Subtest Based on Listwise Deletion of Missing Data

	Model 0: Fully unconditional			Model 1: Adding treatment indicator			Model 2: Adding student and teacher covariates			Model 3: Model 2 + school dummy variables[a]			Model 4: Model 3 + interaction between school dummy variables and treatment indicator[b]		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	516.46	1.85	.00	516.45	1.85	.00	516.46	2.07	.00	516.99	0.68	.00	516.94	0.64	.00
Teacher-level covariates															
Thinking Reader				–3.33	3.75	.38	–0.84	1.50	.58	–0.98	1.28	.45	–1.06	1.36	.45
Years of experience 4–20 (vs. more than 20)							0.70	1.86	.71	2.02	2.30	.38	3.20	2.94	.29
Years of experience less than 4 (vs. more than 20)							4.43	2.66	.10	5.75	3.35	.09	8.81	4.83	.08
Master's degree or higher (vs. bachelor's)							–0.48	1.65	.77	–0.04	1.84	.99	–0.25	2.47	.92
Student-level covariates															
Individualized education program (vs. non-individualized education program)							–7.44	1.69	.00	–7.96	1.70	.00	–7.91	1.71	.00
English language learner (vs. non-English language learner)							–4.16	1.57	.01	–3.46	1.62	.03	–3.54	1.64	.03
Pretest score							0.81	0.01	.00	0.80	0.01	.00	0.80	0.01	.00
Random effect	**Variance**	**χ2 (90)**	***p* value**	**Variance**	**χ2 (89)**	***p* value**	**Variance**	**χ2 (86)**	***p* value**	**Variance**	**χ2 (55)**	***p* value**	**Variance**	**χ2 (24)**	***p* value**
Level 1	990.51			990.51			369.26			368.66			369.01		
Level 2	257.94	590.65	.00	258.66	589.52	.00	29.18	232.86	.00	16.51	112.08	.00	11.34	41.30	.02
N	2,156			2,156			2,147			2,147			2,147		

[a] The 31 school dummy variables are omitted.

[b] The 31 school dummy variables and the interactions between those dummy variables and the treatment indicator are omitted.

Source: Teacher background questionnaire administered by study team; student rosters completed by study teachers; Gates-MacGinitie Reading Tests vocabulary subtest administered by study team.

Table D4.2 Two-Level Impact Model: Results for Gates-MacGinitie Reading Tests Reading Comprehension Subtest Based on Listwise Deletion of Missing Data

	Model 0: Fully unconditional			Model 1: Adding treatment indicator			Model 2: Adding student and teacher covariates			Model 3: Model 2 + school dummy variables[a]			Model 4: Model 3 + interaction between school dummy variables and treatment indicator[b]		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	506.46	1.72	.00	506.45	1.73	.00	509.61	2.61	.00	506.22	0.89	.00	506.22	0.98	.00
Teacher-level covariates															
Thinking Reader				–3.01	3.50	.39	0.46	1.89	.81	0.88	1.69	.61	0.95	2.11	.66
Years of experience 4–20 (vs. more than 20)							–2.53	2.34	.28	–1.47	3.01	.63	–0.52	4.65	.91
Years of experience less than 4 (vs. more than 20)							–1.69	3.36	.62	1.28	4.41	.77	3.28	7.88	.68
Master's degree or higher (vs. bachelor’s)							–1.24	2.09	.55	–2.69	2.40	.27	–0.41	3.83	.92
Student-level covariates															
Individualized education program (vs. non-individualized education program)							–7.16	1.99	.00	–7.87	2.01	.00	–7.92	2.03	.00
English language learner (vs. non-English language learner)							–1.67	1.83	.36	–1.68	1.89	.37	–1.81	1.90	.34
Pretest score							0.71	0.02	.00	0.70	0.02	.00	0.69	0.02	.00
Random effect	**Variance**	**χ2 (90)**	***p* value**	**Variance**	**χ2 (89)**	***p* value**	**Variance**	**χ2 (86)**	***p* value**	**Variance**	**χ2 (55)**	***p* value**	**Variance**	**χ2 (24)**	***p* value**
Level 1	947.39			947.37			505.23			504.96			504.38		
Level 2	219.52	531.91	.00	220.40	530.84	.00	50.43	266.55	.00	34.39	136.86	.00	46.35	70.54	.00
N	2,149			2,149			2,140			2,140			2,140		

[a] The 31 school dummy variables are omitted.
[b] The 31 school dummy variables and the interactions between those dummy variables and the treatment indicator are omitted.
Source: Teacher background questionnaire administered by study team; student rosters completed by study teachers; Gates-MacGinitie Reading Tests comprehension subtest administered by study team.

Table D4.3 Two-Level Impact Model: Results for Metacognitive Awareness of Reading Strategies Inventory Based on Listwise Deletion of Missing Data

	Model 0: Fully unconditional			**Model 1: Adding treatment indicator**			**Model 2: Adding student and teacher covariates**			**Model 3: Model 2 + school dummy variables**[a]			**Model 4: Model 3 + interaction between school dummy variables and treatment indicator**[b]		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.08	0.03	.00	3.08	0.03	.00	3.19	0.06	.00	3.10	0.02	.00	3.10	0.02	.00
Teacher-level covariates															
Thinking Reader				0.02	0.06	.75	0.02	0.046	.72	0.02	0.04	.71	0.02	0.05	.73
Years of experience 4–20 (vs. more than 20)							–0.12	0.057	.04	–0.10	0.07	.20	–0.07	0.10	.50
Years of experience less than 4 (vs. more than 20)							–0.06	0.082	.48	0.00	0.11	1.00	0.26	0.17	.14
Master's degree or higher (vs. bachelor's)							–0.05	0.051	.38	0.06	0.06	.32	0.09	0.09	.30
Student-level covariates															
Individualized education program (vs. non-individualized education program)							–0.07	0.046	.15	–0.06	0.047	.22	–0.07	0.05	.14
English language learner (vs. non-English language learner)							–0.00	0.047	.95	–0.01	0.048	.78	–0.02	0.05	.67
Pretest score							0.56	0.019	.00	0.55	0.020	.00	0.55	0.02	.00
Random effect	**Variance**	**χ2 (90)**	***p* value**	**Variance**	**χ2 (89)**	***p* value**	**Variance**	**χ2 (86)**	***p* value**	**Variance**	**χ2 (55)**	***p* value**	**Variance**	**χ2 (24)**	***p* value**
Level 1	0.49			0.49			0.36			0.36			0.36		
Level 2	0.05	339.09	.00	0.06	338.34	.00	0.03	260.68	.00	0.02	124.85	.00	0.02	53.40	.00
N	2,217			2,217			2,201			2,201			2,201		

[a] The 31 school dummy variables are omitted.

[b] The 31 school dummy variables and the interactions between those dummy variables and the treatment indicator are omitted.

Source: Teacher background questionnaire administered by study team; student rosters completed by study teachers; Metacognitive Awareness of Reading Strategies Inventory survey administered by study team.

Table D4.4 Two-Level Impact Model: Results for Motivation for Reading Questionnaire Based on Listwise Deletion of Missing Data

	Model 0: Fully unconditional			Model 1: Adding treatment indicator			Model 2: Adding student and teacher covariates			Model 3: Model 2 + school dummy variables[a]			Model 4: Model 3 + interaction between school dummy variables and treatment indicator[b]		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	2.77	0.02	.00	2.77	0.02	.00	2.82	0.04	.00	2.78	0.01	.00	2.77	0.01	.00
Teacher-level covariates															
Thinking Reader				–0.02	0.04	.56	0.00	0.03	1.00	0.00	0.03	.94	–0.01	0.03	.66
Years of experience 4–20 (vs. more than 20)							–0.04	0.03	.19	–0.04	0.04	.42	–0.07	0.06	.26
Years of experience less than 4 (vs. more than 20)							–0.04	0.05	.45	0.01	0.07	.94	0.05	0.10	.60
Master's degree or higher (vs. bachelor’s)							–0.04	0.03	.15	0.02	0.04	.55	0.01	0.05	.85
Student-level covariates															
Individualized education program (vs. non-individualized education program)							–0.02	0.03	.46	–0.01	0.03	.66	–0.02	0.03	.43
English language learner (vs. non-English language learner)							–0.05	0.03	.12	–0.06	0.03	.08	–0.05	0.03	.09
Pretest score							0.67	0.02	.00	0.66	0.02	.00	0.66	0.02	.00
Random effect	**Variance**	**χ2 (90)**	***p* value**	**Variance**	**χ2 (89)**	***p* value**	**Variance**	**χ2 (86)**	***p* value**	**Variance**	**χ2 (55)**	***p* value**	**Variance**	**χ2 (24)**	***p* value**
Level 1	0.24			0.24			0.14			0.14			0.14		
Level 2	0.02	250.15	.00	0.02	251.10	.00	0.01	199.02	.00	0.01	109.67	.00	0.01	47.15	.01
N	2,225			2,225			2,208			2,208			2,208		

[a] The 31 school dummy variables are omitted.
[b] The 31 school dummy variables and the interactions between those dummy variables and the treatment indicator are omitted.
Source: Teacher background questionnaire administered by study team; student rosters completed by study teachers; Motivation for Reading Questionnaire survey administered by study team.

D5. Sensitivity Analysis: Impact Results Based on Reduced Sample of One Classroom per Teacher

The fourth sensitivity analysis assessed whether results were sensitive to the number of classes per teacher (multiple classrooms vs. one classroom). Because of the mix of different school configurations in the sample, 62 teachers taught only one classroom and 30 teachers taught multiple classrooms (see Table D5.1).

Table D5.1 Number and Percentage of Teachers With One and Multiple Classes, by Study Condition

	Intervention		Control			
Teacher group	**Number**	**Percent**	**Number**	**Percent**	χ^2 (*df* = 1)	***p* value**
Teachers with multiple classes	16	32.7	14	32. 6		
Teachers with one class	33	67.4	29	67.4	0.00	.99

Note: Total classes equaled 129 (67 in the intervention group and 62 in the control group).

Source: Student rosters completed by study teachers.

Randomization should have balanced the number of classes per teacher. In other words, that some teachers had multiple classes should not have introduced any bias in the estimation of the treatment coefficient. Table D5.2 presents the average number of classes and students per teacher in the intervention and control conditions. The results of this table reveal that the average number of classes and students per teacher in the two conditions were almost identical.

Table D5.2 Comparing Full and Reduced Samples: Average Number of Classes and Students per Teacher, by Study Condition

Classes and students	Total	Intervention	Control
Average classes per teacher			
Full sample	1.4	1.4	1.4
Subsample	1.0	1.0	1.0
Average students per teacher			
Full sample	26.7	26.8	26.7
Subsample	18.5	18.8	18

Source: Student rosters completed by study teachers.

The impact models were first run using the entire sample (reported in Chapter 4). Though multilevel modeling, software can account for unequal cluster sizes, and teachers who taught multiple sections may have contributed disproportionately to the effects observed. As a sensitivity analysis, the impact models were run on a reduced sample, created by randomly selecting and including only a single classroom for the teachers who taught multiple sections (i.e., all teachers had one classroom). This reduced the student sample from 2,407 to 1,699. The characteristics of that subsample are presented in Table D5.3.

Table D5.3 Student Characteristics of the Reduced Sample, by Condition

	Intervention (n = 923)		Control (n = 776)			
Characteristic	**Number**	**Percent**	**Number**	**Percent**	χ^2 (*df*)[a]	***p* value**
Female	472	51.1	411	53.0	χ^2 (1) = 0.56	.45
Race/ethnicity[b]					χ^2 (4) = 4.33	.36
Black	125	14.5	107	14.6		.94
Asian	47	5. 5	52	7.1		.17
Hispanic	261	30.2	202	27.6		.25
White	292	33.8	267	36.5		.27
Other race/ethnicity	138	16.0	104	14.2		.32
Individualized education program	120	13.0	54	7.0	χ^2 (1) = 16.74	.00
English language learner	66	7.2	91	11.7	χ^2 (1) = 10.53	.00

[a] Numbers in parentheses are degrees of freedom for chi-squared tests. The calculation of the statistics does not account for the clustering of students by teacher or teacher by school.
[b] This variable had 104 missing cases (60 in the intervention group and 44 in the control group).
Source: Student self-report section on Gates-MacGinitie Reading Tests administered by study team; student rosters completed by study teachers.

Results of the impact models with the smaller sample of students are presented in Tables D5.4 and D5.5. The direction and magnitude of the treatment effects and overall conclusions did not change between the full and reduced samples.

Table D5.4 Impact Results on Reading Achievement Based on Reduced Sample of One Classroom per Teacher and Listwise Deletion of Missing Data

	GMRT: Vocabulary subtest			GMRT: Comprehension subtest		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	516.34	0.67	.00	506.38	1.05	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–4.74	1.41	.00	–5.35	2.07	.02
Mid-sized: Enrollment 440–575 (vs. less than 440)	–4.47	1.70	.01	–2.86	2.57	.28
Large-sized: Enrollment greater than 575 (vs. less 440)	–6.19	1.68	.00	–4.38	2.52	.09
Teacher-level covariates						
Thinking Reader	–0.75	1.37	.59	0.55	2.04	.79
Years of experience 4–20 (vs. more than 20)	0.96	2.39	.69	–0.65	3.29	.84
Years of experience less than 4 (vs. more than 20)	3.57	3.49	.31	4.64	4.86	.34
Master's degree or higher (vs. bachelor’s)	–0.20	1.91	.92	–0.69	2.62	.79
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–8.83	1.98	.00	–6.32	2.34	.01
English language learner (vs. non-English language learner)	–3.69	1.85	.05	0.35	2.18	.87
Pretest score	0.80	0.02	.00	0.70	0.02	.00
Random effect	**Variance**	**χ^2 (*df*)**	***p* value**	**Variance**	**χ^2(*df*)**	***p* value**
Level 1	363.25			484.95		
Level 2	15.09	42.95 (24)	.01	35.25	66.41 (24)	.00
Level 3	0.29	33.80 (28)	.21	9.77	43.16 (28)	.03
Thinking Reader achievement effect	0.79	32.45 (31)	.40	26.80	37.40 (31)	.20
N	1,511			1,504		

Note: GMRT is Gates-MacGinitie Reading Tests.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table D5.5 Impact Results on Reading Strategies and Reading Motivation Based on Reduced Sample of One Classroom per Teacher and Listwise Deletion of Missing Data

	Reading strategies: MARSI			Reading motivation: MRQ		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.11	0.03	.00	2.78	0.02	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	0.04	0.05	.49	0.04	0.03	.20
Mid-sized: Enrollment 440–575 (vs. less than 440)	0.00	0.07	.98	–0.02	0.04	.64
Large-sized: Enrollment greater than 575 (vs. less 440)	0.00	0.07	.97	–0.02	0.04	.67
Teacher-level covariates						
Thinking Reader	–0.01	0.04	.74	–0.03	0.03	.29
Years of experience 4–20 (vs. more than 20)	–0.10	0.07	.17	–0.04	0.04	.34
Years of experience less than 4 (vs. more than 20)	–0.03	0.10	.78	0.00	0.06	1.00
Master's degree or higher (vs. bachelor’s)	0.06	0.06	.27	0.04	0.03	.27
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–0.05	0.06	.33	–0.04	0.03	.21
English language learner (vs. non-English language learner)	0.04	0.06	.47	–0.04	0.04	.29
Pretest score	0.56	0.02	.00	0.67	0.02	.00
Random effect	**Variance**	**χ^2 (*df*)**	***p* value**	**Variance**	**χ^2 (*df*)**	***p* value**
Level 1	0.35			0.14		
Level 2	0.01	48.03 (25)	.00	0.00	47.14 (25)	.01
Level 3	0.01	67.64 (28)	.00	0.00	56.98 (28)	.00
Thinking Reader achievement effect	0.01	34.15 (31)	.32	0.01	37.11 (31)	.21
N	1,548			1,553		

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; MARSI and MRQ surveys administered by study team.

D6. Sensitivity Analysis: Impact Results Based on Reduced Sample That Excludes Students Whose Teachers Left the Sample

The fifth sensitivity analysis examined whether the results were sensitive to keeping in the sample those students whose original teachers left their school at the beginning of the school year. The teacher sample was reduced by less than four teachers. In all cases, students were distributed into other classrooms. Following an intent-to-treat approach, their movement was tracked, and their original study group assignments in the impact analyses, which are presented in Chapter 4, were preserved. To verify that the impact estimates were unaffected by those classroom reassignments, the same impact models were run on a reduced sample that excluded the 20 students who were in these teachers' classrooms. As shown in Tables D6.1 and D6.2, the impact estimates were nearly identical to the original results.

Table D6.1 Impact Results on Reading Achievement Excluding Students Whose Teachers Left the Sample Based on Listwise Deletion of Missing Data

	GMRT: Vocabulary subtest			GMRT: Comprehension subtest		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	517.14	0.66	.00	506.62	0.93	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–5.12	1.31	.00	–5.26	1.83	.01
Mid-sized: Enrollment 440–575 (vs. less than 440)	–4.35	1.63	.01	–3.13	2.29	.18
Large-sized: Enrollment greater than 575 (vs. less 440)	–5.04	1.58	.00	–4.52	2.23	.05
Teacher-level covariates						
Thinking Reader	–1.24	1.33	.36	0.83	1.74	.64
Years of experience 4–20 (vs. more than 20)	2.58	2.12	.23	–1.71	2.83	.55
Years of experience less than 4 (vs. more than 20)	6.29	3.12	.05	1.61	4.16	.70
Master's degree or higher (vs. bachelor's)	0.46	1.71	.79	–2.10	2.26	.36
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–8.23	1.66	.00	–7.83	1.98	.00
English language learner (vs. non-English language learner)	–3.80	1.56	.02	–1.10	1.84	.55
Pretest score	0.80	0.01	.00	0.71	0.02	.00
Random effect	**Variance**	**χ^2 (*df*)**	***p* value**	**Variance**	**χ^2 (*df*)**	***p* value**
Level 1	367.76			505.17		
Level 2	9.32	42.37 (22)	.01	23.97	63.60 (22)	.00
Level 3	2.74	41.12 (28)	.05	7.73	44.00 (28)	.03
Thinking Reader achievement effect	12.43	41.56 (31)	.10	18.12	37.13 (31)	.21
N	2,131			2,124		

Note: GMRT is Gates-MacGinitie Reading Tests.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table D6.2 Impact Results on Reading Strategies and Reading Motivation Excluding Students Whose Teachers Left the Sample Based on Listwise Deletion of Missing Data

	Reading strategies: MARSI			**Reading motivation: MRQ**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	3.09	0.03	.00	2.77	0.01	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	0.07	0.05	.19	0.05	0.03	.07
Mid-sized: Enrollment 440–575 (vs. less than 440)	0.00	0.06	.96	–0.01	0.03	.88
Large-sized: Enrollment greater than 575 (vs. less 440)	0.00	0.06	.99	–0.02	0.03	.59
Teacher-level covariates						
Thinking Reader	0.00	0.04	.93	–0.01	0.03	.60
Years of experience 4–20 (vs. more than 20)	–0.08	0.07	.27	–0.02	0.04	.61
Years of experience less than 4 (vs. more than 20)	0.03	0.10	.78	0.03	0.06	.65
Master's degree or higher (vs. bachelor's)	0.08	0.06	.16	0.03	0.03	.36
Student-level covariates						
Individualized education program (vs. non–individualized education program)	–0.06	0.05	.19	–0.02	0.03	.45
English language learner (vs. non–English language learner)	–0.03	0.05	.59	–0.05	0.03	.07
Pretest score	0.56	0.02	.00	0.66	0.02	.00
Random effect	**Variance**	**χ^2 (*df*)**	***p* value**	**Variance**	**χ^2 (*df*)**	***p* value**
Level 1	0.35			0.14		
Level 2	0.01	58.37 (23)	.00	0.00	50.49 (23)	.00
Level 3	0.01	62.11 (28)	.00	0.00	51.63 (28)	.00
Thinking Reader achievement effect	0.01	34.59 (31)	.30	0.01	38.20 (31)	.18
N	2,184			2,191		

Note: MARSI is Metacognitive Awareness of Reading Strategies Inventory; MRQ is Motivation for Reading Questionnaire.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; MARSI and MRQ surveys administered by study team.

Appendix E. Exploratory Analyses

This appendix contains information on the exploratory analyses, including the power analysis (Appendix E1), the impact model for the exploratory research questions relating to baseline achievement (Appendix E2), and the impact model for the exploratory research questions relating to baseline motivation (Appendix E3). Appendices E4 and E5 present two sensitivity analyses to explore the robustness of the results to changes in the size and composition of the tertiles.

E1. Power for Exploratory Analysis

To calculate the power for these subgroup analyses, we adjusted the assumptions made for the power calculations of the main study as follows:

- *Statistical significance:* $p < .05$, two-tailed test. For the exploratory analyses, p values were not adjusted for multiple comparison.
- *Teacher and student sample size:* Using baseline counts, each school averages 2.8 teachers. For the power calculations, an average of three teachers was assumed per school. Based on pretest information, each classroom averaged nine students per baseline achievement tertile.[48] For the power calculations, we assumed an average of 8 students per classroom and tertile (accounting for 11% of missing data at posttest).

With these assumptions the minimum detectable effect size (MDES) estimates for within-tertile treatment effects are 0.22, 0.19, and 0.17 for R-squared values of .50, .60, and .70, respectively.

E2. Three-Level Model for Exploratory Research Questions 1 and 2

Level 1 (between-students, within-teacher and school model)

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{Pretest\_T1}_{ijk} - \text{Pretest\_T1...}) + \pi_{2jk}(\text{Pretest\_T2}_{ijk} - \text{Pretest\_T2...}) + \pi_{3jk}(\text{IEP}_{ijk} -$

$\text{IEP..}) + \pi_{4jk}(\text{ELL}_{ijk} - \text{ELL…}) + \varepsilon_{ijk}$, where $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Level 2 (between-teacher, within-school model)

$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{Trt}_{jk} - \text{Trt.}_{k}) + \beta_{02k}(\text{Master}_{jk} - \text{Master.}_{k}) + \beta_{03k}(\text{Exp\_low}_{jk} - \text{Exp\_low.}_{k})$

$+ \beta_{04k}(\text{Exp\_Med}_{jk} - \text{Exp\_Med.}_{k}) + U_{0jk}$, where $U_{0jk} \sim N(0, \tau_{\pi 00})$

$\pi_{1jk} = \beta_{10k} + \beta_{11k}(\text{Trt}_{jk} - \text{Trt.}_{k})$

$\pi_{2jk} = \beta_{20k} + \beta_{21k}(\text{Trt}_{jk} - \text{Trt.}_{k})$

$\pi_{3jk} = \beta_{30k}$

[48] Calculations were made using a sample of 90 teachers.

$\pi_{4jk} = \beta_{40k}$

Level 3 (between-school model)

$\beta_{00k} = \gamma_{000} + \gamma_{001}$ (Hlunch$_k$ – Hlunch.) + γ_{002} (Midsize$_k$ – Midsize.)

+ γ_{003} (Lsize$_k$ – Lsize.) + V_{00k}

$\beta_{01k} = \gamma_{010}$

$\beta_{02k} = \gamma_{020}$

$\beta_{03k} = \gamma_{030}$

$\beta_{04k} = \gamma_{040}$

$\beta_{10k} = \gamma_{100}$

$\beta_{11k} = \gamma_{110}$

$\beta_{20k} = \gamma_{200}$

$\beta_{21k} = \gamma_{210}$

$\beta_{30k} = \gamma_{300}$

$\beta_{40k} = \gamma_{400}$

Y_{ijk} represents the posttest outcome score (reading vocabulary or reading comprehension) for student *i* in teacher *j*'s class in school *k.*

Similar to the main impact model, in the Level 1 equation, the student outcome score is modeled as a function of the English language learner (ELL) status and the special education status (e.g., individualized education program [IEP]). Additionally, and instead of the continuous form of the pretest score, two tertile indicators "Pretest_T1" and "Pretest_T2" are included in the Level 1 equation. The tertile indicators take the value of 1 for a student whose pretest scores are in the lowest and middle tertiles, and 0 otherwise.

In the Level 2 equation, the adjusted mean outcome score π_{0jk} for teacher *j* in school *k* is modeled as varying randomly across teachers and as a function of the treatment indicator (Trt) and teacher characteristics, such as highest level of education (Master's degree) and years of teaching experience dummies (Exp_low and Exp_Med).

In the Level 3 equation, the average outcome in each school is modeled as a function of the same covariates used in the main impact model: a dummy covariate that captures whether the percentage of students who are eligible for free or reduced-price lunch is higher than 74% (Hlunch) and enrollment dummies (Msize and Lsize*).*

The parameters of interest are γ_{110} and γ_{210}. These parameters capture the cross-level interactions between the treatment status and the pretest tertile indicators, after holding constant the other covariates of the model. The differential effect of the treatment on students in the lowest tertile compared with students in the highest tertile is captured by γ_{110}. Similarly, γ_{210} is the differential effect of the treatment on students in the middle tertile compared with students in the highest tertile.

Appendix C includes a detailed description of the coding of the covariates included in the multilevel models.

E3. Three-Level Model for Exploratory Research Questions 3 and 4

Level 1 (between-students, within-teacher and school model)

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}$ (MRQ_T1$_{ijk}$ – MRQ_T1…) + π_{2jk} (MRQ_T2$_{ijk}$ – MRQ_T2...) + π_{3jk} (Pretest$_{ijk}$ – Pretest...) + π_{4jk} (IEP$_{ijk}$ – IEP...) + π_{5jk} (ELL$_{ijk}$ – ELL...) + ε_{ijk}, where $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Level 2 (between-teacher, within-school model)

$\pi_{0jk} = \beta_{00k} + \beta_{01k}$ (Trt$_{jk}$ – Trt.$_{k}$) + β_{02k} (Master$_{jk}$ – Master.$_{k}$) + β_{03k} (Exp_low$_{jk}$ – Exp_low.$_{k}$)

+ β_{04k} (Exp_Med$_{jk}$ – Exp_Med.$_{k}$) + U_{0jk}, where $U_{0jk} \sim N(0, \tau_{\pi 00})$

$\pi_{1jk} = \beta_{10k} + \beta_{11k}$ (Trt$_{jk}$ – Trt.$_{k}$)

$\pi_{2jk} = \beta_{20k} + \beta_{21k}$ (Trt$_{jk}$ – Trt.$_{k}$)

$\pi_{3jk} = \beta_{30k}$

$\pi_{4jk} = \beta_{40k}$

$\pi_{5jk} = \beta_{50k}$

Level 3 (between-school model)

$\beta_{00k} = \gamma_{000} + \gamma_{001}$ (Hlunch$_{k}$ – Hlunch.) + γ_{002} (Midsize$_{k}$ – Midsize.)

+ γ_{003} (Lsize$_{k}$ – Lsize.) + V_{00k}

$\beta_{01k} = \gamma_{010}$

$\beta_{02k} = \gamma_{020}$

$\beta_{03k} = \gamma_{030}$

$\beta_{04k} = \gamma_{040}$

$\beta_{10k} = \gamma_{100}$

$\beta_{11k} = \gamma_{110}$

$\beta_{20k} = \gamma_{200}$

$\beta_{21k} = \gamma_{210}$

$\beta_{30k} = \gamma_{300}$

$\beta_{40k} = \gamma_{400}$

$\beta_{50k} = \gamma_{500}$

In addition to the pretest score, and the indicators that capture whether the student is classified as IEP or ELL, the Level 1 equation also includes two dummy indicators—"MRQ_T1" and "MRQ_T2"—which take the value of 1 for students in the lowest and middle tertiles of the baseline motivation to read measure, respectively, and 0 otherwise.

The parameters of interest are γ_{110} and γ_{210}. These parameters measure the cross-level interaction between the treatment indicator and the Motivation for Reading Questionnaire (MRQ) lowest and middle tertile indicators, after holding constant the other covariates of the model. The differential effect of the treatment on students in the lowest tertile compared with students in the highest tertile is captured by γ_{110}. Similarly, γ_{210} captures the differential effect of the treatment on students in the middle tertile compared with students in the highest tertile.

Table E3.1 Exploratory Impact Results on Reading Achievement (Research Questions 1 and 2)

	GMRT: Vocabulary subtest			**GMRT: Comprehension subtest**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	516.66	0.74	.00	506.45	0.99	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–7.51	1.52	.00	–6.91	2.01	.00
Mid-sized: Enrollment 440–575 (vs. less than 440)	–4.55	1.89	.02	–4.16	2.52	.11
Large-sized: Enrollment greater than 575 (vs. less than 440)	–6.46	1.83	.00	–6.02	2.46	.02
Teacher-level covariates						
Thinking Reader	–0.61	1.36	.65	–0.54	1.76	.76
Years of experience 4–20 (vs. more than 20)	2.37	2.44	.34	–0.42	3.14	.90
Years of experience less than 4 (vs. more than 20)	5.62	3.55	.12	1.21	4.59	.79
Master's degree or higher (vs. bachelor's)	–0.47	1.95	.81	–1.35	2.50	.59
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–11.18	1.96	.00	–11.59	2.11	.00
English language learner (vs. non-English language learner)	–5.96	1.81	.00	–2.39	1.94	.22
Tertile 1: Lowest achievement tertile (vs. highest tertile)	–57.86	1.36	.00	–49.30	1.37	.00
Tertile 1 *Thinking Reader*	3.19	2.57	.21	2.69	2.69	.32
Tertile 2 : Middle achievement tertile (vs. highest tertile)	–31.74	1.21	.00	–28.01	1.32	.00
Tertile 2 *Thinking Reader*	–0.32	2.42	.90	–3.07	2.65	.25
Tertile 1: Lowest achievement tertile (vs. middle tertile)[a]	–26.13	1.27	.00	–21.29	1.34	.00
Tertile 1 *Thinking Reader*[a]	3.51	2.51	.16	5.77	2.69	.03
Random effect	**Variance**	**χ^2(*df*)**	***p* value**	**Variance**	**χ^2(*df*)**	***p* value**
Level 1	507.57			568.04		
Level 2	14.59	95.79(55)	.00	36.52	141.42(55)	.00
Level 3	1.94	38.70(28)	.09	5.76	41.60(28)	.05
N	2,147			2,140		

Note: GMRT is Gates-MacGinitie Reading Tests.
[a] The coefficient, standard error and *p* value for this contrast were obtained by changing the reference group from Tertile 3 to Tertile 2 in the multilevel model.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table E3.2 Exploratory Impact Results on Reading Achievement (Research Questions 3 and 4)

	GMRT: Vocabulary subtest			**GMRT: Comprehension subtest**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	517.02	0.64	.00	506.64	0.85	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–5.16	1.31	.00	–5.45	1.72	.00
Mid-sized: enrollment 440–575 (vs. less than 440)	–4.80	1.62	.01	–3.05	2.14	.17
Large-sized: enrollment greater than 575 (vs. less than 440)	–6.47	1.57	.00	–5.08	2.07	.02
Teacher-level covariates						
Thinking Reader	–1.12	1.23	.37	0.79	1.65	.63
Years of experience 4–20 (vs. more than 20)	2.16	2.20	.33	–1.39	2.92	.63
Years of experience less than 4 (vs. more than 20)	6.40	3.21	.05	2.10	4.28	.63
Master's degree or higher (vs. bachelor's)	0.27	1.76	.88	–2.15	2.33	.36
Student-level covariates						
Individualized education program (vs. non–individualized education program)	–8.02	1.69	.00	–7.32	2.00	.00
English language learner (vs. non–English language learner)	–3.78	1.55	.02	–0.83	1.82	.65
Pretest	0.80	0.01	.00	0.70	0.02	.00
Motivation to read Tertile 1: lowest tertile (vs. highest tertile)	–2.24	1.06	.04	–2.91	1.25	.02
Motivation to read Tertile 1**Thinking Reader*	–1.15	2.13	.59	–2.37	2.50	.35
Motivation to read Tertile 2: middle tertile (vs. highest tertile)	–1.22	1.03	.24	–0.09	1.21	.94
Motivation to read Tertile 2**Thinking Reader*	0.43	2.06	.84	–2.04	2.42	.40
Motivation to read Tertile 1: lowest tertile (vs. highest tertile)[a]	–1.02	1.03	.33	–2.82	1.21	.02
Motivation to read Tertile 1**Thinking Reader*[a]	–1.58	2.09	.45	–0.33	2.45	.89
Random effect	**Variance**	**χ2(df)**	***p* value**	**Variance**	**χ2(df)**	***p* value**
Level 1	367.33			500.21		
Level 2	13.94	111.97(55)	.00	31.29	136.48(55)	.00
Level 3	0.42	33.70(28)	.21	0.75	35.33(28)	.16
N	2,135			2,129		

Note: GMRT is Gates-MacGinitie Reading Tests.

[a] The coefficient, standard error and *p* value for this contrast were obtained by changing the reference group from Tertile 3 to Tertile 2 in the multilevel model.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

E4. Sensitivity Analysis 1: Exploratory Impact Results Using Semi-Open Intervals to Define the Tertiles

The first sensitivity analysis defined the baseline achievement and motivation to read tertiles as semi-open intervals. More precisely, in the analyses, Tertiles 1 and 2 no longer include the upper bound values equivalent to the cumulative percentages of 33% and 66%, respectively. Tertile 1 includes all the pretest scores that fall below the cumulative percentage of 33%, and Tertile 2 includes all the scores at or above the 33% but below the 66% cumulative percentages. The exploratory impact analyses were re-estimated using these new cut points. The following tables present the new distributions of the tertiles for baseline reading vocabulary (Table E4.1), reading comprehension (Table E4.1), and motivation to read (Table E4.2). Tables E4.3 and E4.4 present results of the multilevel models for the tertiles based on semi-open intervals.

Table E4.1 Sensitivity Analysis 1: Reading Achievement Pretest Measure, by Baseline Achievement Tertile

Baseline achievement tertiles	N	Percent	Mean	Standard deviation	Minimum	Maximum
Reading Vocabulary						
Tertile 1	714	29.90	465.13	17.69	367	483
Tertile 2	820	34.34	498.40	7.86	486	510
Tertile 3	854	35.76	536.73	22.55	514	653
Reading Comprehension						
Tertile 1	784	32.83	466.16	15.81	396	487
Tertile 2	728	30.49	498.47	6.86	488	509
Tertile 3	876	36.68	534.44	19.76	512	652

Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table E4.2 Sensitivity Analysis 1: Motivation to Read Pretest Scores, by Baseline Reading Motivation Tertile

Baseline motivation tertiles	N	Percent	Mean	Standard deviation	Minimum	Maximum
Tertile 1	762	31.92	2.29	0.31	1.19	2.67
Tertile 2	816	34.19	2.87	0.12	2.67	3.06
Tertile 3	809	33.89	3.32	0.19	3.08	3.96

Source: Motivation for Reading Questionnaire survey administered by study team.

Table E4.3 Sensitivity Analysis 1: Exploratory Impact Results on Reading Achievement (Research Questions 1 and 2)

	GMRT: Vocabulary subtest			GMRT: Comprehension subtest		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	516.58	0.79	.00	506.48	1.02	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–7.75	1.62	.00	–6.93	2.07	.00
Mid-sized: Enrollment 440–575 (vs. less than 440)	–4.47	2.01	.03	–4.29	2.59	.11
Large-sized: Enrollment greater than 575 (vs. less than 440)	–6.48	1.95	.00	–6.03	2.53	.02
Teacher-level covariates						
Thinking Reader	–0.50	1.42	.72	–0.42	1.81	.82
Years of experience 4–20 (vs. more than 20)	1.87	2.54	.46	–0.16	3.22	.96
Years of experience less than 4 (vs. more than 20)	4.98	3.70	.18	1.56	4.72	.74
Master's degree or higher (vs. bachelor's)	0.01	2.03	1.00	–1.89	2.57	.46
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–11.76	2.00	.00	–10.96	2.14	.00
English language learner (vs. non-English language learner)	–6.32	1.85	.00	–2.80	1.97	.16
Tertile 1: Lowest achievement tertile (vs. highest tertile)	–56.41	1.40	.00	–48.16	1.40	.00
Tertile 1 **Thinking Reader*	3.71	2.65	.16	3.01	2.75	.27
Tertile 2: Middle achievement tertile (vs. highest tertile)	–31.68	1.21	.00	–28.56	1.32	.00
Tertile 2 **Thinking Reader*	–0.70	2.41	.77	–4.07	2.64	.12
Tertile 1: Lowest achievement tertile (vs. middle tertile)[a]	–24.73	1.33	.00	–19.60	1.38	.00
Tertile 1 **Thinking Reader*[a]	4.41	2.65	.10	7.09	2.80	.01
Random effect	**Variance**	**χ2(df)**	***p* value**	**Variance**	**χ2(df)**	***p* value**
Level 1	524.53			581.53		
Level 2	17.11	102.40 (55)	.00	39.35	145.02 (55)	.00
Level 3	2.87	39.67(28)	.07	6.24	41.23(28)	.05
N	2,147			2,140		

Note: GMRT is Gates-MacGinitie Reading Tests.

[a] The coefficient, standard error and *p* value for this contrast were obtained by changing the reference group from Tertile 3 to Tertile 2 in the multilevel model.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table E4.4 Sensitivity Analysis 1: Exploratory Impact Results on Reading Achievement (Research Questions 3 and 4)

	GMRT: Vocabulary subtest			**GMRT: Comprehension subtest**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	517.01	0.64	.00	506.62	0.84	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–5.20	1.31	.00	–5.54	1.72	.00
Mid-sized: Enrollment 440–575 (vs. less than 440)	–4.79	1.61	.01	–3.02	2.14	.17
Large-sized: Enrollment greater than 575 (vs. less than 440)	–6.47	1.56	.00	–5.09	2.07	.02
Teacher-level covariates						
Thinking Reader	–1.12	1.23	.36	0.80	1.64	.63
Years of experience 4–20 (vs. more than 20)	2.19	2.19	.32	–1.31	2.91	.65
Years of experience less than 4 (vs. more than 20)	6.44	3.20	.05	2.20	4.26	.61
Master's degree or higher (vs. bachelor's)	0.29	1.75	.87	–2.13	2.32	.36
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–7.93	1.69	.00	–7.28	2.00	.00
English language learner (vs. non-English language learner)	–3.78	1.55	.02	–0.84	1.82	.65
Pretest	0.80	0.01	.00	0.70	0.02	.00
Motivation to read Tertile 1: Lowest tertile (vs. highest tertile)	–2.58	1.07	.02	–3.57	1.26	.01
Motivation to read Tertile 1 **Thinking Reader*	–1.10	2.14	.61	–1.94	2.51	.44
Motivation to read Tertile 2: Middle tertile (vs. highest tertile)	–0.90	1.02	.37	–0.17	1.19	.89
Motivation to read Tertile 2 **Thinking Reader*	1.01	2.04	.62	–1.76	2.39	.46
Motivation to read Tertile 1: Lowest tertile (vs. middle tertile)[a]	–1.68	1.04	.11	–3.41	1.22	.01
Motivation to read Tertile 1 **Thinking Reader*a	–2.11	2.11	.32	–0.18	2.48	.94
Random effect	**Variance**	**χ2(df)**	***p* value**	**Variance**	**χ2(df)**	***p* value**
Level 1	367.12			499.63		
Level 2	13.67	111.03(55)	.00	30.9	135.77(55)	.00
Level 3	0.4	33.58(28)	.22	0.88	35.52(28)	.16
N	2,135			2,129		

Note: GMRT is Gates-MacGinitie Reading Tests.
[a] The coefficient, standard error and *p* value for this contrast were obtained by changing the reference group from Tertile 3 to Tertile 2 in the multilevel model.
Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

E5. Sensitivity Analysis 2: Exploratory Impact Results Using the Gates-MacGinitie Reading Tests Norming Distribution of Extended Scale Scores

The second sensitivity analysis defined the tertiles using the norming distribution of the extended scale scores on the Gates-MacGinitie Reading Tests (GMRT) as external benchmarks.[49] To derive the cut points, we used information presented in Tables E5.1 and E5.2. Table E5.1 shows the correspondence between z-scores, normal curve equivalent (NCE) scores, and cumulative percentages in a normal distribution. Table E5.2 shows the equivalent NCE scores for the different extended scale scores of the Grade 6 GMRT.[50] From data in Table E5.1, we approximated the NCE scores that corresponded to the 33 and 66 cumulative percentages (the first and second tertiles), which are 40 and 60, respectively. From Table E5.2, we identified the GMRT extended scale scores that corresponded to the NCE scores of 40 and 60. These scale scores are 496 and 531 for reading vocabulary and 498 and 534 for reading comprehension.

We used these new cut points to define the tertiles and rerun Exploratory Research Questions 1 and 2.[51] Table E5.3 presents the distribution of the tertiles for the second sensitivity analysis. Per Table E5.3, using the external benchmark to define the tertiles' cut points generated subgroups with unbalanced samples. The first subgroup, which includes students with the lowest baseline achievement scores, represents 47% and 49% of the study sample of students with baseline reading vocabulary and comprehension scores, respectively. The proportion of students in the second tertile is 37%, and in the third tertile, 16% and 14% for reading vocabulary and comprehension, respectively. These numbers indicate that the study sample has a larger proportion of students with lower scores relative to the sample used by the GMRT to develop their scores.

Table E5.4 presents results for Exploratory Research Questions 1 and 2 based on the new tertiles. Finally, all of the different cut points used to define the subgroups of students are illustrated in Tables E5.5 and E.5.6.

[49] The fourth edition of the GMRT used a sample of 37,000 students for evaluation of all test questions. The field testing was carried out in fall 1997, and the schools selected for participation represented all regions of the country, large and small school districts, and public and non-public schools. For more details about the characteristics of the sample used by this test, see MacGinitie, MacGinitie, Maria and Dreyer (2002).

[50] This table was extracted from page 54 of MacGinitie, MacGinitie, Maria, Dreyer, and Hughes (2007).

[51] For reading vocabulary, the first subgroup includes students with pretest scores below or equal to 496, the second group contains students with scores above 496 but less than or equal to 531, and the third group includes students with pretest scores above 531. For reading comprehension, the first subgroup includes students with pretest scores less than or equal to 498, the second group includes students with scores above 498 but less than or equal to 534, and the third group includes students with pretest scores above 534.

Table E5.1 Z-Scores, Cumulative Percentages and Normal Curve Equivalent Scores

Z-score	Cumulative percent	Normal curve equivalents
–2.00	2%	7.9
–1.00	16%	28.9
–0.50	30.9%	39.5
–0.45	32.6%	40.5
–0.43	33.4%	40.9
0.00	50.0%	50.0
0.43	66.6%	59.1
0.45	67.4%	59.5
0.50	69.1%	60.5
1.00	84%	71.1
2.00	98%	92.1
–2.00	2%	7.9

Note: NCE is defined as (approximately) 50 + 21.06z where "z" is the standard score.
Source: Crocker and Algina (1986).

Table E5.2 Gates-MacGinitie Reading Tests Scores Form S: Normal Curve Equivalent and Extended Scale Scores

GMRT: Reading vocabulary		GMRT: Reading comprehension	
Normal curve equivalents	**Extended scale scores**	**Normal curve equivalents**	**Extended scale scores**
1	354	1	353
1	373	1	364
1	389	1	375
1	401	1	386
1	412	1	396
1	423	1	406
1	433	1	413
7	440	1	420
10	447	1	426
13	453	1	432
17	458	1	437
19	463	7	442
23	468	10	447
25	472	13	451
28	476	15	456
30	479	19	460
32	483	20	464
34	486	22	467
37	490	24	471
39	493	26	475
40	496	28	478
42	500	30	481
44	503	33	485
47	507	34	488
48	510	36	491
51	514	38	494
52	517	41	498
54	520	42	500
56	524	44	503
58	527	45	506
60	531	47	509
61	534	48	512
64	538	51	516
66	542	52	519
68	546	54	522

GMRT: Reading vocabulary		GMRT: Reading comprehension	
Normal curve equivalents	Extended scale scores	Normal curve equivalents	Extended scale scores
70	550	56	526
72	554	58	530
75	559	60	534
77	564	68	538
80	569	65	543
83	575	68	547
87	582	70	553
90	590	74	559
93	600	78	566
99	619	83	574
99	653	87	583
		93	594
		99	617
		99	652

Note: GMRT is Gates-MacGinitie Reading Tests. The boxes outline the NCE scores that correspond to the 33 and 66 cumulative percentages.
Source: Gates-MacGinitie Reading Tests Manual for Scoring and Interpretation

Table E5.3 Sensitivity Analysis 2: Reading Achievement Pretest Scores, by Baseline Achievement Subgroups Generated on the Basis of Gates-MacGinitie Reading Tests Norming Distribution of Extended Scale Scores

Baseline achievement tertiles	N	Percent	Mean	Standard deviation	Minimum	Maximum
Reading Vocabulary						
Subgroup 1	1,121	46.94	474.69	19.10	367	496
Subgroup 2	876	36.68	513.65	9.38	500	531
Subgroup 3	391	16.37	555.17	21.07	534	653
Reading Comprehension						
Subgroup 1	1,161	48.62	474.82	18.14	396	498
Subgroup 2	891	37.31	515.09	10.33	500	534
Subgroup 3	336	14.07	554.50	16.90	538	652

Source: Gates-MacGinitie Reading Tests vocabulary and comprehension subtests administered by study team.

Table E5.4 Sensitivity Analysis 2: Exploratory Impact Results on Reading Achievement (Research Questions 1 and 2)

	GMRT: Vocabulary subtest			**GMRT: Comprehension subtest**		
Fixed effect	**Coefficient**	**Standard error**	***p* value**	**Coefficient**	**Standard error**	***p* value**
Intercept	516.74	0.77	.00	506.49	1.07	.00
School-level covariates						
More than 74% of enrollment eligible for free or reduced-price lunch	–8.18	1.57	.00	–7.73	2.16	.00
Mid-sized: Enrollment 440–575 (vs. less than 440)	–2.95	1.96	.14	–3.97	2.71	.16
Large-sized: Enrollment greater than 575 (vs. less than 440)	–6.28	1.91	.00	–5.59	2.65	.04
Teacher-level covariates						
Thinking Reader	–1.61	1.36	.24	0.45	1.87	.81
Years of experience 4–20 (vs. more than 20)	0.54	2.43	.82	–0.41	3.31	.90
Years of experience less than 4 (vs. more than 20)	2.86	3.54	.42	1.97	4.85	.69
Master's degree or higher (vs. bachelor’s)	–0.81	1.94	.68	–2.41	2.64	.37
Student-level covariates						
Individualized education program (vs. non-individualized education program)	–13.87	1.90	.00	–12.39	2.14	.00
English language learner (vs. non-English language learner)	–9.52	1.76	.00	–4.30	1.97	.03
Tertile 1: Lowest achievement tertile (vs. highest tertile)	–63.93	1.44	.00	–55.10	1.67	.00
Tertile 1 **Thinking Reader*	4.54	2.77	.10	–4.45	3.23	.17
Tertile 2: Middle achievement tertile (vs. highest tertile)	–34.53	1.39	.00	–29.12	1.62	.00
Tertile 2 **Thinking Reader*	1.19	2.76	.67	–3.72	3.19	.24
Tertile 1: Lowest achievement tertile (vs. middle tertile)[a]	–29.23	1.10	.00	–25.82	1.21	.00
Tertile 1 **Thinking Reader*[a]	3.75	2.18	.09	–0.95	2.40	.69
Random effect	**Variance**	**χ2(df)**	***p* value**	**Variance**	**χ2(df)**	***p* value**
Level 1	481.41			588.68		
Level 2	15.60	101.98 (55)	.00	43.42	149.47 (55)	.00
Level 3	3.34	42.05 (28)	.04	7.40	43.81 (28)	.03
N	2,156			2,149		

Note: GMRT is Gates-MacGinitie Reading Tests.

[a] The coefficient, standard error and *p* value for this contrast were obtained by changing the reference group from Tertile 3 to Tertile 2 in the multilevel model.

Source: Connecticut Department of Education (n.d. c); Massachusetts Department of Education (n.d. b); Rhode Island Department of Education (n.d. b); teacher background questionnaire administered by study team; student rosters completed by study teachers; GMRT vocabulary and comprehension subtests administered by study team.

Table E5.5 Distribution of Gates-MacGinitie Reading Tests Reading Vocabulary Pretest Scores

Reading vocabulary baseline	Number	Percent	Cumulative percent	
367& 401	7	0.29	0.41	
412 & 415	16	0.67	1.08	
423 & 426	13	0.54	1.62	
433& 436	21	0.88	2.5	
440 & 443	25	1.04	3.54	
447	27	1.13	4.56	
450	4	0.17	4.73	
453 & 456	45	1.88	6.61	
458 & 461	50	2.09	8.7	
463 & 466	74	3.1	11.8	
468	72	3.02	14.82	
472 & 474	84	3.52	18.34	
476 & 478	82	3.43	21.77	
479	96	4.02	25.8	
483	98	4.1	29.9	Upper bound Tertile 1, Sensitivity A. 1
486	94	3.94	33.84	Upper bound Tertile 1
490	103	4.31	38.15	
493	95	3.98	42.13	
496	115	4.82	46.94	Upper bound Tertile 1, Sensitivity A. 2
500	99	4.15	51.09	
503	95	3.98	55.07	
507	104	4.36	59.42	
510	115	4.82	64.24	Upper bound Tertile 2, Sensitivity A. 1
514	96	4.02	68.26	Upper bound Tertile 2
517	94	3.94	72.19	
520	84	3.52	75.71	
524	57	2.39	78.1	
527	71	2.97	81.07	
531	61	2.55	83.63	Upper bound Tertile 2, Sensitivity A. 2
534	59	2.47	86.1	
538	50	2.09	88.19	
542	38	1.59	89.78	
546	41	1.72	91.5	
550	36	1.51	93.01	
554	26	1.09	94.1	
559	26	1.09	95.18	
564	20	0.84	96.02	
569	24	1.01	97.03	
575	16	0.67	97.7	
582	15	0.63	98.32	
590	16	0.67	98.99	

Reading vocabulary baseline	Number	Percent	Cumulative percent
600	15	0.63	99.62
619 & 653	9	0.37	99.99
N	2,388		

Source: Gates-MacGinitie Reading Tests Manual for Scoring and Interpretation.

Table E5.6 Distribution of Gates-MacGinitie Reading Tests Reading Comprehension Pretest Scores

Reading comprehension baseline	Number	Percent	Cumulative percent	
396 & 402	3	0.12	0.13	
406	4	0.17	0.29	
413 & 420	4	0.16	0.46	
426 & 428	8	0.33	0.8	
432	20	0.84	1.63	
437	8	0.34	1.97	
442	31	1.3	3.27	
447	28	1.17	4.44	
451	47	1.97	6.41	
456	49	2.05	8.46	
460 & 461	59	2.47	10.93	
464	67	2.81	13.74	
467 & 469	57	2.39	16.12	
471	70	2.93	19.05	
473	3	0.13	19.18	
475	75	3.14	22.32	
478 & 480	79	3.31	25.63	
481 & 483	92	3.85	29.48	
485 & 487	80	3.35	32.83	Upper bound Tertile 1, Sensitivity A. 1
488	95	3.98	36.81	Upper bound Tertile 1
491	86	3.6	40.41	
494	101	4.23	44.64	
498	95	3.98	48.62	Upper bound Tertile 1, Sensitivity A. 2
500	91	3.81	52.43	
503	80	3.35	55.78	
506	89	3.73	59.51	
509	91	3.81	63.32	Upper bound Tertile 2, Sensitivity A. 1
512	92	3.85	67.17	Upper bound Tertile 2
516	66	2.76	69.93	
519	87	3.64	73.58	
522	96	4.02	77.6	
526	68	2.85	80.44	
530	73	3.06	83.5	

Reading comprehension baseline	Number	Percent	Cumulative percent	
534	58	2.43	85.93	Upper bound Tertile 2, Sensitivity A. 2
538	66	2.76	88.69	
543	50	2.09	90.79	
547	44	1.84	92.63	
553	53	2.22	94.85	
559	45	1.88	96.73	
566	31	1.3	98.03	
574	17	0.71	98.74	
583	17	0.71	99.46	
594	8	0.34	99.79	
617 & 652	5	0.21	100	
N	2,388			

Source: Gates-MacGinitie Reading Tests Manual for Scoring and Interpretation.

References

Anderson-Inman, L., Horney, M., Chen, D. & Lewin, L. (1994). Hypertext literacy: Observations from the ElectroText project. *Language Arts, 71*, 279–287.

Au, K. H., & Raphael, T. E. (1998). Curriculum and teaching in literature-based programs. In T. E. Raphael & K. H. Au (Eds.), *Literature-based instruction: Reshaping the curriculum* (pp. 123–148). Norwood, MA: Christopher-Gordon Publishers.

Biancarosa, G., & Snow, C. E. (2004). *Reading next: A vision for action and research in middle and high school literacy. A report to the Carnegie Corporation of New York.* Washington, DC: Alliance for Education.

Bitter, C., O'Day, J., Gubbins, P., & Socias, M. (2009). What works to improve student literacy achievement? An examination of instructional practices in a balanced literacy approach. *Journal of Education for Students Placed at Risk, 14*(1), 17–44.

Blachowicz, C. L. Z., Fisher, P. J. L., Ogle, D., & Watts-Taffe, S. (2006). Vocabulary: Questions from the classroom. *Reading Research Quarterly, 41*(4), 524–539.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). *Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions.* New York: MDRC.

Boone, R., & Higgins, K. (1993). Hypermedia basal readers: Three years of school-based research. *Journal of Special Education Technology, 12*(2), 86–106.

Brown, A. L., & Palincsar, A. S. (1985). *Reciprocal teaching of comprehension strategies: A natural history of one program for enhancing learning.* (Tech. Rep. No. 334). Urbana, IL: University of Illinois, Center for the Study of Reading.

Brown, A. L., & Palincsar, A. S. (1989). Guided, cooperative learning and individual knowledge acquisition. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 393–451). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carnegie Council on Advancing Adolescent Literacy. (2010). *Time to act: An agenda for advancing adolescent literacy for college and career success.* New York, NY: Carnegie Corporation of New York.

Cioffi, G., & Carney, J. J. (1997). Dynamic assessment of composing abilities in children with learning disabilities. *Educational Assessment 4*(3), 175–202.

Connecticut Department of Education. (n.d. a). *2008 Connecticut Mastery Test, 4th generation.* Retrieved February 1, 2010, from http://www.ctreports.com/.

Connecticut Department of Education. (n.d. b). *Public school enrollment, race and gender by grade, school and district 2007–08.* Retrieved May 29, 2009, from http://www.csde.state.ct.us/public/cedar/edfacts/enrollment/public.htm

Connecticut Department of Education. (n.d. c). *Strategic school profiles school data table 2007–08.* Retrieved May 29, 2009, from http://www.csde.state.ct.us/public/cedar/profiles/ssp_data.htm

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: Harcourt Brace Jovanovich College Publishers.

Dalton, B., Pisha, B., Eagleton, M., Coyne, P., & Deysher, S. (2002). *Engaging the text: Reciprocal teaching and questioning strategies in a scaffolded learning environment.* Peabody, MA: CAST, Inc.

Dalton, B., & Strangman, N. (2006). Improving struggling readers' comprehension through scaffolded hypertexts and other computer-based literacy programs. In D. Reinking, M. C. McKenna, L. D. Labbo, & R. D. Keiffer (Eds.), *Handbook of literacy and technology* (2nd ed., pp. 75–92). Mahwah, NJ: Lawrence Earlbaum Publishers.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192.

Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 205–242). Newark, DE: International Reading Association.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring students progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28*(4), 659–671.

Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children's comprehension and recall. *Reading Research Quarterly, 28*(3), 264–276.

Greenlee-Moore, M. E., & Smith, L. L. (1996). Interactive computer software: The effect on young children's reading achievement. *Reading Psychology: An International Quarterly,* Vol. 17 pp 43 -64.

Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 403–422). Mahwah, NJ: Erlbaum.

Guthrie, J. T., Wigfield, A., Metsala, J. L., & Cox, K. E. (1999). Motivational and cognitive predictors of text comprehension and reading amount. *Scientific Studies of Reading*, *3*(3), 231–256.

Harris, T. L. & Hodges, R. E. (1995). *The literacy dictionary: The vocabulary of reading and writing*. Newark, DE: International Reading Association.

Joffe, V. L., Cain, K., & Marić, N. (2007). Comprehension problems in children with specific language impairment: Does mental imagery training help? *International Journal of Language & Communication Disorders, 42*(6), 648–664.

Kamil, M. L. (2003). *Adolescents and literacy: Reading for the 21st century*. Washington, DC: Alliance for Excellent Education.

Kamil, M. L., Borman, G. D., Dole, J., Kral, C. C., Salinger, T., & Torgesen, J. (2008). *Improving adolescent literacy: Effective classroom and intervention practices: A Practice Guide* (NCEE #2008-4027). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved June 17, 2009, from http://ies.ed.gov/ncee/wwc

Levene, H. (1960). Robust tests for the equality of variance. In I. Olkin, S. G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honour of Harold Hotelling* (pp. 278-292). Stanford, CA: Stanford University Press.

Lipson, M. Y., & Wixson, K. K. (2003). *Assessment and instruction of reading and writing difficulty: An interactive approach* (3rd ed.). New York: Allyn and Bacon.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John W. Wiley and Sons.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2002). *Gates-MacGinitie Reading Tests, technical report Forms S and T.* Itasca, IL: Riverside Publishing.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (1999). *Gates-MacGinitie Reading Tests* (4th ed.). Itasca, IL: Riverside Publishing.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2007). *Gates-MacGinitie Reading Tests, Levels 4-6, Forms S&T: Manual for scoring and interpretation.* Rolling Meadows, IL: Riverside Publishing.

Massachusetts Department of Education, Department of Elementary and Secondary Education. (n.d. a). *School/district profiles: 2008 MCAS report.* Retrieved February 1, 2010, from http://profiles.doe.mass.edu/state_report/mcas.aspx

Massachusetts Department of Education. (n.d. b). *School/district profiles: 2007–08 enrollment by grade, race, and selected populations.* Retrieved May 29, 2009, from http://profiles.doe.mass.edu/state_report/enrollmentbygrade.aspx

McKenna, M. (1998). Electronic text and the transformation of beginning reading. In D. Reinking, M. C. McKenna, L. D. Labbo, & R. D. Kieffer (Eds.), *Handbook of literacy and technology: Transformations in a post-typographic world* (pp. 45–60). Mahwah, NJ: Lawrence Erlbaum Associates.

Meltzer, J., Smith, N. C., & Clark, H. (2001). *Adolescent literacy resources: Linking research and practice.* Providence, RI: LAB at Brown University. Retrieved February 1, 2010 from http://www.alliance.brown.edu/pubs/adlit/alr_lrp.pdf

Mokhtari, K., & Reichard, C. A. (2002). Assessing students' metacognitive awareness of reading strategies. *Educational Psychology, 94*(2), 249–259.

Moran, J., Ferdig, R. E., Pearson, P. D., Wardrop, J., & Blomeyer, R. L. (2008). Technology and reading performance in the middle-school grades: A meta-analysis with recommendations for policy and practice. *Journal of Literacy Research, 40*(1), 6–58.

Nagy, W. E., & Scott, J. A. (2004). Vocabulary processes. In R. B. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 574–593). Newark, DE: International Reading Association.

National Center for Education Statistics. (2009a). *The Nation's Report Card: Reading 2009* (NCES 2010–458). Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved April 20, 2010, from http://nces.ed.gov/nationsreportcard/pdf/main2009/2010458.pdf

National Center for Education Statistics. (2009b). *The Nation's Report Card*: *Reading 2009*: *Grade 8 target population and sample size.* Retrieved April 20, 2010, from http://nationsreportcard.gov/reading_2009/target_pop.asp?subtab_id=Tab_2&tab_id=tab1#chart

National Council of Teachers of English. (2006). *NCTE principals of adolescent literacy reform: A policy research brief.* Urbana, IL: NCTE. Retrieved February 2, 2010, from http://www.ncte.org/library/NCTEFiles/Resources/PolicyResearch/AdolLitPrinciples.pdf

National Governors Association. (2005). *Reading to achieve: A Governor's guide to adolescent literacy*. Washington, DC: Author.

National Governors Association & Council of Chief State School Officers. (2010). *Key points in English language arts draft: Key takeaways from the draft K–12 Common Core State Standards in English language arts.* Washington, DC: Author. Retrieved April 20, 2010, from http://www.corestandards.org/Files/K12ELAStandards.pdf

National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Report of the National Reading Panel.* Washington, DC: U.S. Government Printing Office.

Oakhill, J., & Patel, S. (1991). Can imagery training help children who have comprehension difficulties? *Journal of Research in Reading, 14*(2), 106–115.

Palincsar, A. S. (1982). *Improving the reading comprehension of junior high students through the reciprocal teaching of comprehension-monitoring strategies* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition & Instruction, 1*(2), 117–175.

Palincsar, A. S., & Brown, A. L. (1989). Classroom dialogues to promote self-regulated comprehension. In J. Brophy (Ed.), *Teaching for meaningful understanding and self-regulated learning* (Vol. 1, pp. 35–72). Greenwich, CT: JAI Press.

Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology, 8,* 317–344.

Pearson, P. D., Hiebert, E. H, & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly, 42*(2), 282–297.

Planty, M., Hussar, W., Snyder, T., Kena, G., Kewalramani, A., Kemp, J., . . . Dinkes, R. (2009). *The Condition of Education 2009* (NCES 2009-081). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Pressley, G. M. (1976). Mental imagery helps eight year olds remember what they read. *Journal of Educational Psychology, 68,* 355–359.

Proctor, C. P., Dalton, B., & Grisham, D. L. (2007). Scaffolding English language learners and struggling readers in a universal literacy environment with embedded strategy instruction and vocabulary support. *Journal of Literacy Research, 39*(1), 71–93.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Santa Monica, CA: RAND.

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

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2008). *HLM: hierarchical linear and nonlinear modeling* [Computer program]. Lincolnwood, IL: Scientific Software International.

Reinking, D. (1988). Computer-mediated text and comprehension differences: The role of reading time, reader preference, and estimation of learning. *Reading Research Quarterly, 23,* 484–498.

Reinking, D., & Rickman, S. S. (1990). The effects of computer-mediated texts on the vocabulary learning and comprehension of intermediate-grade readers*, Journal of Literacy Research, 22(*4), 395–411.

Rhode Island Department of Education. (n.d. a). *2007–08 NECAP reports.* Retrieved February 1, 2010, from http://reporting.measuredprogress.org/NECAPpublicRI/

Rhode Island Department of Education. (n.d. b). *Information Works!: School reports.* Retrieved May 29, 2009, from http://infoworks.ride.uri.edu/2009/default.asp

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning.* Alexandria, VA: Association of Supervision and Curriculum Development.

Rosenblatt, L. M. (1994). The transactional theory of reading and writing. In R. Ruddell, M. Ruddell, & H. Singer, (Eds.), *Theoretical models and processes of reading* (4th ed., pp. 1057–1092). Newark, DE: International Reading Association.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*(4), 479–530.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research, 66*(2), 181–221.

Schafer, J. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3–15.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87 .

Schochet, P. Z. (2008). *The late pretest problem in randomized control trials of education interventions* (NCEE 2009-4033). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The handbook of quantitative methods for the social sciences* (pp. 259–280). Thousand Oaks, CA: Sage Publications.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shin, E. C., Schallert, D. L., & Savenye, W. C. (1994). Effects of learner control, advisement, and prior knowledge on young students' learning in a hypertext environment. *Educational Technology Research and Development, 42*(1), 33–46.

Snyder, T. D., Dillow, S. A., & Hoffman, C. M. (2009). *Digest of education statistics 2008* (NCES 2009-020). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72–110.

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford Press.

Stuart, E., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children’s mental health. *American Journal of Epidemiology, 169*(9), 1133–1139.

Taylor, B. M. (2004). *School change classroom observation manual.* Minneapolis, MN: University of Minnesota.

Taylor, B. M., & Pearson, P. D. (2000). *The CIERA school change classroom observation scheme.* Minneapolis, MN: University of Minnesota.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal, 104*(1), 3–28.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2004). The CIERA school change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly, 40*(1), 40–69.

Tom Snyder Productions. (2006a). *Thinking Reader* [Computer software]. Watertown, MA: Scholastic.

Tom Snyder Productions. (2006b). *Thinking Reader teacher's guide.* Watertown, MA: Scholastic.

Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes.* Cambridge, MA: Harvard University Press.

What Works Clearinghouse. (2010). *WWC intervention report: Adolescent literacy: Reciprocal teaching.* Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved September 21, 2010, from http://ies.ed.gov/ncee/wwc/pdf/wwc_rec_teach_091410.pdf

Wigfield, A., & Guthrie, J. T. (1997). Relations of children's motivation for reading to the amount and breadth of their reading. *Journal of Educational Psychology, 89*(3), 420–432.

Wijekumar, K., Hitchcock, J., Turner, H., Lei, P. W., & Peck, K. (2009). *A multisite cluster randomized trial of the effects of Compass Learning Odyssey® Math on the math achievement of selected grade 4 students in the mid-Atlantic region* (NCEE 2009-4068). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Winograd, P., Flores-Dueñas, L., & Arrington, H. (2003). Best practices in literacy assessment. In L. Morrow, L. Gambrell, & M. Pressley (Eds.), *Best practices in literacy instruction* (2nd ed.). New York: Guilford Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 17*(2), 89–100.

www.ed.gov ies.ed.gov